Chapter 2

The Art of Laboratory Experimentation

TIMOTHY D. WILSON, ELLIOT ARONSON, AND KEVIN CARLSMITH

If you are reading this chapter you may well be a graduate student in social psychology, perhaps at the beginning of your graduate career. If so, congratulations and welcome to the field! We assume that your graduate school advisors have already taught you the secret Social Psychology Handshake that will identify you as a member of our special guild. What, they forgot? No worries—in this chapter we will teach you the handshake, metaphorically speaking. It is the methods we use that define our guild, and once you learn about these methods and begin to use them yourself, you won't need a special handshake.

To be a successful social psychologist you need good ideas, of course—astute predictions about how people will behave and brilliant explanations about why they behave that way. But in some ways that's the easy part. Transforming your ideas into hypotheses that can be tested with an elegant, tightly controlled experiment is the real challenge. As with any challenge, it can be both frustrating and great fun. As Leon Festinger once said, it is like solving a difficult puzzle:

> I love games. I think I could be very happy being a chess player or dealing with some other kinds of games. But I grew up in the Depression. It didn't seem one could survive on chess, and science is also a game. You have very strict ground rules in science, and your ideas have to check out with the empirical world. That's very tou—gh and also very fascinating. (quoted in D. Cohen, 1977, p. 133)

In this chapter we hope to convey some of this fascination with a particular kind of scientific game, namely the laboratory experiment. This is not the only method available to social psychologists, of course. We can and do use correlational designs and conduct our research in other settings (e.g., in the field). These approaches have provided us with some of our richest and most fascinating data about social phenomena, as described in another chapter in this volume (Reis & Gosling). In this chapter we hope to convey something about the approach that has been the workhorse of social psychological research, the laboratory experiment.

We have two main goals in this chapter. First, we will discuss why the laboratory experiment is often the method of choice. What are its advantages and disadvantages? Why is it used so frequently when it has some clear drawbacks? This is a timely question, because it is our impression that the use of the lab experiment has become less frequent in many areas of psychology, including social psychology. One reason for this is that social psychologists have ventured into areas in which it is more difficult to do experiments, such as the study of culture, close relationships, and the areas of the brain that correspond to social cognition and behavior (social neuroscience). Another reason is that sophisticated statistical techniques (e.g., structural equation modeling) are now available, allowing more precise tests of the relationships between variables in correlational designs. Although we welcome these advances, we fear that the unique power and value of the experimental method sometimes gets lost in the enthusiasm generated by new topics and techniques. In the first part of the chapter we will discuss the advantages of experiments in general terms.

The second part of the chapter is more of a "how-to" manual describing, in some detail, how to conduct an experiment. It is our hope that, during the first part of the chapter, we will have convinced the reader of the continued value of experiments; then, in the second part, we hope to provide detailed instructions about "how to do it" for those new to this method. We hasten to add that the best way to learn to do an experiment is to do so under the guidance of an expert, experienced researcher. Experimentation is

Address correspondence to Dr. Timothy D. Wilson, Department of Psychology, 102 Gilmer Hall, University of Virginia, Charlottesville, Virginia 22904-4400. Phone: 434-924-0674, fax: 434-982-4766, e-mail: twilson@virginia.edu.

very much a trade, like plumbing or carpentry or directing a play; the best way to learn to do it is by apprenticing oneself to a master. Nonetheless, just as it helps to read manuals explaining how to fix a leaky faucet or stage a production of Hamlet, our "how to do an experiment" manual might prove to be a helpful adjunct to a hands-on apprenticeship.

WHY DO LABORATORY EXPERIMENTS?

You have probably had something like the following experience (nearly every social psychologist has): You are at a party and someone asks you what you do. The first thing you explain is that you are not THAT kind of psychologist; you aren't Dr. Phil and you are not analyzing everyone in the room. OK, your questioner gets that, and understands that you are a research psychologist who focuses on, say, stereotyping and prejudice. "But how do you study that?" your friend asks. Now comes the hard part——explaining why you do the kind of studies you do.

Suppose, for example, that you had just published a well-known study of stereotype activation by Gilbert and Hixon (1991). "Here's an example of one of my studies," you tell your friend. The participants were white college students who watched a videotape of a woman holding up a series of cards with word fragments on them, such as P_ST. The participants' job was to make as many words from these fragments as they could within 15 seconds, such as POST or PAST. Now, unbeknownst to the participants, there were two versions of the videotape. In one the woman holding up the cards was Caucasian, whereas in the other she was Asian. This was one of the *independent variables,* you explain to your friend, looking around for some chalk and a blackboard. "That's the variable that the researcher varies to see if it has an effect on some other variable of interest (the *dependent variable*)." Your friend nods, so you continue. The other independent variable, you explain, was how "cognitively busy" or distracted people were while watching the videotape. People in the "busy" condition were asked to remember an eight-digit number, which made it difficult for them to think carefully about what they were doing. People in the "nonbusy" condition were not asked to remember any numbers.

The hypothesis was that people who had to remember the eight-digit number would not have the cognitive resources to activate their stereotype of Asians, and thus would judge the Asian woman no differently than the Caucasian women. Not busy participants, however, would have the resources to call to mind their stereotype, and thus would judge the Asian woman differently than the Caucasian woman. "But how did you measure stereotype activation?" your friend asks. Ah, you say, this was the point of the word completion task. It just so happened that five of the word fragments on the cards people saw could be completed to form words that were consistent with American college students' stereotypes about Asians. For example, the fragment S_Y could be completed to make the word "SHY," and the fragment "POLI_E" could be completed to form the word "POLITE." The measure of stereotype activation was the number of times people completed the fragments with the words that reflected the Asian stereotype.

The results were as predicted, you proudly tell your friend: People who were not busy and saw the Asian woman generated the most stereotypic words. People who were cognitively busy did not generate any more stereotypical words for the Asian as opposed to the Caucasian woman. Even better, you (actually, Gilbert & Hixon) did a second study that distinguished between the *activation* and the *application* of a stereotype, and found that the people's ratings of the Asian woman's personality were most stereotypic when they were not busy while viewing the videotape (allowing their stereotypes to be activated) but cognitively busy while listening to the assistant describe her typical day (allowing the stereotype to be applied to the woman with no inhibition).

Assuming that your friend has not left to go talk with the literature professor across the room, she is likely to have several questions. "People are being discriminated against every day on the basis of their race or gender or sexual preference and wars are being fought over ethnic identity," she says. "On the other hand, an African American has been elected president for the first time in the history of the United States. With such rich and important material to study in everyday life, why on earth are you doing a lab study in which college students watch videotapes and complete word fragments?"

Good question. Even to seasoned social psychologists, lab studies sometimes seem far removed from the problems that inspired them. Most social psychologists would agree that the perfect study would be one that was conducted in a naturalistic setting, with a diverse sample of participants, that revealed the nature and causes of an important social psychological phenomenon (such as stereotyping and prejudice). Unfortunately, such a study is like a Platonic ideal that can rarely be achieved. Experimentation almost always involves a trade-off between competing goals: the desire to study a real problem in its natural context, on the one hand, and the desire to have enough control over the setting to be able to learn something about that problem on the other. There are several important methodological points to be made here, beginning with the distinction between correlational and experimental studies.

Correlational Versus Experimental Studies

One of the points of the Gilbert and Hixon (1991) study was to examine whether the amount of cognitive resources people have influences the activation of their stereotypes. Like most social psychological questions this is a causal hypothesis, namely that one psychological variable (cognitive busyness) will have an interesting effect on another (stereotype activation). In order to test causal hypotheses, the researcher needs to have enough control over the situation to manipulate the independent variable (in this case, cognitive busyness) while keeping everything else constant. Although that is sometimes possible to do in field studies, it is much easier to accomplish in the laboratory.

To illustrate this point, think about ways in which we could test Gilbert and Hixon's hypotheses about stereotype activation in a more realistic setting. It wouldn't be easy, but maybe we could pull off a study such as the following: At a large state university, we attend the first day of classes that are taught by graduate student teaching assistants—some of whom happen to be Caucasian and some of whom happen to be Asian. We take advantage of the fact that some of the classes are held in a building that is being renovated, such that the high-pitched whine of power saws and the explosions of nail guns intrude into the classrooms. We can assume that students in these rooms are cognitively busy, because the noise makes it difficult to pay close attention to the instructor. Other classes are held in buildings in which there is no construction noise, and these students are assumed to be relatively "nonbusy." At the end of each class we ask the students to rate their instructor on various trait dimensions, including some that are part of the Asian stereotype (e.g., shyness). Suppose that the results of this study were the same as Gilbert and Hixon's: Students in the "nonbusy" (quiet) classrooms rate Asian instructors more stereotypically than students in the "busy" (noisy) classrooms (e.g., they think the instructors are more shy). There is no difference between busy and nonbusy students in their ratings of Caucasian instructors.

To many readers, this study probably seems to have some definite advantages over the one conducted by Gilbert and Hixon (1991). The measure of stereotyping—students' ratings of their TA—seems a lot more realistic and important than the word fragments people complete after watching a videotape in a psychology experiment. But, whereas the study would be interesting and possibly worth doing, it would have definite limitations. It would demonstrate a *correlation* between cognitive busyness and stereotypic ratings of Asians (at least as these variables were measured in this study), but there would be no evidence of a *causal* relationship between these variables. For example, students who took the classes in the noisy building might differ in numerous ways from students who took the classes in the quiet building. Maybe some departments offer classes in one building but not the other, and maybe students interested in some subjects have more stereotypic views of Asians than other students do. If so, the differences in ratings of the Asian instructors might reflect these differences in endorsement of the stereotype and have nothing to do with cognitive busyness. Further, there is no way of knowing whether the instructors who teach in the different buildings have similar personalities. Perhaps the Asian instructors teaching in the noisy building really were more shy than the Asian instructors teaching in the quiet building. In short, there is simply no way of telling whether students' ratings of the Asian instructors in the different buildings were due to (a) differences in their level of cognitive busyness; (b) the fact that different students took classes in the different buildings, and these students differed in their endorsement of the Asian stereotype; or (c) the fact that different instructors taught in the different buildings, and these instructors had different personalities.

One of the great advantages of an experiment is the ability to ensure that the stimuli in experimental conditions are similar. The fact that Gilbert and Hixon showed all participants the same videotape of an Asian or Caucasian woman solved one of the problems with our hypothetical correlational study: personality differences between the instructors of the courses. The fact that people who were nonbusy showed more evidence of stereotyping than people who were busy cannot be attributed to differences in the personality of the Asian women they saw on the videotape, because participants in both conditions saw the same woman.

But how do we know that this difference was not due to the fact that the students in the nonbusy condition happened to be more prejudiced toward Asians? Gilbert and Hixon (1991) solved this problem with the most important advantage of experimental designs: The ability to randomly assign people to conditions. Unlike the correlational study, people did not "self-select" themselves into the busy or nonbusy condition (i.e., by deciding which courses to take). Everyone had an equal chance of being in either condition, which means that people who were especially prejudiced against Asians were as likely to end up in one condition as the other. Random assignment is the great equalizer: As long as the sample size is sufficiently large, researchers can be relatively certain that differences in the personalities or backgrounds of their participants are distributed evenly across conditions. Any differences that are observed, then, are likely to be due to the independent variable encountered in the experiment, such as their different levels of cognitive busyness.

Our discussion of the limits of correlational designs—and the advantage of experiments—is no different from that in any introductory course in statistics or research methodology. As straightforward and obvious as these points may seem, however, they are often overlooked, by both lay people and professional researchers. To understand why, consider the following two (fictitious) investigations of the same problem. In the first, a team of researchers finds that school performance in a group of inner-city children is related to the frequency with which they eat breakfast in the morning. The more often the kids eat breakfast, the better their school performance, with a highly significant correlation of .30 (this means that the relationship between eating breakfast and school performance is moderately strong and highly unlikely to be due to chance). As far as you can tell the researchers used good measures and the study was well conducted. What do you think of this finding? Does it make you more confident that programs that provide free breakfasts for underprivileged children are having positive effects on their academic performance? If you were reviewing a report of this study for a journal, how likely would you be to recommend publication? Most of us, we suspect, would find this to be an interesting and well-conducted study that should be in the literature.

Now consider this study: A team of researchers conducts an experiment with a group of inner-city children. Half of the kids are randomly assigned to a condition in which they receive free breakfasts at school every morning, whereas the other half are in a control group that does not receive this intervention. Unfortunately, the researchers introduced a confound into their design: While the kids in the first group eat their breakfast, teachers also read to them and help them with their homework. After a few months, the researchers assess the kids' school performance, and find that those in the breakfast condition are doing significantly better than those in the control condition. The measure of academic performance is the same as in the previous study and the magnitude of the effect is the same. What do you think of this experiment? How likely would you be to recommend that it be published? The confound in the design, we would guess, is likely to be apparent and appalling to most of us. Is it eating breakfast that improved the kids' performance or the reading and extra attention from the teachers? Many of us would feel that the design of this study is so flawed that it should not be published.

But let's compare the two studies more carefully. The key question is how confident we can be that eating breakfast causes improved academic performance. The flaw in the experiment is that we cannot be sure whether eating breakfast or extra attention from a teacher or both were responsible for the improved performance. But how confident can we be from the correlational study? Kids who eat breakfast probably differ in countless ways from kids who do not. They may come from more functional families, get more sleep—or, for that matter, have parents or teachers who are more likely to help them with their homework! The experimental study, despite its flaw, rules out every single one of these alternative explanations except for one. Admittedly, this is a serious flaw; the researchers did err by confounding breakfast eating with extra attention from the teachers. But the fact remains that the correlational study leaves the door open to the same confound, and dozens or hundreds of others besides. If the goal is to reduce uncertainty about causality, surely the correlational study is much more flawed than the experimental one. Why, then, does it seem like more can be learned from the correlational study? One reason is that the correlational study was done well, by the standards of correlational designs, whereas the experimental study was done poorly, by the standards of experimental designs. Our point is that the same standard should be applied to both types of studies: How much do they reduce uncertainty about causality?

The ability to determine relationships between variables in correlational designs has improved, we should add, with the advent of sophisticated statistical techniques such as structural equation modeling. These methods allow researchers to test complex relationships between several variables and are useful techniques for distinguishing between competing models. We do not have the space to review all of the pros and cons of structural equation modeling (for excellent reviews see Kenny, this volume and Reis, 1982). Our point is that as useful as this technique is, it cannot, in the absence of experimental manipulations with random assignment, determine causal relationships. One reason for this is obvious but sometimes overlooked: It is impossible to measure all variables in a correlational design, and the researchers might have omitted one or more crucial causal variables. Thus, although there may be a direct path two variables in a structural model, one can never be sure whether this is because one variable really causes the other or whether there are unmeasured variables that are the true causes and happen to correlate highly with the measured variables. The only way to definitely rule out such alternative explanations is to use experimental designs, in which people are randomly assigned to different experimental conditions.

Validity and Realism in Experiments

We hope we have convinced the reader of the great advantage of the experiment—its ability to answer causal questions. Some, however, might still be a little uncomfortable with our conclusions, in that there is one way in

which experiments are often inferior to observational and correlational studies: They are typically done in the "artificial" confines of a psychology laboratory and involve behaviors (e.g., forming words from word fragments, remembering eight-digit numbers) that seem to have little to do with the kinds of things people do in everyday life. This is, perhaps, the most common objection to social psychological experiments—they seem "artificial" and "unrealistic." How can we generalize from such artificial situations to everyday life?

Types of Validity

Campbell and his colleagues (Campbell, 1957; Campbell & Stanley, 1963; Cook & Campbell, 1979) distinguished among different types of validity. In Campbell's taxonomy, the interpretation of research results may be assessed with respect to at least three different kinds of validity—internal validity, external validity, and construct validity.

Internal Validity Basically, *internal validity* refers to the confidence with which we can draw cause-and-effect conclusions from our research results. To what extent are we certain that the independent variable, or treatment, manipulated by the experimenter is the sole source or cause of systematic variation in the dependent variable? Threats to the internal validity of research results arise when the conditions under which an experiment is conducted produce systematic sources of variance that are irrelevant to the treatment variable and not under control of the researcher. The internal validity of a study is questioned, for instance, if groups of participants exposed to different experimental conditions are not assigned randomly and are different from each other in some important ways other than the independent variable (as in our hypothetical breakfast-eating study).

As we have seen, it is usually much easier to maintain high internal validity in a laboratory experiment, because the researcher has more control over extraneous variables that might compromise the design. Even when internal validity is high, however, there may be questions about the validity of interpretations of causal effects obtained in any given study. It is here that the distinction between external validity and construct validity becomes relevant.

External Validity This term refers to the robustness of a phenomenon: the extent to which a causal relationship, once identified in a particular setting with particular research participants, can safely be generalized to other times, places, and people. Threats to external validity arise from potential interaction effects between the treatment variable of interest and the context in which it is delivered or the type of participant population involved. When laboratory experimentation in social psychology is criticized as being "the study of the psychology of the college sophomore," what is being called into question is the external validity of the findings. Because so many laboratory experiments are conducted with college students as participants, the truth of the causal relationships we observe may be limited to that particular population (Sears, 1986). If it happens that college students—with their youth, above-average intelligence, and nonrepresentative socioeconomic backgrounds—respond differently to our experimental treatment conditions than other types of people, then the external (but not internal) validity of our findings would be low.

The issue is actually a little more subtle. No one would seriously deny that Princeton students might respond differently to a particular experimental treatment than would a sample of 50-year-old working-class immigrants or college students in another culture. External validity refers to the extent to which a particular causal relationship is robust across populations, cultures, and settings. Thus, if we were interested in the effects of lowered self-esteem on aggression, we might have to use different techniques to lower self-esteem in the two populations. Being informed that one has failed a quiz about the history of Ivy League football is likely to lower self-esteem more for Princeton sophomores than for working-class immigrants. But if we can find another technique of lowering self-esteem among that second sample, we still must ask whether this lowered self-esteem will have the same effects on aggression in both samples.

External validity is related to settings as well as to participant populations. How do we know whether the results we find in one situation (e.g., a psychology laboratory) will generalize to another situation (e.g., everyday life)? For example, Milgram's (1974) initial studies of obedience were conducted in a research laboratory at Yale University, and a legitimate question is the extent to which his findings would generalize to other settings. Because participants were drawn from outside the university and because many had no previous experience with college, the prestige and respect associated with a research laboratory at Yale may have made the participants more susceptible to the demands for compliance that the experiment entailed than they would have been in other settings. To address this issue Milgram undertook a replication of his experiment in a very different physical setting. Moving the research operation to a "seedy" office in the industrial town of Bridgeport, Connecticut, adopting a fictitious identity as a psychological research firm, Milgram hoped to minimize the reputational factors inherent in the Yale setting. In comparison with data obtained in the original study, the Bridgeport replication resulted in slightly lower but still

dramatic rates of compliance to the experimenter. Thus, setting could be identified as a contributing but not crucial factor to the basic findings of the research.

Construct Validity To question the external validity of a particular finding is not to deny that a cause and effect relationship has been demonstrated in the given research study, but rather to express doubt that the same effect could be demonstrated under different circumstances or with different participants. Similarly, concerns with *construct validity* do not challenge the fact of an empirical relationship between an experimentally manipulated variable and the dependent measure, but rather question how that fact is to be interpreted in conceptual terms. Construct validity refers to the correct identification of the nature of the independent and dependent variables and the underlying relationship between them. To what extent do the operations and measures embodied in the experimental procedures of a particular study reflect the theoretical concepts that gave rise to the research in the first place? Threats to construct validity derive from errors of measurement, misspecification of research operations, and, in general, the complexity of experimental treatments and measures. One of the most difficult parts of experimental design is constructing a concrete independent variable (e.g., memorizing an eight-digit number) that is a good instantiation of the conceptual variable (cognitive busyness). This is essentially an issue of construct validity: How well does the independent variable capture the conceptual variable?

The same issue holds for the dependent variable. When we devise an elaborate rationale for inducing our participants to express their attitudes toward the experiment or toward some social object in the form of ratings on a structured questionnaire, how can we be sure that these responses reflect the effect variable of conceptual interest rather than (or in addition to) the myriad of other complex decision rules our participants may bring to bear in making such ratings? And how do we know that the functional relationships observed between treatment and effect, under a particular set of operations, represent the conceptual processes of interest?

We can now see that the experimenter is faced with a daunting task: designing a study that is well-controlled (high in internal validity), includes independent and dependent variables that are good reflections of the conceptual variables of interest (high in construct validity), and is generalizable to other settings and people (high in external validity). Internal validity may be considered a property of a single experimental study. With sufficient knowledge of the conditions under which an experiment has been conducted, of the procedures associated with assignment of participants, and of experimenter behavior, we should be able to assess whether the results of that study are internally valid.

Issues involving construct validity and external validity, on the other hand, are more complicated. Researchers do the best they can in devising independent and dependent variables that capture the conceptual variables perfectly. But how can external validity be maximized? How can researchers increase the likelihood that the results of the study are generalizable across people and settings? One way is to make the setting as realistic as possible, which is, after all, one point of field research: to increase the extent to which the findings can be applied to everyday life, by conducting the study in real-life settings. The issue of realism however, is not this straightforward. There are several different types of realism with different implications.

Mundane Realism Versus Experimental Realism Versus Psychological Realism

Aronson and Carlsmith (1968) distinguished between two ways in which an experiment can be said to be realistic. In one sense, an experiment is realistic if the situation is involving to the participants, if they are forced to take it seriously, if it has impact on them. This kind of realism they called *experimental realism.* In another sense, the term "realism" can refer to the extent to which events occurring in the research setting are likely to occur in the normal course of the participants' lives, that is, in the "real world." They called this type of realism *mundane realism.* The fact that an event is similar to events that occur in the real world does not endow it with importance. Many events that occur in the real world are boring and unimportant in the lives of the actors or observers. Thus, it is possible to put a participant to sleep if an experimental event is high on mundane realism but remains low on experimental realism.

Mundane realism and experimental realism are not polar concepts; a particular technique may be high on both mundane realism and experimental realism, low on both, or high on one and low on the other. Perhaps the difference between experimental and mundane realism can be clarified by citing a couple of examples. Let us first consider Asch's (1951) experiment on perceptual judgment. Here the participants were asked to judge the length of lines and then were confronted with unanimous judgments by a group of peers that contradicted their own perceptions. For most participants this experiment seems to have contained a good deal of experimental realism. Whether participants yielded to group pressure or stood firm, the vast majority underwent a rather difficult experience that caused them to squirm, sweat, and exhibit other signs of tension and discomfort. They were involved, upset, and deeply concerned about the evidence being presented to them. We may assume that they were reacting to a situation

that was as "real" for them as any of their ordinary experiences. However, the experiment was hardly realistic in the mundane sense. Recall that the participants were judging a very clear physical event. In everyday life it is rare to find oneself in a situation where the direct and unambiguous evidence of one's senses is contradicted by the unanimous judgments of one's peers. Although the judging of lines is perhaps not important or realistic in the mundane sense, one cannot deny the impact of having one's sensory input contradicted by a unanimous majority.

On the other hand, consider an experiment by Walster, Aronson, and Abrahams (1966) that, although high on mundane realism, was low indeed on experimental realism. In this experiment, participants read a newspaper article about the prosecution of criminal suspects in Portugal. In the article, various statements were attributed to a prosecuting attorney or to a convicted criminal. The article was embedded in a real newspaper and hence, the participants were doing something they frequently do—reading facts in a newspaper. Thus the experiment had a great deal of mundane realism. However, nothing was happening to the participant. Very few U.S. college students are seriously affected by reading a rather pallid article about a remote situation in a foreign country. The procedure did not have a high degree of experimental realism.

Aronson, Wilson, and Akert (1994) introduced a third type of realism that they termed *psychological realism.* This is the extent to which the psychological processes that occur in an experiment are the same as psychological processes that occur in everyday life. It may be that an experiment is nothing like what people encounter in everyday life (low in mundane realism) and fails to have much of an impact on people (low in experimental realism). It could still be high in psychological realism, however, if the psychological processes that occur are similar those that occur in everyday life. Consider the Gilbert and Hixon (1991) study we described earlier. This study was low in mundane realism; in everyday life people rarely, if ever, watch a videotape of a woman holding up cards with word fragments on them and think of words to complete the fragments. It was also relatively low in experimental realism, in that the study was not very impactful or engaging. Watching the woman was probably of mild interest, but surely the study was less impactful than the Milgram or Asch studies. The study was high in psychological realism, however, to the extent that the psychological processes of stereotype activation and application were the same as those that occur in everyday life. It is common to encounter a member of a group and for a stereotype of that group to come to mind automatically. To the extent that this psychological process is the same as what occurred in Gilbert and Hixon's (1991) study, they succeeded in devising a situation that was high in psychological realism.

There is some overlap between experimental and psychological realism, in that many of the psychological processes of interest to psychologists are ones that occur when people are reacting to impactful events in their environments. Thus, the situations in everyday life in which cognitive dissonance, prejudice, or aggression occur are usually ones in which people are quite engaged. Thus, when studying these phenomena, it is imperative to devise experimental settings that are equally impactful. Such studies would be high in both experimental and psychological realism (although not necessarily high in mundane realism). Increasingly, however, social psychologists have become interested in psychological processes that occur when people are not actively engaged or motivated to process information carefully. Examples include the study of automatic processing (as in the Gilbert & Hixon study), peripheral or heuristic processing of persuasive messages (Chaiken, 1987, Petty & Cacioppo, 1986), or "mindlessness" (Langer, 1989). To study these phenomena it is important to devise experimental settings that are high in psychological realism but low in experimental realism.

External Validity: Is it Always a Goal?

Before leaving this topic it is important to make one more point about external validity and generalizability. It is often assumed that all studies should be as high as possible in external validity, in the sense that we should be able to generalize the results as much as possible across populations and settings and time. Sometimes, however, the goal of the researcher is different. Mook (1983) published a provocative article entitled, "In defense of external invalidity," in which he argued that the goal of many experiments is to test a theory, not to establish external validity. Theory-testing can take a variety of forms, some of which have little to do with how much the results can be generalized. For example, a researcher might construct a situation in which a specific set of results should occur if one theory is correct, but not if another is correct. This situation may be completely unlike any that people encounter in everyday life, and yet, the study can provide an interesting test of the two theories.

Mook (1983) gives the example of Harlow's classic study of emotional development in rhesus monkeys (Harlow & Zimmerman, 1958). Infant monkeys were separated from their mothers and placed in cages with wire-covered contraptions that resembled adult monkeys. Some of the wire monkeys were covered with terry cloth and were warmed with a light bulb, whereas others were bare and uninviting. Nourishment (in the form of a baby bottle) was sometimes available from one type of monkey and sometimes from the other. Harlow found that the monkeys clung to the terry cloth "mother," regardless of

whether it contained the bottle of milk. These results were damaging to drive-reduction theories that argued that the monkeys should prefer nourishment over emotional comfort. Was this study high in external validity? Clearly not. There was no attempt to randomly select the monkeys from those reared in the wild, or to simulate conditions that monkeys encounter in real-life settings. Nonetheless, if theories of drive reduction that were prevalent at the time were correct, the monkeys should have preferred the nourishment, regardless of which "monkey" it came from. The researchers succeeded in devising a situation in which a specific set of actions should have occurred if a particular theory was right—even though the situation was not one that would be found in everyday life. The purpose of this experiment, then, was to disprove an accepted theory rather than to establish external validity.

Mook also points out that some experiments are valuable because they answer questions about "what can happen," even if they say little about "what does happen" in everyday life. Consider Milgram's experiments on obedience to authority. As we've seen, there was little attempt to simulate any kind of real-life setting in these studies; outside of psychology experiments, people are never asked to deliver electric shocks to a stranger who is performing poorly on a memory test. The results were very informative, however, because it was so surprising that people would act the way they did under *any* circumstances. This is sometimes referred to as "proof of principle." The fact that people *can* be made to harm a complete stranger, because an authority figure tells them to, is fascinating (and frightening) despite the artificiality of the setting.

Mook's (1983) position is persuasive, and we heartily agree that the goal of many experiments is to test a theory, rather than to establish external validity. Nonetheless, we believe that even if external validity is not the main goal of study, it should never be completely forgotten. The importance of a theory, after all, depends on its applicability to everyday life. The reason Harlow's study is so important is because the theories it addresses—drive-reduction and emotional attachment—are so relevant to everyday life. The theories apply to humans as well as monkeys, and to many situations beyond cages and wire mothers. It is precisely because the *theories* are generalizable (i.e., applicable to many populations and settings) that a test of those theories is important. Thus, a specific study might test a theory in an artificial setting that is low in external validity, but why would we conduct such a study if we didn't believe that the theory was generalizable? Similarly, Milgram's results are so compelling because we can generate important, real-life examples of times when similar processes occurred. Indeed, the inspiration for Milgram's study was the Holocaust, in which seemingly normal individuals (e.g., guards at prison camps) followed the orders of authority figures to the point of committing horrific acts. Thus, if we were to conclude that the psychological processes Milgram uncovered never occur in everyday life, we could justifiably dismiss his findings. The fact that these processes appear to be similar to those that occurred at some of humankind's darkest moments—such as the Holocaust—is what makes his results so compelling.

We are essentially reiterating the importance of psychological realism in experimentation. To test a theory it may be necessary to construct a situation that is extremely artificial and low in mundane realism. As long as it triggers the same psychological processes as occur outside of the laboratory, however, it can be generalized to those real-life situations in which the same psychological processes occur. Of course, as discussed earlier, claims about psychological realism cannot be taken completely on faith; only by replicating a study in a variety of settings can external validity be firmly established.

Problem-Oriented versus Process-Oriented Research: Studying the Phenomenon Versus Studying the Process

One of your mythical friend's objections to the Gilbert and Hixon experiment was that it didn't really study the phenomena that inspired it, stereotyping and prejudice. Seeing how people complete the word fragment S_Y is far removed from ethnic conflict or housing discrimination. We have already addressed this criticism to some extent: Although it is true that people do not complete word fragments in everyday life (the study was low in mundane realism), the study may well have captured the processes by which stereotypes are activated in everyday life—that is, the study was high in psychological realism. But there is another answer to this criticism, which is related to our discussion of Mook's (1983) defense of external invalidity—whether the researcher's goal is to study a phenomenon that he or she wants to understand and possibly change, such as prejudice, or to study the underlying mechanisms responsible for the phenomenon. This distinction may seem a little odd, in that it probably seems that these goals are interdependent—and they are. To understand and change a phenomenon, it is necessary to understand the mechanisms that cause it. How can we reduce prejudice, example, without understanding the psychological mechanisms that lead to stereotype activation? In practice, however, there is a distinction to be made between research that focuses on the problem itself and research that focuses on mechanisms.

Part of this distinction involves still another one: the difference between *basic* and *applied* research. With basic research, investigators try to find the best answer to the

question of why people behave the way they do, purely for reasons of intellectual curiosity. No direct attempt is made to solve a specific social or psychological problem. In contrast, the goal in applied research is to solve a specific problem. Rather than investigating questions for their own sake, constructing theories about why people do what they do, the aim is to find ways of alleviating such problems as racism, sexual violence, and the spread of AIDS. Thus, the basic researcher is more likely to be interested in the mechanisms underlying an interesting phenomenon than the applied researcher. If applied researchers find something that works they might not be as concerned with why. In medicine, for example, there are many examples of cures that work for unknown reasons, such as the effects of aspirin on body temperature.

The distinction between problem-oriented and process-oriented research, however, involves more than the distinction between applied and basic research. To illustrate this, consider two basic researchers who are equally interested in understanding the causes of prejudice and racism. (As with many social psychological topics this is, of course, an eminently applied one as well, in that the researchers are interested in finding ways of reducing prejudice.) One researcher conducts a field study in which members of different races interact under different conditions (e.g., cooperative vs. competitive settings), to study the conditions under which reductions in prejudicial behavior occur. The other conducts a laboratory experiment on automatic processing and categorization, or the way in which people categorize the physical and social world immediately, spontaneously, and involuntarily. The stimulus materials, however, have nothing to do with race per se; in fact, the issue of race never comes up in this experiment. Participants judge a white stimulus person, under conditions thought to trigger automatic evaluations and conditions thought to trigger more controlled, thoughtful evaluations (e.g. Bargh, 1989; Brewer, 1988; Uleman, 1989).

Which study is a better investigation of prejudice and racism? Most people, we suspect, would say the former study. What does the second study have to do with prejudice? How can you possibly study racism, one might argue, without looking at behavior and attitudes of one race toward another? Herein lies our point: For researchers interested in process and mechanisms, the study of a phenomenon (such as prejudice) can involve the study of basic, psychological processes that are several steps removed from the phenomenon itself. In our view both types of studies are important: Those that study the phenomenon (e.g., racism) itself and work backward to try to discover its causes, and those that study the basic mechanisms of human perception, cognition, motivation, emotion, and behavior, and then work forward to apply these concepts to important problems (e.g., racism).

Like our earlier distinctions, we hasten to add that this one is not entirely clear-cut. Sometimes research is both problem- and process-oriented; it explores a problem and the mechanisms responsible for it simultaneously. Often, however, the focus of research on a particular problem changes as research on it progresses. As noted by Zanna and Fazio (1982), initial investigations of a problem tend to explore "is" questions: What is the phenomenon? Does it exist? These studies are, in our terms, very much problem-oriented; they establish the existence of a particular phenomenon (e.g., whether there is a stereotype based on physical attractiveness). When this question is answered, researchers typically move on to questions that have more to do with the underlying mechanisms, namely studies exploring variables that moderate or mediate the effect. Interestingly, these process-oriented studies sometimes do not study the original problem at all, focusing instead on general mechanisms that produce many different effects (as in our example of basic research on categorization and impression formation that do not study interactions between people of different races, but which are quite relevant to stereotyping and prejudice).

The Basic Dilemma of the Social Psychologist

It should be clear by now that the perfect social psychology study would be experimental instead of correlational, be extremely high in psychological realism, and study the psychological processes underlying an important phenomenon. Ideally, the study would be conducted in a naturalistic setting in which participants were randomly assigned to experimental conditions and all extraneous variables were controlled. Unfortunately, it is next to impossible to design an experiment that meets all of these demands. Indeed, almost no study ever has. One of the few exceptions, perhaps, is the Lepper, Greene, and Nisbett (1973) classic study of the overjustification effect, which was conducted in a naturalistic setting (a preschool) in which participants (3- and 4-year old children) were randomly assigned to various conditions of rewards or no rewards for drawing with felt-tip pens, and the dependent variable was how much the kids played with the pens 2 weeks later during a normal classroom activity. (An interesting social psychological parlor game is to see if you can come up with any other studies that meet all of the conditions we have laid out for the Platonic Social Psychological Experiment—there are not many.) Aronson and Carlsmith (1968) called this the basic dilemma of the experimental social psychologist. On the one hand, we want maximal control over the independent variable, to maintain internal validity. But, by maximizing internal validity, we often reduce external validity (e.g., by conducting our study in the lab instead of the field).

Programmatic Research

A solution to the basic dilemma of the social psychologist is to not try to "do it all" in one experiment. Instead, a programmatic series of studies can be conducted in which different experimental procedures are used, in different settings, to explore the same conceptual relationship. It is in this realm of conceptual replication with different scenarios that the interplay between lab and field experimentation is most clear. However, in considering these interrelationships, the tradeoff mentioned earlier between control and impact in different experimental settings becomes especially salient. In order to be defensible, weaknesses in one aspect of experimental design must be offset by strengths or advantages in other features, or the whole research effort is called into question. This dictum is particularly applicable to field experiments in which inevitable increases in cost and effort are frequently accompanied by decreases in precision and control that can be justified only if there are corresponding gains in construct validity, impact, or the generalizability of findings.

Multiple Instantiations of the Independent Variable Essentially, there are two properties that we demand of a series of experiments before we are convinced that we understand what the conceptual interpretation should be. First, we ask for a number of empirical techniques that differ in as many ways as possible, having in common only our basic conceptual variable. If all these techniques yield the same result, then we become more and more convinced that the underlying variable that all techniques have in common is, in fact, the variable that is producing the results. For example, the construct of cognitive dissonance (Festinger, 1957) has been operationalized in a wide variety of ways in both laboratory and field studies, including having people read lists of obscene words, write counter-attitudinal essays, eat unpleasant foods, and make a difficult choice between which horse to bet on at a racetrack.

Multiple Instantiations of the Dependent Variable Second, we must show that a particular empirical realization of our independent variable produces a large number of different outcomes, all theoretically tied to the independent variable. Again, we point to research on cognitive dissonance, in which a wide array of dependent variables has been used. For example, asking people to engage in unpleasant activities, under conditions of high perceived choice, has been found to influence their attitudes, their galvanic skin response while receiving electric shocks, and how hungry they are.

Systematic Replications When it comes to interpretation, there is a fundamental asymmetry between positive and negative results of replications. If proper techniques have been employed to preclude bias, successful replications speak for themselves. Failures to replicate are ambiguous, however, and therefore require supplementary information. For these reasons, good programmatic research involves replication with systematic differences in procedures and operations so that differences in results are potentially interpretable. In many cases, including exact replication along with conceptual variations are useful. Suppose, for example, that Jones, a hypothetical psychologist at the University of Illinois, produces a specific experimental result using Illinois undergraduates as participants. In addition, suppose that Smith, at Yale University, feels that these results were not a function of the conceptual variable proposed by Jones but rather were a function of some artifact in the procedure. Smith then repeats Jones's procedure in all respects save one: She changes the operations in order to eliminate this artifact. She fails to replicate and concludes that this demonstrates that Jones's results were artifactual. This is only one of many possible conclusions. Smith's failure to replicate has several possible causes and is therefore uninterpretable. It may be a function of a change in experimenter, a different participant population (Yale students may be different on many dimensions from Illinois students), or countless minor variations in the procedure such as tone of voice. Most of this ambiguity could be eliminated by a balanced design that includes an "exact" replication of the conditions run by the original experimenter. That is, suppose Smith's design had included a repeat of Jones's conditions with the suspected artifact left in, and her results approximated those of Jones's experiment. If, as part of the design, Smith changed the experiment slightly and produced no differences, or differences in the opposite direction, one could then be sure that this result was not merely a function of incidental differences like the experimenter or the participant population but must be a function of the change in the procedure. If she failed even to replicate Jones's basic experiment, the results would be much harder to interpret, because the different results could be due to any number of factors (different experimenters, different population of participants, etc.).

Non-Systematic Replications In many situations it is difficult to selectively and systematically modify the particular operational definition of the independent variable without changing the entire experimental setting. This is most dramatically true when conceptual replication involves a shift from laboratory setting to field setting. The potential complementary aspects of different research paradigms are best exemplified when operations of independent and dependent variables in laboratory procedures are significantly modified to take advantage of field settings so as to embed them appropriately in this altered context. Such modifications often involve fundamental rethinking about the conceptual

variables; it is "back to square one," with attendant costs in time and effort. If the result is a successful conceptual replication, the effort has paid off handsomely in enhanced validity for our theoretical constructs. But what if the replication fails to confirm our original findings? In this case, the multitude of procedural differences that could have increased our confidence (with a successful replication) now contributes to the ambiguity.

Now that we have discussed the value of laboratory experiments, we turn to a discussion of how to conduct one. In discussing the nuts and bolts of experimentation we will not lose sight of the important questions about the advantages and disadvantages of experiments and will return to these issues frequently.

PLANNING AND CONDUCTING A LABORATORY EXPERIMENT

The best way to describe how to conduct an experiment is to take a real study and dissect it carefully, examining how it was done and why it was done that way. We have chosen for illustrative purposes a classic laboratory experiment by Aronson and Mills (1959). We use this experiment for several reasons. First, it illustrates clearly both the advantages and the challenges of attempting to do experimental research in social psychology; we did not select it as a Platonic ideal. Second, we discuss it as an example of basic, process-oriented research that is applicable to many different phenomena, including the one near the beginning of this chapter—prejudice. At first glance this might be difficult to see, in that the Aronson and Mills (1959) study investigated the effects of the severity of an initiation on liking for a discussion group—a topic that seems far removed from the kinds of prejudice and racism we see around us today. Indeed, some aspects of the Aronson and Mills study might even seem old-fashioned; it was, after all, conducted over 50 years ago. Nonetheless, it deals with basic issues that are as fresh and important today as they were in 1959: What happens when people invest time and effort in something, such as joining a social group, that turns out to be much less enjoyable than they thought it would be? Can the psychological processes that are triggered add to our understanding of why people in contemporary society tend to be attached to their own groups to an extreme degree, and why they derogate members of other groups? The fact is that a laboratory experiment—even one conducted five decades ago—does have a lot to say about a variety of current real-world phenomena, including prejudice, because it illuminates basic, psychological processes.

Aronson and Mills set out to test the hypothesis that individuals who undergo a severe initiation in order to be admitted to a group will find the group more attractive than they would if they were admitted to that group with little or no initiation. To test this hypothesis, they conducted the following experiment. Sixty-three college women were recruited as volunteers to participate in a series of group discussions on the psychology of sex. This format was a ruse in order to provide a setting wherein people could be made to go through either mild or severe initiations in order to gain membership in a group.

Each participant was tested individually. When a participant arrived at the laboratory, ostensibly to meet with her group, the experimenter explained to her that he was interested in studying the "dynamics of the group discussion process" and that, accordingly, he had arranged these discussion groups for the purpose of investigating these dynamics, which included such phenomena as the flow of communications, who speaks to whom, and so forth. He explained that he had chosen as a topic "The Psychology of Sex" in order to attract many volunteers, as many college people were interested in the topic. He then went on to say that, much to his dismay, he subsequently discovered that this topic presented one great disadvantage; namely, that many volunteers, because of shyness, found it more difficult to participate in a discussion about sex than in a discussion about a more neutral topic. He explained that his study would be impaired if a group member failed to participate freely. He then asked the participant if she felt able to discuss this topic freely. Each participant invariably replied in the affirmative.

The instructions were used to set the stage for the initiation that followed. The participants were randomly assigned to one of three experimental conditions: a severe-initiation condition, a mild-initiation condition, or a no-initiation condition. The participants in the no-initiation condition were told, at this point, that they could now join a discussion group. It was not that easy for the participants in the other two conditions. The experimenter told these participants that he had to be absolutely certain that they could discuss sex frankly before admitting them to a group. Accordingly, he said that he had recently developed a test that he would now use as a "screening device" to eliminate those students who would be unable to engage in such a discussion without excessive embarrassment. In the severe-initiation condition, the test consisted of having people recite (to the male experimenter) a list of 12 obscene words and two vivid descriptions of sexual activity from contemporary novels. In the mild-initiation condition, the women were merely required to recite words related to sex that were not obscene.

Each of the participants was then allowed to "sit in" on a group discussion that she was told was being carried on by members of the group she had just joined. This group was described as one that had been meeting for several weeks; the participant was told that she would be replacing

a group member who was leaving because of a scheduling conflict.

To provide everyone with an identical stimulus, the experimenter had them listen to the same tape-recorded group discussion. At the same time, the investigators felt it would be more involving for participants if they didn't feel that they were just listening to a tape recording but were made to believe that this was a live-group discussion. In order to accomplish this and to justify the lack of visual contact necessitated by the tape recording, the experimenter explained that people found that they could talk more freely if they were not being looked at; therefore, each participant was in a separate cubicle, talking through a microphone and listening in on headphones. Since this explanation was consistent with the other aspects of the cover story, all the participants found it convincing.

Needless to say, it was important to discourage participants from trying to "talk back" to the tape, since by doing so they would soon discover that no one was responding to their comments. In order to accomplish this, the experimenter explained that it would be better if she did not try to participate in the first meeting, since she would not be as prepared as the other members who had done some preliminary readings on the topic. He then disconnected her microphone.

At the close of the taped discussion, the experimenter returned and explained that after each session all members were asked to rate the worth of that particular discussion and the performance of the other participants. He then presented each participant with a list of rating scales. The results confirmed the hypothesis. The women in the severe-initiation condition found the group much more attractive than did the women in the mild-initiation or the no-initiation conditions.

At first glance, this procedure has some serious problems. As with the Gilbert and Hixon (1991) study we discussed earlier, the experimenters constructed an elaborate scenario bearing little relation to the "real-life" situations in which they were interested. The "group" which people found attractive was, in fact, nothing more than a few voices coming in over a set of earphones. The participant was not allowed to see her fellow group members nor was she allowed to interact with them verbally. This situation is a far cry from group interaction as we know it outside the laboratory. In addition, reciting a list of obscene words is undoubtedly a much milder form of initiation to a group than most actual initiation experiences outside the laboratory (e.g., a college fraternity or into the Marine Corps). Moreover, the use of deception raises serious ethical problems as well as more pragmatic ones such as whether the deception was successful.

The reasons why Aronson and Mills (1959) opted to do a laboratory experiment should be clear from our earlier discussion of experimental versus correlational methods and laboratory versus field research: the ability to control extraneous variables and to randomly assign people to the different conditions. They could have opted to study real groups, such as fraternities and sororities, measuring the severity of their initiations and the attractiveness of each group to its members. Although such a study would have some advantages, we trust its disadvantages are by now clear: the inability to determine causality. Because of the inability to control extraneous variables (i.e., the actual attractiveness of the different fraternities and sororities), and the inability to randomly assign people to condition (i.e., to groups with mild or severe initiations), there would be no way of knowing whether severe initiations caused more attraction to the group. For example, it may be that desirable fraternities are inundated with applicants; because of this, they set up severe initiations to discourage people from applying. Once word gets around, only those who are highly motivated to join those particular fraternities are willing to subject themselves to severe initiations. If this were the case, it is not the severity of the initiation that caused people to find the fraternities attractive; rather, it is the attractiveness of the fraternities that produced the severity of the initiation!

Choosing the Type of Experiment to Perform

Let us assume that you are a novice researcher with a terrific idea for an experiment. The first decision you would want to make is whether to design your experiment for the laboratory or the field. It is our position that all experiments should be conducted in a variety of settings. Thus, we advocate that, ideally, all experimentally researchable hypotheses should be tested in both the laboratory and the field. As we have mentioned, each approach has its advantages and disadvantages, although it is often easier to maintain internal validity in the laboratory. So, let's suppose that you decide to start there. The next decision you must make is how to ensure that psychological realism is high in your study. Often this determines whether the experiment is to be an *impact* or a *judgment* type. In impact experiments people are active participants in an unfolding series of events and have to react to those events as they occur. Often, these events have a substantial impact on their self-views, and people thus become deeply involved in the experiment. In judgment experiments participants are more like passive observers; they are asked to recognize, recall, classify, or evaluate stimulus materials presented by the experimenter. Little direct impact on participants is intended, except insofar as the stimulus materials capture people's attention and elicit meaningful judgmental responses. Which type of study should you do?

As we mentioned, it depends on the psychological phenomenon you are studying. A researcher who was interested in the effects of sexual arousal on persuasibility would be in the domain of the impact study. It would be absurd to conduct an experiment on the effects of sexual arousal without doing something aimed at affecting the degree of sexual arousal among some of the participants. On the other hand, some hypotheses are judgmental in nature. For example, as we saw, Gilbert and Hixon (1991) hypothesized that stereotypes are more likely to be activated when people are not cognitively busy. They pointed out that interacting with a member of a stereotyped group can itself make people "busy," in that people have to think about their own actions and the impressions they are making at the same time they are forming an impression of the other person. Thus, to see whether stereotypes are more likely to be triggered when people are *not* cognitively busy, it was important to have people judge a member of a stereotyped group but not to interact with this person—in short, to make it more of a judgment than an impact study. They accomplished this by showing people a videotape of an Asian or Caucasian woman, instead of having them actually meet and interact with the woman.

The point is that researchers should tailor their method to their hypothesis. Judgment experiments are usually easier to do, because they require a less elaborate "setting of the stage" to involve the participants in an impactful situation. If researchers are interested in what happens when a person's self-concept is engaged by a series of events that happen to that person, however, there is no substitute for the impact experiment.

Experimentation and the Internet

Recent developments in Internet applications have made it possible to order groceries, adjust a thermostat, and program a television recorder from one's home, office, or wilderness retreat. It is thus not surprising that researchers have also found innovative methods to leverage this medium for psychological research. Our take on these innovations is that although they usefully expand the researcher's toolbox, the fundamental principles of experimentation that we are laying out in this chapter remain unchanged. The mechanics are different, of course, but the underlying principles are not.

Researchers have been using desktop computers for decades to conduct experiments, of course, and the addition of Internet connectivity enhances those capabilities. The most obvious and common application involves opinion surveys that can be easily administered to large groups of people with more demographic diversity than is typically found within college populations. This approach has the added bonus of being able to easily modify the surveys to create experimental manipulations, or to even dynamically react to respondents' answers.

At this point in time, Internet-based research has primarily focused on judgment-type experiments. However, there have been several instances in which researchers have designed impact type experiments that occur over the Internet. For example, Kip Williams successfully ported a laboratory paradigm for ostracism (Williams & Sommer, 1997) into the online realm (Williams, Cheung, & Choi, 2000). Participants in original paradigm sat in a waiting room with two confederates who, by design, initiated a three-way game of catch with a rubber ball. After a predetermined amount of time, the confederates excluded the participant and only tossed the ball to each other. Excluded participants found the experience to be intensely unpleasant and revealed marked cognitive deficits as a result. In the online version of this game, participants control their own computer avatar and play "catch" with other players. As in the real-life game, programmed confederates eventually exclude the participant. Williams et al. (2000) report that although the psychological impact of the ostracism is less pronounced, it is nonetheless undeniably real in the sense of having "experimental realism."

Researchers have also begun to explore the potential of immersive virtual reality experiences that promise to allow researchers to control the laboratory experiment with even more precision (e.g., Blascovich et al., 2002). When these techniques are coupled with Internet capability, it promises to increase both the impact of laboratory studies as well as the range of experimental milieus and subject populations.

The Four Stages of Laboratory Experimentation

The process of planning a laboratory experiment consists of four basic stages: (l) setting the stage for the experiment, (2) constructing the independent variable, (3) measuring the dependent variable, and (4) planning the post-experimental follow-up. In this section we will suggest ways of developing a sensible and practical *modus operandi* for each of those stages. We will be looking at both the impact experiment and the judgment experiment. It should be mentioned at the outset that the four phases listed above apply to both types of laboratory experiment. Almost without exception, however, the impact experiment is more complex and involves a wider scope of planning than does the judgment experiment. Accordingly, much of our discussion will be devoted to the high-impact type of study, not because we consider such experiments as necessarily more important but because we consider them more complex.

Setting the Stage

In designing any laboratory experiment, a great deal of ingenuity and invention must be directed toward the context,

or stage, for the manipulation of the independent variable. Because of the fact that our participants tend to be intelligent, adult, curious humans, the setting must make sense to them. It not only must be consistent with the procedures for presenting the independent variables and measuring their impact but also can and should enhance that impact and help to justify the collection of the data.

Many experiments involve deception; if deception is used, the setting must include a sensible, internally consistent pretext or rationale for the research as well as a context that both supports and enhances the collection of the data and reduces the possibility of detection. This false rationale is often referred to as a *cover story.*

In a judgment experiment, the cover story is typically less elaborate and more straightforward than in an impact experiment. Although deception is frequently used in a judgment experiment, it is usually minimal and aimed primarily at increasing the interest of the participants and providing a credible rationale for the data collection procedures and judgment task. For example, Aronson, Willerman, and Floyd (1966) performed a judgment experiment to test the hypothesis that the attractiveness of a highly competent person would be enhanced if that person committed a clumsy blunder—because the clumsy blunder would tend to humanize the person. To provide an adequate test of the hypothesis, it was necessary to expose people to one of four experimental conditions: (l) a highly competent person who commits a clumsy blunder, (2) a highly competent person who does not commit a clumsy blunder, (3) a relatively incompetent person who commits a clumsy blunder, or (4) a relatively incompetent person who does not.

What would be a reasonable context that would justify exposing people to one of these stimulus persons and inducing them to rate the attractiveness of that person? The experimenters simply informed the participants (who were students at the University of Minnesota) that their help was needed in selecting students to represent the university on the *College Bowl,* a television program pitting college students from various universities against one another in a test of knowledge. They told the participants that they could evaluate the general knowledge of the candidates objectively, but that this was only one criterion for selection. Another criterion was judgments from the participants concerning how much they liked the candidates. The experimenter then presented the participant with an audio tape recording of a male stimulus person being interviewed. This stimulus person answered a number of questions either brilliantly or not so brilliantly and either did or did not clumsily spill a cup of coffee all over himself. The participant then rated the stimulus person on a series of scales. The cover story in this experiment was simple and straightforward and did succeed in providing a credible rationale for both the presentation of the stimulus and the collection of the data.

Providing a convincing rationale for the experiment is almost always essential, since participants attempt to make sense of the situation and to decipher the reasons for the experiment. A good cover story is one that embraces all the necessary aspects of the experiment in a plausible manner and thus eliminates speculation from a participant about what the experimenter really has in mind. It also should capture the attention of the participants so that they remain alert and responsive to the experimental events. This is not meant facetiously; if a cover story strikes the participants as being a trivial or silly reason for conducting an experiment, they may simply tune out. If the participants are not attending to the independent variable, it will have little impact on them.

The setting may be a relatively simple one, or it may involve an elaborate scenario, depending on the demands of the situation. Obviously, the experimenter should set the stage as simply as possible. If a simple setting succeeds in providing a plausible cover story and in capturing the attention of the participants, there is no need for greater elaboration. A more elaborate setting is sometimes necessary, especially in a high-impact experiment. For example, suppose researchers want to make people fearful. They might achieve this goal by simply telling the participants that they will receive a strong electric shock. Yet the chances of arousing strong fear are enhanced if one has set the stage with a trifle more embellishment. This can be done by providing a medical atmosphere, inventing a medical rationale for the experiment, having the experimenter appear in a white laboratory coat, and allowing the participant to view a formidable, scary-looking electrical apparatus as in Schachter's (1959) experiments on the effects of anxiety on the desire to affiliate with others. One might go even further by providing the participant with a mild sample shock and implying that the actual shocks will be much greater.

The point we are making is that in a well-designed experiment, the cover story is an intricate and tightly woven tapestry. With this in mind, let us take another look at the Aronson and Mills (1959) experiment. Here we shall indicate how each aspect of the setting enhanced the impact and/or plausibility of the independent and dependent variables and contributed to the control of the experiment. The major challenge presented by the hypothesis was to justify an initiation for admission to a group. This was solved, first, by devising the format of a sex discussion, and second, by inventing the cover story that the experimenters were interested in studying the dynamics of the discussion process. Combining these two aspects of the setting, the experimenter could then, third, mention that because

shyness about sex distorts the discussion process, it was, fourth, necessary to eliminate those people who were shy about sexual matters by, fifth, presenting the participants with an embarrassment test.

All five aspects of the setting led directly to the manipulation of the independent variable in a manner that made good sense to the participants, thereby allaying any suspicions. Moreover, this setting allowed the experimenter to use a tape-recorded group discussion (for the sake of control) and at the same time to maintain the fiction that it was an ongoing group discussion (for the sake of impact).

This fiction of an already formed group served another function in addition to that of enhancing the involvement of the participants. It also allowed the experimenter to explain to the participant that all the other members had been recruited before the initiation was made a requirement for admission. This procedure eliminated a possible confounding variable, namely, that participants might like the group better in the severe-initiation condition because of the feeling that they had shared a common harrowing experience.

Finally, because of the manner in which the stage had been set, the dependent variable (the evaluation of the group) seemed a very reasonable request. In many experimental contexts, obtaining a rating of attractiveness tends to arouse suspicion. In this context, however, it was not jarring to the participant to be told that each member stated her opinion of each discussion session, and therefore it did not surprise the participant when she was asked for her frank evaluation of the proceedings of the meeting. Ultimately, the success of a setting in integrating the various aspects of the experiment is an empirical question: Do the participants find it plausible? In the Aronson-Mills experiment only one of 64 participants expressed any suspicions about the true nature of the experiment.

The testing of some hypotheses is more difficult than others because of their very nature. But none is impossible; with sufficient patience and ingenuity a reasonable context can be constructed to integrate the independent and dependent variables regardless of the problems inherent in the hypothesis.

Constructing the Independent Variable

One of the most important and difficult parts of experimental design is constructing an independent variable that manipulates only what you want it to manipulate. The experimenter begins with what we will call the *conceptual variable,* which is a theoretically important variable that he or she thinks will have a causal effect on people's responses. In the Aronson and Mills study, for example, the conceptual variable might be thought of as cognitive dissonance caused by an embarrassing initiation. There are many ways to translate an abstract conceptual variable such as this into a concrete experimental operation. One of the most important parts of experimental design is to devise a procedure that "captures" the conceptual variable perfectly without influencing any other factors. If we have our participants recite a list of obscene words and then listen to a boring group discussion, how can we be sure that this is, in fact, an empirical realization of our conceptual variable? Sometimes this is very difficult, and after an experiment is done, the researcher realizes that whereas participants in Conditions A and B were thought to differ only in one conceptual variable (the amount of cognitive dissonance people experienced), they also differed in some other way.

Controversy over the correct interpretation of the results obtained in the Aronson and Mills initiation experiment discussed earlier provides an example of this problem. The complex social situation used by Aronson and Mills has many potential interpretations, including the possibility that reading obscene materials generated a state of sexual arousal that carried over to reactions to the group discussion. If that were the case, it could be that transfer of arousal, rather than effort justification, accounted for the higher attraction to the group.

A replication of the initiation experiment by Gerard and Mathewson (1966) ruled out this interpretation. Their experiment was constructed so as to differ from the Aronson and Mills study in many respects. For example, Gerard and Mathewson used electric shocks instead of the reading of obscene words as their empirical realization of severe initiation (and the dissonance it produced); the shocks were justified as a test of "emotionality" rather than as a test of embarrassment; the tape recording concerned a group discussion of cheating rather than of sex; and the measure of attractiveness of the group differed slightly. Thus, sexual arousal was eliminated as a concomitant of the experimental procedures. The results confirmed the original findings: People who underwent painful electric shocks in order to become members of a dull group found that group to be more attractive than did people who underwent mild shocks. Such a confirmation of the basic initiation effect under quite different experimental operations supports, at least indirectly, the idea that it was cognitive dissonance produced by a severe initiation, and not some other conceptual variable, that was responsible for the results. A considerable amount of research in social psychology has been motivated by similar controversies over the valid interpretation of results obtained with complex experimental procedures.

The Issue of Standardization Ideally, the empirical realization of an independent variable is forceful enough to have maximum impact and clear enough to generate the intended interpretation in all participants. This section has sought to establish some important general guidelines. There

is, however, one crucial, yet frequently misunderstood, point: It is extremely important for all participants to be in the same psychological state as a result of the manipulation of the independent variable. This does not necessarily mean that all participants should be exposed to the identical independent variable. This *does* mean that the experimenter's skill and wisdom should be used to make sure that all participants arrive at a similar understanding of the instructions (or the implications of the "event" manipulation). To achieve this goal, the experimenter should take considerable latitude in delivering the instructions or experimental manipulation. This is a tricky issue and is one that may raise doubts in the minds of many investigators. Our point is this: In their zeal for standardization, many experimenters make an effort to have all instructions to the participants tape-recorded, printed, or computerized, to ensure that all participants are exposed to identical stimuli. Such an effort is admirable, but in practice it ignores the fact that people are different, and as a consequence, the same instructions do not mean the same thing to all participants. More prosaic, yet more important, participants differ greatly in their ability to understand instructions. For example, one of the most common mistakes the novice experimenter makes is to present the instructions too succinctly; consequently, a large percentage of the participants fail to understand what is going on in an experiment (especially one as complicated as most social psychological experiments are), a good deal of *redundancy* is necessary.

A brief analogy to drug trials will make our point. When a clinician tests the efficacy of an anti-anxiety medication, she administers a precise quantity of the drug to ensure that each participant develops the same concentration of the drug in his or her bloodstream. The 250-pound linebacker will be administered a significantly higher dose than the 90-pound dancer; but in the end both participants have the identical therapeutic dose of the drug in their system. In social psychological experiments it may sometimes be necessary to "titrate" the independent variable in a similar manner; one must focus on the *outcome* of the manipulation rather than the *input* of the manipulation.

We anticipate that many experimenters will disagree with us, suggesting that standardization is the hallmark of an experiment. We agree, but exactly what is it that should be standardized? What the experimenter says, or what the participant understands? We feel that the more variability there is in the participants' comprehension of the experimental operations, the more likely it will be that the changes caused by the independent variable will be obscured. This discussion again echoes the constant tension in experimentation between control and impact. However, in this case we would argue that by giving up rigid conformity to a script, one in fact can gain increased impact and decreased variability in the psychological experience of participants. Of course, by allowing the experimenter to depart from a standardized script, one may increase the possibility of introducing a systematic bias. But if proper techniques are employed to eliminate bias, this ceases to be a problem. In particular, if the experimenter who is giving the instructions is unaware of the participant's experimental condition, there is no way in which variations in the presentation can systematically bias the results.

We return now to a discussion of independent variables and how they should be administered. Recall that the essence of an experiment is the random assignment of participants to experimental conditions. For this reason, it should be obvious that any characteristics that the participants bring to the experiment cannot be regarded as independent variables in the context of a true experiment. It is not infrequent to find an "experiment" purporting to assess the effects of a participant variable (like level of self-esteem, for example) on some behavior in a specific situation. It should be clear that although such a procedure may produce interesting results, it is not an experiment because the variable was not randomly assigned. Nonrandom assignment of participants to experimental conditions is not confined to the use of personality measures in lieu of experimental treatments. It usually takes place in more subtle ways. One of the most common occurs when the experimenter is forced to perform an "internal analysis" in order to make sense out of his or her data.

The term "internal analysis" refers to the following situation. Suppose that an experimenter has carried out a true experiment, randomly assigning participants to different treatment conditions. Unfortunately, the treatments do not produce any measurable differences on the dependent variable. In addition, suppose that the experimenter has had the foresight to include an independent measure of the effectiveness of the experimental treatment. Such "manipulation checks" are always useful in providing information about the extent to which the experimental treatment had its intended effect on each individual participant. Now, if the manipulation check shows no differences between experimental treatments, the experimenter may still hope to salvage his or her hypothesis. That is, the manipulation check shows that for some reason the treatments were unsuccessful in creating the internal states in the participants that they were designed to produce. Since they were unsuccessful, one would not expect to see differences on the dependent variable. In this case, the experimenter may analyze the data on the basis of the responses of the participants to the manipulation check, resorting participants into "treatment" according to their responses to the manipulation check. This is an internal analysis.

For example, Schachter (1959) attempted to alter the amount of anxiety experienced by his participants by varying

the description of the task in which the participants were to engage. However, in some of the studies, many participants who had been given the treatment designed to produce low-anxiety actually reported higher anxiety levels than some who had been given the treatment designed to produce high anxiety. From the results of an internal analysis of these data, it does seem that anxiety is related to the dependent variable. Again, these data can be useful and provocative, but since the effect was not due to the manipulated variable, no causal statement can be made. Although many of the "highly anxious" participants were made anxious by the "high-anxiety" manipulation, many were highly anxious on their own. Because people who become anxious easily may be different from those who do not, we are dealing with an individual difference variable. This means that we can no longer claim random assignment—and, in effect, we no longer have an experiment.

Another situation in which the treatments are assigned nonrandomly occurs when the participants assign themselves to the experimental conditions. That is, in certain experimental situations the participant, in effect, is given a choice of two procedures in which to engage. The experimenter then compares the subsequent behavior of participants who choose one alternative with those who choose the other. For example, Carlsmith, Wilson, and Gilbert (2008) examined the emotional consequences of exacting revenge. They created a situation in which the participant could readily punish an offending person, and then measured participants' affect shortly thereafter. One approach would have been to create two groups—punishers and non-punishers—by allowing the participants to decide for themselves whether to punish. Such a procedure, however, would have confounded the act of punishment with participant variables. For example, it could be that people who entered the experiment in a grouchy mood would be more likely to punish and more likely to report negative affect. Thus, in order to maintain the experimental nature of the design, it was necessary to create a true control group that had no opportunity to punish and to create a punishment group in which every participant opted to punish, although technically they had the option not to. The problem of free choice is a particularly sticky one because, if the hypothesis involves the effects of choice, it is obviously important to give the participant a perception of clear choice. Yet this perception must remain nothing more than a perception, for as soon as the participant takes advantage of it, we are beset with the problems of nonrandom assignment.

Between- versus Within-Subject Designs Another decision facing the experimenter is whether to manipulate the independent variable on a between-subject or within-subject basis. In a between-subject design people are randomly assigned to different levels of the independent variable, as in the Aronson and Mills study, in which different groups of people received different levels of initiation. In a within-subjects design all participants receive all levels of the independent variable. For example, in the literature on detecting deception, participants are typically shown a videotape of another person and are asked to judge whether that person is lying or telling the truth. A number of factors have been manipulated to see how easy it is to tell whether the person is lying, such as whether the person on the tape is saying something good or bad about another person and whether the person had the opportunity to think about and plan the lie before delivering it (e.g. DePaulo, Lanier, & Davis, 1983). These factors are often manipulated on a within-subjects basis. In the DePaulo et al. (1983) study, for example, participants watched people make four statements: a planned lie, a spontaneous lie, a planned true statement, and a spontaneous true statement. The participants did not know which statement was which, of course; their job was to guess how truthful each statement was. As it turned out, people were able to detect lies at better than chance levels, but spontaneous lies were no easier to detect than planned lies.

Within-subject designs are often preferred, because fewer participants are required to achieve sufficient statistical power. Imagine that DePaulo et al. (1983) had used a between-subjects design, such that four separate groups of participants saw statements that were either planned lies, unplanned lies, planned truthful statements, or unplanned truthful statements. They probably would have had to include at least 15 people per condition, for a total of 60 participants. By using a within-design in which every participant was run in each of the four conditions, fewer people were needed (there were only 24 people who judged the statements in this study).

One reason fewer participants are needed is because each participant serves as his or her own control; each person's responses in one condition are compared to that same person's responses in the other conditions. This controls for any number of individual difference variables that are treated as error variance in a between-subjects design. Suppose, for example, that one participant has a very suspicious view of the world and thinks that people are lying most of the time. Another participant is very trusting and thinks that people seldom lie. Suppose further that a between-subjects design was used, and the distrustful and trusting people are randomly assigned to different conditions. In this design, it would be difficult to separate the effects of the independent variable (e.g., whether the person on the tape was lying or telling the truth) from how suspicious participants' are in general. With random assignment, of course, individual differences tend to cancel

out across condition; the number of suspicious versus trusting people should be roughly the same in all conditions. Nonetheless the "noise" produced by personality differences makes it difficult to detect the "signal" of the effects of the independent variable, and a large number of participants often have to be run to detect the signal. In a within-subjects design, this problem is solved by running every person in every condition. The suspicious person's responses to the lies are compared to his or her responses to the non-lies, thereby "canceling out" his or her tendency to rate everyone as deceptive.

If a within-subject design is used it is important, of course, to vary the order of the experimental conditions, to make sure that the effects of the independent variable are not confounded with the order in which people receive the different manipulations. This is referred to as "counterbalancing," whereby participants are randomly assigned to get the manipulations in different orders. In the DePaulo et al. (1983) study, for example, the presentation of the deceptive versus nondeceptive statements and planned versus unplanned statements was counterbalanced, such that different participants saw the statements in different orders.

In many social psychological experiments within-subject designs are not feasible, because it would not make sense to participants to evaluate the same stimulus more than once under slightly different conditions. For example, in the experiment by Aronson, Willerman, and Floyd, once a participant was exposed to a tape recording of a competent person spilling coffee, it would have been ludicrous to present that same participant with an otherwise identical tape of a competent person who doesn't spill coffee. Who would believe that there are two people in the world who are identical in all ways except for their coffee-spilling behavior? By the same token, in the vast majority of impact experiments, the nature of the impactful manipulation precludes utilization of the same participants in more than one condition. For example, in the Aronson and Mills experiment, once the experimenters put a participant through a severe initiation in order to join a group and then asked her to rate the attractiveness of that group, it would have been silly to ask her to start all over and go through a mild initiation. Thus, within-subjects designs are preferable if at all possible, but in many studies—especially impact experiments—they are not feasible.

Avoiding Participant Awareness Biases It is arguably more challenging to perform a meaningful experiment in social psychology than in any other scientific discipline for one simple and powerful reason: In social psychology, we are testing our theories and hypotheses on adult human beings who are almost always intelligent, curious, and experienced. They are experienced in the sense that they have spent their entire lives in a social environment and—because of their intelligence and curiosity—they have formed their own theories and hypotheses about precisely the behaviors we are trying to investigate. That is to say, everyone in the world, including the participants in our experiments, is a social psychological theorist.

In a nutshell, the challenge (and the excitement) of doing experiments in social psychology lies in the quest to find a way to circumvent or neutralize the theories that the participants walk in with so that we can discover their true behavior under specifiable conditions, rather than being left to ponder behavior that reflects nothing more than how the subjects think they should behave in a contrived attempt to confirm their own theory.

One special form of participant awareness is closely related to the idea of "demand characteristics" as described by Orne (1962). The term refers to features introduced into a research setting by virtue of the fact that it *is* a research study and that the participants know that they are part of it. As aware participants, they are motivated to make sense of the experimental situation, to avoid negative evaluation from the experimenter, and perhaps even to cooperate in a way intended to help the experimenter confirm the research hypothesis (Sigall, Aronson, & Van Hoose, 1970). Such motivational states could make participants responsive to any cues—intended or unintended—in the research situation that suggest what they are supposed to do to appear normal or "to make the study come out right." It is for this reason that experimenters frequently employ deception, elaborate cover stories, and the like, in an attempt to keep participants unaware of the experimental manipulations in play.

Another aspect of the problem of demand characteristics and participant awareness is the possibility that the experimenter's own behavior provides inadvertent cues that influence the responses of the participants. In our experience novice researchers often dismiss this possibility; they smile knowingly and say, "Of course *I* wouldn't act in such a way to bias people's responses." Decades of research on expectancy effects, however, show that the transmission of expectations from researchers to participants is subtle and unintentional, and that this transmission can have dramatic effects on participants' behavior. It can occur even between a human experimenter and an animal participant; in one study, for example, rats learned a maze quickly when the experimenter thought they were good learners and slowly when the experimenter thought they were poor learners (Rosenthal, 1994; Rosenthal & Lawson, 1964).

Therefore, steps must be taken to avoid this transmission of the experimenter's hypotheses to the research participants. One way of doing so is to keep the experimenter unaware of the hypothesis of the research. The idea here is that if the experimenter does not know the hypothesis, he or she

cannot transmit the hypothesis to the research participants. In our judgment, however, this technique is inadequate. One characteristic of good researchers is that they are hypothesis-forming organisms. Indeed, as we mentioned earlier, this is one characteristic of all intelligent humans. Thus, if not told the hypothesis, the research assistant, like a participant, attempts to discover one. Moreover, keeping the assistant in the dark reduces the value of the educational experience. Since many experimenters are graduate students, full participation in an experiment is the most effective way of learning experimentation. Any technique involving the experimenter's ignorance of the hypothesis or a reduction in contact with the supervisor is a disservice to him or her. A more reasonable solution involves allowing the experimenters to know the true hypothesis but keeping them ignorant of the specific experimental condition of each participant. This is typically referred to as a "double-blind" study in which both the experimenter and participant are unaware of the experimental condition. In theory, this is a simple and complete solution to the problem and should be employed whenever possible.

In a study by Wilson et al. (1993), for example, the independent variable was whether people were asked to think about why they felt the way they did about some art posters, to examine the effects of introspection on attitude change and satisfaction with consumer choices. Participants were told that the purpose of the study was to examine the different types of visual effects that people like in pictures and drawings and that they would be asked to evaluate some posters. The critical manipulation was whether people wrote down why they felt the way they did about each poster (the reasons condition) or wrote why they had chosen their major (the control condition). To assign people to condition randomly, the experimenter simply gave them a questionnaire from a randomly ordered stack. To make sure the experimenter did not know whether it was the reasons or control questionnaire, an opaque cover sheet was stapled to each one. The experimenter left the room while the participant completed the questionnaire, and thus throughout the experiment was unaware whether the participant was in the reasons or control condition.

In other types of experiments, the experimental manipulations cannot be delivered simply by having people read written instructions, making it more difficult to keep the experimenter unaware of condition. In studies on intrinsic motivation, for example, the critical manipulation is the level of reward people believe they will get for performing a task. This could be conveyed in written form, but there is a risk that participants will not read the questionnaire carefully enough, missing the crucial information about the reward. A frequently used solution to this problem is to tape record the instructions, and to keep the experimenter unaware of which recorded instructions each participant receives (e.g. Harackiewicz, Manderlink, & Sansone, 1984).

In other studies, however—particularly high impact ones—the experimenter must deliver the independent variable in person, making it more difficult for him or her to be unaware of participant's experimental condition. In the Aronson and Mills experiment, for example, people's condition was determined by which list of words they had to read aloud to the experimenter. The experimenter could have given people a list and asked them to read the words to themselves, but this obviously would have reduced the impact of the manipulation considerably. In studies such as these, where it is necessary for the experimenter to "deliver" the independent variable, several steps can still be taken to avoid demand characteristics, participant awareness biases, and experimenter expectancy effects. First, the experimenter should be kept ignorant of people's condition until the precise moment of crucial difference in manipulations. That is, in most studies, the experimenter need not know what condition the participant is in until the crucial manipulation occurs. When the choice point is reached, a randomizing device can be used, and the remainder of the experiment is, of course, not carried out in ignorance. For example, in the Aronson and Mills study, it would have been easy to delay assignment of participant to condition until the point of initiation; by reaching into a pocket and randomly pulling out one of three slips of paper, the experimenter could determine whether the participant would recite the obscene words, the mild words, or no words at all. Thus, all the pre-manipulation instructions would be unbiased.

This is only a partial solution because the experimenter loses his or her ignorance midway through the experiment. However, if the experimenter left the room immediately after the recitation and a different experimenter (unaware of the participant's experimental condition) collected the data, this solution would approach completeness. The use of multiple experimenters, each ignorant of some part of the experiment, offers a solution that is frequently viable. For example, Wilson and Lassiter (1982) were interested in whether prohibiting people from engaging in unattractive activities would increase the appeal of those activities; that is, whether the Aronson and Carlsmith (1963) "forbidden toy" effect would apply when the prohibited activity was undesirable at the outset. The participants were preschool children who were seen individually. In one condition the experimenter showed the child five toys and said that he or she could play with any of them but a plastic motorcycle, which was known to be unattractive to the children. In the control condition the children were allowed to play with all five toys. As we have discussed, the experimenter randomly assigned people to condition at the last possible

moment, namely after he had shown the children all the toys and demonstrated how they worked.

To assess children's subsequent interest in the toys, the children were seen again a week later and given two of the toys to play with—the plastic motorcycle and another, attractive toy. At this session the same experimenter could not be used, however, because he was no longer unaware of the child's experimental condition. Further, his presence might cause children to base their choice on factors other than their liking; for example, they might be concerned that he still did not want them to play with the motorcycle. Thus, a different experimenter (unaware of the child's condition) was used, and the children were not told that this session was part of the same study as the first session. As predicted, the children who were prohibited from playing with the motorcycle in the first session played with it significantly more at the second session than did people in the control condition.

Returning to the more general issue of demand characteristics, it should be clear that the most effective type of deception in an impact experiment involves the creation of an independent variable as an event that appears not to be part of the experiment at all. Creating such an independent variable not only guarantees that the participant will not try to interpret the researcher's intention but also that the manipulation has an impact on the participant. Several classes of techniques have been used successfully to present the independent variable as an event unrelated to the experiment. Perhaps the most effective is the "accident" or "whoops" manipulation, in which the independent variable is presented as part of what appears to be an accident or unforeseen circumstance. Wilson, Hodges, and LaFleur (1995) used a variation on this procedure to influence people's memory for behaviors performed by a target person. These researchers showed people a list of positive and negative behaviors the target person had performed and then wanted to make sure that people found it easiest to remember either the positive or negative behaviors. They did so by simply showing people either the positive or negative behaviors a second time. The danger of this procedure, however, is that it would be obvious to people that the researchers were trying to influence their memory. If Wilson et al. had said, "OK, now we are going to show you only the positive (negative) behaviors again," participants would undoubtedly have wondered why and possibly figured out that the point was to influence their memory for these behaviors. To avoid this problem, Wilson et al. told people that they would see all of the behaviors again on slides. After only positive (or negative) ones had been shown, it just so happened that the slide projector malfunctioned. The projector suddenly went dark, and after examining it with some frustration, the experimenter declared that the bulb was burned out. He searched for another for a while, unsuccessfully, and then told participants that they would have to go on with the study without seeing the rest of the slides. By staging this "accident," the researchers ensured that people were not suspicious about why they saw only positive or negative behaviors a second time.

Another way to make the independent variable seem like a spontaneous, unrelated event is to have a confederate, apparently a fellow participant, introduce the manipulation. For example, Schachter and Singer (1962) attempted to manipulate euphoria by having a confederate waltz around the room shooting rubber bands, play with hula hoops, and practice hook shots into the wastebasket with wadded paper. Presumably, this behavior was interpreted by the participant as a spontaneous, unique event unrelated to the intentions of the experimenter. A third method is to use the whole experimental session as the independent variable and to measure the dependent variable at some later time. For example, in the Wilson and Lassiter (1984) study mentioned earlier, the independent variable (whether people were constrained from playing with an unattractive toy) was introduced at one session, and the dependent variable (how long people played with the toy) was assessed at another session a week later. It is unlikely that the participants realized that what happened in the first study was the independent variable of interest. Even within the same experimental session it is possible to convince people that they are taking part in separate, unrelated experiments. A common ruse is the "multiple study" cover story, in which people are told that for reasons of convenience several unrelated mini-experiments are being conducted at the same session. This ruse is commonly employed in priming experiments, in which it is very important that people not connect the independent variable (the priming of a semantic category) with the dependent variable (ratings of a target person whose standing on that category is ambiguous). Higgins, Rholes, and Jones (1977), for example, had people memorize words related to adventurousness or recklessness as part of an initial "Study 1" concerned with perception and memory, and then had people rate a stimulus person, whose behavior was ambiguous as to whether it was adventurous or reckless, as part of a "Study 2" on impression formation.

Optimizing the Impact of the Independent Variable As we mentioned, one problem with keeping experimenters unaware of condition, by delivering the independent variable in written form, is that the impact of the independent variable will be reduced. One of the most common mistakes the novice experimenter makes is to present instructions too briefly; consequently, a large percentage of the participants fail to understand some important aspects of the instructions.

To ensure that all participants understand what is going on in an experiment (especially one as complicated as most social psychological experiments), the instructions should be repeated in different ways.

More important than simple redundancy, however, is ensuring the instructions are expressed precisely so that each participant fully understands them and the events that occur in the experiment. This can be accomplished by a combination of written and verbal instructions, in which the experimenter repeats or paraphrases key parts of the instructions until satisfied that the participant is completely clear about all of them. Although the point seems obvious, it has been our experience that many experiments fail precisely because the instructions were never made clear enough to become understandable to all the participants.

In the well-designed impact experiment, there is less likely to be a question about whether the participant is paying attention to the relevant stimulus conditions. Nonetheless the experimenter should be as certain as possible that the complex bundle of stimuli constituting the independent variable produce the intended phenomenological experience in the participants. For this purpose, there is no substitute for the thorough pretesting of the manipulation. During the pretesting, the experimenter can conduct long, probing interviews with the participant after the test run of the experiment is completed or, better yet, after the manipulation of the independent variable.

One of the most frequently misunderstood aspects of experimentation is the amount of pretesting that is often required to make sure that the independent variable is having the desired impact. When students read published experiments in psychological journals, they often have the impression that the researchers had an idea, designed a study, collected the data in a few weeks, analyzed the data, and presto, found exactly what they predicted. Little do they know that in most cases the experiment was preceded by a good deal of pretesting, whereby different versions of the independent variable were "tried out." For example, in the Wilson, Hodges, and LaFleur (1995) study mentioned earlier, in which the researchers staged a malfunction of a slide projector, a good deal of pretesting was required to "fine tune" this manipulation. Different versions of the manipulation were tried before one was found that worked convincingly.

This might seem to be misleading, in that the researchers ended up reporting only the version of the independent variable that had the desired effect. It is important to note, however, that there are two meanings of the phrase "desired effect": (a) whether the researchers manipulated what they intended to manipulate and (b) whether the independent variable had the predicted effect on the dependent variable. An experiment cannot test a hypothesis unless the independent variable manipulates what it is supposed to manipulate. For example, in the Wilson et al. (1995) study, the point was to see what happens when people analyze the reasons for their impressions of a person and either positive or negative thoughts about that person are most accessible in memory. The hypotheses of the study could only be tested if the manipulation of people's memory succeeded in making positive or negative thoughts more accessible. The ability to play with a design so that the manipulations change the right variables is a skill similar to that of a talented director who knows exactly how to alter the staging of a play to maximize its impact on the audience. This is where some of the most important work in experimental design occurs, but it is rarely reported in published articles, because it would not be very informative or interesting to begin the methods section by saying, "We will first tell you about all the ways of manipulating the independent variable that didn't work. The first mistake we made was…"

It is another matter, however, if the manipulation works as intended but does not influence the dependent variable in the predicted manner. Another reason that a manipulation can fail to have an effect is because the researcher's hypothesis is wrong. The manipulation might work exactly as intended (as indicated, for example, on a manipulation check), but have a different effect on the dependent variable than predicted. This *is* informative, because it suggests that the hypothesis might be wrong. The catch is that it is often difficult to tell whether an experiment is not working because the manipulation is ineffective or because the hypothesis is wrong. The answer to this question often becomes clear only after extensive tinkering and restaging of the experimental situation.

Once it becomes clear that the manipulation is working as intended but the hypothesis is off the mark, a second talent comes into play: The ability to learn from one's mistakes. Some of the most famous findings in social psychology did not come from reading the literature and deducing new hypotheses, or from "aha" insights while taking a shower. Rather, they came about from the discovery that one's hypotheses were wrong and the data suggest a very different hypothesis—one that is quite interesting and worth pursuing faithfully.

Choosing the Number of Independent Variables We have been talking thus far of the independent variable in the social psychological experiment as if it were a simple two-level variation on a single dimension. Yet many, if not most, experiments involve procedures that simultaneously manipulate two or more variables. Once one has taken the time and trouble of setting up a laboratory experiment, recruiting participants, and training research assistants, it seems only efficient to take the occasion to assess the effects of more than one experimental treatment.

There are no pat answers to the question of how many independent variables can or should be manipulated at one time, but our own rule is that an experiment should be only as complex as is required for important relationships to emerge in an interpretable manner. Sometimes it is essential to vary more than one factor because the phenomenon of interest appears in the form of an interaction. Petty, Cacioppo, and Goldman (1981), for example, hypothesized that the way in which people process information in a persuasive communication depends on the personal relevance of the topic. When the topic was highly relevant, people were predicted to be most influenced by the strength of the arguments in the communication, whereas when it was low in relevance, people were predicted to be most influenced by the expertise of the source of the communication. To test this hypothesis the authors had to manipulate (a) the personal relevance of the topic, (b) the strength of the arguments in the message, and (c) the expertise of the source of the message. Only by including each of these independent variables could the authors test their hypothesis, which was confirmed in the form of a three-way interaction.

Measuring the Dependent Variable

The basic decision facing the researcher in planning the measurement of dependent variables is whether to rely on participants' self-reports or observations by others as the means of assessing a person's responses to the experimental situation. Actually, it is not that simple, for it is possible to imagine a continuum ranging from behaviors of great importance and consequence for the participant down to the most trivial paper-and-pencil measures about which the participant has no interest. At one extreme the experimenter could measure the extent to which participants actually perform a great deal of tedious labor for a fellow student (as a reflection of, say, their liking for that student). At the other extreme one could ask them to circle a number on a scale entitled "How much did you like that other person who participated in the experiment?" Close to the behavioral end of the continuum would be a measure of the participant's commitment to perform a particular action without actually performing it. We call this a "behavioroid" measure.

It is a fair assumption to say that most social psychologists care the most about social behavior: how people treat each other and how they respond to the social world. The goal is not to explain and predict which number people will circle on a scale or which button on a computer they will press, but people's actual behavior toward another person or the environment. Thus, the first choice of a dependent measure in a social psychological experiment is usually overt behavior. The ideal measure of prejudice is the way in which members of different groups treat each other, the ideal measure of attitude change is behavior toward an attitude object, and the ideal measure of interpersonal attraction is affiliative behaviors between two individuals. If you pick up a copy of a recent social psychological journal, however, you will find that measures of actual behavior are hard to come by (Baumeister, Vohs, & Funder, 2007; de la Haye, 1991). The dependent measures are more likely to be such things as questionnaire ratings of people's thoughts, attitudes, emotions, and moods; their recall of past events; the speed with which they can respond to various types of questions; or, as we saw in the Gilbert and Hixon (1991) study, the ways in which people complete word fragments.

There are four main reasons why social psychologists often measure things other than actual behavior. The first is convenience: It is much easier to give people a questionnaire on which they indicate how much they like a target person, for example, then to observe and code their actual behavior toward the target person. Of course, convenience is no excuse for doing poor science, and the assumption that questionnaire responses are good proxies for actual behavior should not be taken on faith. In the early years of attitude research, for example, it was assumed that people's questionnaire ratings of their attitudes were good indicators of how they would actually behavior toward the attitude object. It soon became apparent that this was often not the case (e.g., Wicker, 1969), and many researchers devoted their energies to discovering when questionnaire measures of attitudes predict behavior and when they do not. A large literature on attitude-behavior consistency was the result, and it is now clear that self-reported attitudes predict behavior quite well under some circumstances but not others (e.g., Fazio, 1990; Wilson, Dunn, Kraft, & Lisle, 1989).

Needless to say, there are some situations in which obtaining a direct measure of the behavior of interest is not simply inconvenient, it is virtually impossible. For example, Aronson and his students conducted a series of laboratory experiments aimed at convincing sexually active teenagers to use condoms as a way of preventing AIDS and other sexually transmitted diseases (Aronson, Fried, & Stone, 1991). The ideal behavioral dependent variable is obvious: whether the participants in the experimental condition actually used condoms during sexual intercourse to a greater extent than participants in the control condition. Think about it for a moment: How would you collect those data? Even experimental social psychologists feel obliged to stop short of climbing into bed with their subjects in order to observe their sexual behavior directly. Aronson and his students were forced to use proxies. In some of their studies, they used self-report as a proxy. In others, in addition to self-report, they set up a situation where, at the close of the experiment, the experimenter while leaving the room, indicated that the participants, if they wanted, could purchase condoms (at a bargain price) by helping themselves

from huge a pile of condoms on the table and leaving the appropriate sum of money. Although the participants had no way of suspecting that their behavior was being monitored, as soon as they left the laboratory, the experimenter returned and re-counted the condoms on the table to ascertain exactly how many they had purchased. Admittedly, the number of condoms *purchased* is not quite as direct a measure as the actual *use* of condoms, but especially given the fact that this measure was consistent with self-report measures, it seems like a reasonable proxy.

A second reason behavioral measures are sometimes avoided has to do with our earlier distinction between problem-oriented and process-oriented research. If the research is problem-oriented, then the dependent measures should correspond as closely to that phenomenon (e.g., prejudice, consumer behavior, condom use) as possible. If it is process-oriented, however, the goal is to understand the mediating processes responsible for a phenomenon, and the dependent measures are often designed to tap these processes, not the phenomena they produce. For example, to understand when people will act in a prejudiced manner toward a member of a social group, it is important to know when their stereotype of that group is activated. As we saw earlier, Gilbert and Hixon (1991) addressed this question by showing people a videotape of a woman holding up cards with word fragments on them and asking people to complete the fragments to make as many words as they could. The main dependent measure was the number of times people completed the fragments with words that were consistent with Caucasians' stereotypes of Asians to see if this differed according to whether the woman on the tape was Asian and whether people were under cognitive load. Note that the researchers never measured people's behavior toward Asians—participants never interacted with anyone except the experimenter. How, then, can this be an experiment on stereotyping and prejudice? It is by studying some of the psychological processes (stereotype activation) hypothesized to mediate prejudicial behavior.

A third reason why nonbehavioral measures are often used is that, in many situations, that they can be a more precise measure of intervening processes than overt behavior. Behavior is often complex and multidetermined, making it difficult to know the exact psychological processes that produced it. For example, suppose in an experiment a confederate (posing as a fellow participant) either praises the participant, implying that he or she is brilliant, or insults the participant, implying that he or she is stupid. Suppose our dependent variable is how much the participant likes the confederate. We can measure it by handing participants a rating scale and asking them to rate their liking for the confederate, from +5 to –5. Or, on a more behavioral level, we can observe the extent to which the participant makes an effort to join a group to which the confederate belongs. This latter behavior seems to be a reflection of liking, but it may reflect other things instead. For example, it may be that some participants in the "insult" condition want to join the group in order to prove to the confederate that they are not stupid. Or it may be that some want an opportunity to see the insulting person again so that they can return the favor. Neither of these behaviors reflects liking, and consequently, may produce results different from those produced by the questionnaire measure.

Fourth, some psychological states are best measured by self-report instruments—that is, by asking people—than by observing their behavior. In recent years, for example, there has been a considerable amount of interest in human happiness, such as what causes it, how well people can predict it, and whether it can be changed (e.g., Diener & Biswas-Diener, 2008; Gilbert, 2006; Wilson & Gilbert, 2003). Researchers have conducted a great deal of psychometric work on how best to measure how happy people are, and it turns out that the most valid and reliable way is to ask them (Andrews & Robinson, 1991; Diener, 1994; Fordyce, 1988). Thus, in some cases self-report instruments are the best measure of the phenomenon researchers are trying to assess.

Nonetheless, it is important to note some limitations of questionnaire measures. Most fundamentally, people may not know the answer to the questions they are asked. This is especially true of "why" questions, whereby people are asked to report the reasons for their behavior and attitudes. Rather than reporting accurately, people might be relying on cultural or idiosyncratic theories about the causes of their responses that are not always correct (Nisbett & Wilson, 1977; Wilson, 2002).

Disguising the Measure Even if people know the answer to a question, they may not answer truthfully. As previously mentioned, people might distort their responses due to self-presentational concerns or because they have figured out the hypothesis and want to tell the experimenters what they want to hear. It is thus often important to disguise the fact that a particular collection of data is actually the measurement of the dependent variable. This presents problems very similar to those involved in attempting to disguise the independent variable, as discussed in the earlier section on guarding against demand characteristics. Again, there are several classes of solutions that can be applied to the problem of disguising the dependent variable.

One approach is to measure the dependent variable in a setting that participants believe is totally removed from the remainder of the experiment. For example, in research on intrinsic motivation it is common to assess people's interest

in an activity by observing how much time they spend on that activity during a "free time" period. Participants believe that this time period is not part of the experiment and do not know that they are being observed. Lepper, Greene, and Nisbett (1973), for instance, measured children's interest in a set of felt-tip pens by unobtrusively observing how much time they spent playing with the pens during a free-play period of their preschool class.

Another example of how the dependent measure can be disguised comes from the Wilson et al. (1993) study mentioned earlier, in which people either analyzed why they liked some posters or did not. One hypothesis of this study was that people who analyzed reasons would change their minds about which posters they preferred the most, and would thus choose different types of posters to take home than people in the control condition. To test this hypothesis the experimenter told people, at the end of the study, that as a reward for their participation, they could choose one poster to take home. Asking people to make their choice in front of the experimenter would have been problematic, because self-presentational biases might have come into play, whereby people chose a poster on the basis of how this made them look to the experimenter, rather than on the basis of which one they really liked the best (Baumeister, 1982; DePaulo, 1992; Schlenker, 1980). The posters were of different types; some were reproductions of classic paintings, whereas others were more contemporary, humorous posters. Participants might have thought, "I would prefer one of the humorous posters but this might make me look shallow and inane, so I will go ahead and take the one by Monet."

To minimize self-presentational biases, Wilson et al. took the following steps to make the choice of poster as private as possible: After telling the participant that she could choose a poster to take home, the experimenter said that she had to go get the explanation sheet describing the purpose of the study. She told the participant to pick out a poster from bins that contained several copies of each poster, and then left the room. The participant expected the experimenter to return shortly, and thus may still have been concerned that the experimenter would see which poster she chose. To minimize such a concern, the researchers placed multiple copies of each poster in each bin. Further, all the posters were rolled up so that only the reverse, blank side was showing, making it impossible (in the minds of the participants) for the experimenter to tell which poster she had chosen. (After the participant had left, the experimenter was able to tell which poster people chose by counting the number left in each bin.) It is possible that despite these rather elaborate precautions, some participants were still motivated to choose posters that would make them look good rather than ones they really liked. It is important to minimize such self-presentational concerns, however, as much as possible. As it happened, Wilson et al.'s predictions were confirmed: People who analyzed reasons chose different types of posters than people who did not.

A similar approach is to tell participants that the dependent variable is part of a different study than the one in which the independent variable was administered. As mentioned earlier the "multiple study" cover story can be used, in which participants think they are taking part in separate studies (e.g., Higgins et al., 1977).

If the independent and dependent variables are included in the same study, steps are often taken to disguise the purpose of the dependent measure. For example, there is a family of techniques for measuring a dependent variable that is parallel to the "whoops" procedure for manipulating an independent variable. The most common member of this family involves claiming that the pretest data were lost so that a second set of measures must be collected. In attitude-change experiments, the most typical solution is to embed the key items in a lengthy questionnaire that is given to the participant. One may have some qualms about the extent to which this always disguises the measurement from the participant, yet it has been used effectively in some instances.

Dependent Measures that are Uncontrollable All of the above ways of disguising the dependent measure make the assumption that if people knew what was being measured, they might alter their responses. The prototypical example of such a measure is the questionnaire response; if people are asked on a 7-point scale whether they would help someone in an emergency, they might indicate how they would like to respond, or how they think they should respond, instead of according to how they really would respond. There is another way of avoiding this problem: Use dependent measures that by their very nature are uncontrollable, such that people could not alter their responses even if they wanted to—obviating the need to disguise the measure. Controllability is a matter of degree, of course; it is more difficult to control one's heart rate than one's response on a 7-point scale, but even heart rate can be controlled to some degree (e.g., by holding one's breath). Social psychologists have broadened their arsenal of dependent measures considerably in recent years, and for present purposes it is interesting to note that many of these measures are more difficult for people to control than questionnaire responses, and less susceptible to demand characteristics or self-presentational concerns. Examples include measures of physiological and neurological responses, as well as the virtual explosion of measures of automatic cognitive and affective responses (e.g. Bargh, 1990; de Houwer, in press; Gilbert, 1991; Greenwald & Banaji, 1995; Nosek, Greenwald, & Banaji, 2007; Wegner, 1994).

An obvious advantage of such measures is that they are difficult for people to control. Another is that they may tap psychological constructs that are distinct from what people are able to report. We do not have the space here to review the vast literatures on implicit measures and unconscious mental processes (see the chapters in this volume by Banaji and Dijksterhuis). Suffice it to say that this is an exciting time for social psychological theory and methods; new frontiers are opening as researchers develop new tools and methods to tap psychological processes.

One of these frontiers is social neuroscience, whereby researchers measure the neural correlates of social psychological processes, typically using functional Magnetic Resonance Imaging (fMRI) to measure blood flow in the brain and electroencephalography (EEG) to measure event-related potentials (ERP). These measures can be quite useful in identifying the regions of the brain that are active during a particular task, and thus to infer some of the underlying social psychological processes that drive a given behavior. For example, Greene et al. (2001; 2004) demonstrated that different sorts of moral dilemmas activate different regions of the brain, suggesting a dual process model of moral reasoning. They found that "impersonal moral dilemmas" elicited utilitarian responses (judging personal moral violations to be acceptable when they serve a greater good), which were associated with brain activation patterns involving the abstract reasoning centers of the dorsolateral prefrontal cortex. By contrast, "personal moral dilemmas" elicited quite different responses that focused on fairness and appropriateness rather than "greatest good" considerations. These latter cases generally drew quick decisions and involved heightened brain activity in the emotion and social cognition areas (specifically, the medial prefrontal cortex, posterior cingulate/precuneus, and superior temporal sulcus/temporoparietal junction).

As exciting as some of these advances have been, we would be remiss not to point out that the value of brain-imaging studies to date are incremental rather than revolutionary. Indeed, many of the technical requirements of current fMRI and ERP studies directly contradict our advice on previous pages. For example, participants in fMRI studies are necessarily alone and confined within a rather large magnet. They cannot have meaningful and authentic social interactions, and are often acutely aware of the psychological variables that are being measured. Current technology limits these studies to judgment-type experiments, and typically they rely on wholly within-participant designs. At their best these studies can clarify the region of the brain involved in judgments; at their worst they merely identify the neural correlates of behavior without yielding new psychological insight regarding the behavior of humans. This debate is exacerbated by the tremendous fiscal cost associated with imaging studies. Critics point out that many of the published results could have been obtained with more traditional dependent measures and could have been obtained for a mere fraction of the cost. We have no doubt that much will be learned from social neuroscience, particularly as new technologies are developed (e.g., that permit brain scans during social interactions). But we hope that researchers and funding agencies will conduct cost-benefit analyses of the value of such studies, and not lose sight of the advantages of experimental studies that "merely" include self-report and behavioral measures.

The Postexperimental Follow-up

The experiment does not end when the data have been collected. Rather, the prudent experimenter will want to remain with the participants to talk and listen in order to accomplish three important goals: (a) to ensure that the participants are in a good and healthy frame of mind; (b) to be certain that the participants understand the experimental procedures, the hypotheses, and their own performance so that they gain a valuable educational experience as a result of having participated; (c) to avail themselves of the participant's unique skill as a valuable consultant in the research enterprise; that is, only the participants know for certain whether the instructions were clear, whether the independent variable had the intended impact on them, and so on; (d) to probe for any suspicion on the part of the participants, such as whether they believed the cover story.

It is impossible to overstate the importance of the postexperimental follow-up. The experimenter should never conduct it in a casual or cavalier manner. Rather, the experimenter should probe gently and sensitively to be certain that all of the above goals are accomplished. This is especially and most obviously true if any deception has been employed. In this case, the experimenter needs to learn if the deception was effective or if the participant was suspicious in a way that could invalidate the data based on his or her performance in the experiment. Even more important, where deception was used, the experimenter must reveal the true nature of the experiment and the reasons why deception was necessary. Again, this cannot be done lightly. People do not enjoy learning that they have behaved in a naive or gullible manner. The experimenter not only must be sensitive to the feelings and dignity of the participants but also should communicate this care and concern to them. We have found that people are most receptive to experimenters who are open in describing their own discomfort with the deceptive aspects of the procedure. Then, in explaining why the deception was necessary, the experimenter not only is sharing his or her dilemma as an earnest researcher (who is seeking the truth through the use of deception) but also is contributing to the participants'

educational experience by exploring the process as well as the content of social psychological experimentation.

Although it is important to provide people with a complete understanding of the experimental procedures, this is not the best way to begin the postexperimental session. In order to maximize the value of the participants as consultants, it is first necessary to explore with each the impact of the experimental events. The value of this sequence should be obvious. If we tell the participants what we expected to happen before finding out what the participants experienced, they may have a tendency to protect us from the realization that our procedures were pallid, misguided, or worthless. Moreover, if deception was used, the experimenter, before revealing the deception, should ascertain whether the participant was suspicious and whether particular suspicions were of such a nature as to invalidate the results.

This should not be done abruptly. It is best to explore the feelings and experiences of the participants in a gentle and gradual manner. Why the need for gradualness? Why not simply ask people if they suspected that they were the victims of a hoax? Subjects may not be responsive to an abrupt procedure for a variety of reasons. First, if a given person *did* see through the experiment, he or she may be reluctant to admit it out of a misplaced desire to be helpful to the experimenter. Second, as mentioned previously, since most of us do not feel good about appearing gullible, some participants may be reluctant to admit that they can be easily fooled. Consequently, if participants are told pointedly about the deception, they might imply that they suspected it all along, in order to save face. Thus, such an abrupt procedure may falsely inflate the number of suspicious participants and may, consequently, lead the experimenter to abandon a perfectly viable procedure. Moreover, as mentioned previously, abruptly telling people that they have been deceived is a harsh technique that can add unnecessarily to their discomfort and, therefore, should be avoided.

The best way to begin a postexperimental interview is to ask the participants if they have any questions. If they do not, the experimenter should ask if the entire experiment was perfectly clear—the purpose of the experiment as well as each aspect of the procedure. The participants should then be told that people react to things in different ways, and it would be helpful if they would comment on how the experiment affected them, why they responded as they did, and how they felt at the time, for example. Then each participant should be asked specifically whether there was any aspect of the procedure that he or she found odd, confusing, or disturbing.

By this time, if deception has been used and any participants have any suspicions, they are almost certain to have revealed them. Moreover, the experimenter should have discovered whether the participants misunderstood the instructions or whether any responded erroneously. If no suspicions have been voiced, the experimenter should continue: "Do you think there may have been more to the experiment than meets the eye?" This question is virtually a giveaway. Even if the participants had not previously suspected anything, some will probably begin to suspect that the experimenter was concealing something. In our experience, we have found that many people will take this opportunity to say that they did feel that the experiment, as described, appeared too simple (or something of that order). This is desirable; whether the participants were deeply suspicious or not, the question allows them an opportunity to indicate that they are not the kind of person who is easily fooled. The experimenter should then explore the nature of the suspicion and how it may have affected the participant's behavior. From the participant's answers to this question, the experimenter can make a judgment as to how close a participant's suspicions were to the actual purpose of the experiment and, consequently, whether the data are admissible. Obviously, the criteria for inclusion should be both rigorous and rigid and should be set down before the experiment begins; the decision should be made without knowledge of the participants' responses on the dependent variable.

The experimenter should then continue with the debriefing process by saying something like this: "You are on the right track, we *were* interested in exploring some issues that we didn't discuss with you in advance. One of our major concerns in this study is…" The experimenter should then describe the problem under investigation, specifying why it is important and explaining clearly exactly how the deception took place and why it was necessary. Again, experimenters should be generous in sharing their own discomfort with the participant. They should make absolutely certain that the participant fully understands these factors before the postexperimental session is terminated.

It is often useful to enlist the participant's aid in improving the experiment. Often the participant can provide valuable hints regarding where the weaknesses in the manipulation occurred and which one of these caused competing reactions to the one the experimenter intended. These interviews can and should, of course, be continued during the time the experiment is actually being run, but it is usually during pretesting that the most valuable information is obtained.

Finally, regardless of whether deception is used, the experimenter must attempt to convince the participants not to discuss the experiment with other people until it is completed. This is a serious problem because even a few sophisticated participants can invalidate an experiment. Moreover, it is not a simple matter to swear participants to secrecy; some have friends who may subsequently volunteer

for the experiment and who are almost certain to press them for information. Perhaps the best way to reduce such communication is to describe graphically the colossal waste of effort that would result from experimenting with people who have foreknowledge about the procedure or hypothesis of the experiment and who thus can rehearse their responses in advance. The experimenter should also explain the damage that can be done to the scientific enterprise by including data from such participants. It often helps to provide participants with some easy but unrevealing answers for their friends who ask about the study (e.g., "it was about social perception"). If we experimenters are sincere and honest in our dealings with the participants during the postexperimental session, we can be reasonably confident that few will break faith.

To check on the efficacy of this procedure, Aronson (1966) enlisted the aid of three undergraduates who each approached three acquaintances who had recently participated in one of his experiments. The confederates explained that they had signed up for that experiment, had noticed the friend's name on the sign-up sheet, and wondered what the experiment was all about. The experimenter had previously assured these confederates that their friends would remain anonymous. The results were encouraging. In spite of considerable urging and cajoling on the part of the confederates, none of the former participants revealed the true purpose of the experiment; two of them went as far as providing the confederates with a replay of the cover story, but nothing else.

What if the participant *has* been forewarned before entering the experimental room? That is, suppose a participant does find out about the experiment from a friend who participated previously. Chances are the participant will not volunteer this information to the experimenter before the experiment. Once again, we as experimenters must appeal to the cooperativeness of the participant, emphasizing how much the experiment will be compromised if people knew about it in advance. We cannot overemphasize the importance of this procedure as a safeguard against the artifactual confirmation of an erroneous hypothesis because of the misplaced cooperativeness of the participant. If the participants are indeed cooperative, they will undoubtedly cooperate with the experimenter in this regard also and will respond to a direct plea of the sort described.

Ethical Concerns in Experimentation

Experimental social psychologists have been deeply concerned about the ethics of experimentation for a great many years precisely because our field is constructed on an ethical dilemma. Basically, the dilemma is formed by a conflict between two sets of values to which most social psychologists subscribe: a belief in the value of free scientific inquiry and a belief in the dignity of humans and their right to privacy. We will not dwell on the historical antecedents of these values or on the philosophical intricacies of the ethical dilemma posed by the conflict of these values. It suffices to say that the dilemma is a real one and cannot be dismissed either by making pious statements about the importance of not violating a person's feelings of dignity or by glibly pledging allegiance to the cause of science. It is a problem every social psychologist must face squarely, not just once, but each time he or she constructs and conducts an experiment, since it is impossible to delineate a specific set of rules and regulations governing all experiments. In each instance the researcher must decide on a course of action after giving careful consideration to the importance of the experiment and the extent of the potential injury to the dignity of the participants.

Obviously, some experimental techniques present more problems than others. In general, experiments that employ deception cause concern because of the fact that lying, in and of itself, is problematic. Similarly, procedures that cause pain, embarrassment, guilt, or other intense feelings present obvious ethical problems. In addition, any procedure that enables the participants to confront some aspect of themselves that may not be pleasant or positive is of deep ethical concern. For example, many of Asch's (1951) participants learned that they could conform in the face of implicit group pressure; many of Aronson and Mettee's (1968) participants learned that they would cheat at a game of cards; and many of Milgram's (1974) participants learned that they could be pressured to obey an authority even when such obedience involved (apparently) inflicting severe pain on another human being. Even more imposing are the findings of the Stanford prison study in which college students learned that, even in the absence of direct explicit commands, they would behave cruelly and even sadistically toward fellow students (Haney, Banks, & Zimbardo, 1973).

It can be argued that such procedures are therapeutic or educational for the participants. Indeed, many of the participants in these experiments have made this point. But this does not, in and of itself, justify the procedure primarily because the experimenter could not possibly know in advance that it would be therapeutic for all participants. Moreover, it is arrogant for the scientist to decide that he or she will provide people with a therapeutic experience without their explicit permission.

The use of deception, when combined with the possibility of "self-discovery," presents the experimenter with a special kind of ethical problem. In a deception experiment it is impossible, by definition, to attain informed consent from the participants in advance of the experiment. For example, how could Milgram or Asch have attained informed consent from their participants without revealing

aspects of the procedure that would have invalidated any results they obtained? An experimenter cannot even reveal in advance that the purpose of an experiment is the study of conformity or obedience without influencing the participant to behave in ways that are no longer "pure." Moreover, we doubt that the experimenter can reveal that deception might be used without triggering vigilance and, therefore, adulterating the participant's response to the independent variable.

A number of guidelines have been developed to protect the welfare of research participants. In 1973 the American Psychological Association (APA) published a set of guidelines for the conduct of research involving human participants, which have since been revised and updated a number of times. It behooves all investigators to read these guidelines carefully before conducting research (American Psychological Association, 2002). Further, as stated in the guidelines, ethical decisions should not be made alone. Researchers may not always be in the best position to judge whether their procedures are ethically permissible. Because of this fact, all research using human subjects that is funded by the federal government, or conducted at colleges and universities, must receive approval from an Institutional Review Board (IRB). This is a panel of scientists and nonscientists who judge whether the risks to participants outweigh the potential gains of the research. It is not uncommon for an IRB to ask researchers to revise their procedures to minimize risks to participants.

It is worth noting that there have been some empirical investigations of the impact of deception experiments on participants. These studies have generally found that people do not object to the kinds of mild discomfort and deceptions typically used in social psychological research (e.g., Christensen, 1988; Sharpe, Adair, & Roese, 1992; Smith & Richardson, 1983). If mild deception is used, and time is spent after the study discussing the deception with participants and explaining why it was necessary, the evidence is that people will not be adversely affected. Nonetheless, the decision as to whether to use deception in a study should not be taken lightly, and alternative procedures should always be considered.

As we noted in our discussion of the postexperimental interview, it is critical to explain to participants in a deception study, at its conclusion, the true nature of the procedures and the reasons for them. We strongly recommend, however, that a thorough explanation of the experiment be provided regardless of whether deception or stressful procedures are involved. The major reason for this recommendation is that we cannot always predict the impact of a procedure; occasionally, even procedures that appear to be completely benign can have a powerful impact on some participants. An interesting example of such an unexpectedly powerful negative impact comes from a series of experiments on social dilemmas by Dawes and his students (Dawes, McTavish, & Shaklee, 1977). In these experiments, typically, the participant must make a decision between cooperating with several other people or "defecting." The contingencies are such that if all participants choose to cooperate, they all profit financially; however, if one or more defect, defection has a high payoff and cooperation produces little payoff. Each person's response is anonymous and remains so. The nature of the decision and its consequences is fully explained to the participants at the outset of the experiment. No deception is involved.

Twenty-four hours after one experimental session, an elderly man (who had been the sole defector in his group and had won $19) telephoned the experimenter trying to return his winnings so that it could be divided among the other participants (who, because they chose to cooperate, had each earned only $1). In the course of the conversation, he revealed that he felt miserable about his greedy behavior and that he had not slept all night. After a similar experiment, a woman who had cooperated while others defected revealed that she felt terribly gullible and had learned that people were not as trustworthy as she had thought. In order to alleviate this kind of stress, Dawes went on to develop an elaborate and sensitive follow-up procedure.

We repeat that these experiments were selected for discussion precisely because their important and powerful impact could not have been easily anticipated. We are intentionally not focusing on experiments that present clear and obvious problems like the well-known obedience study (Milgram, 1974), or the Stanford prison study (Haney et al., 1973). We have purposely selected an experiment that involves no deception and is well within the bounds of ethical codes. Our point is simple but important. No code of ethics can anticipate all problems, especially those created through participants discovering something unpleasant about themselves or others in the course of an experiment. However, we believe a sensitive postexperimental interview conducted by a sincere and caring experimenter not only instructs and informs, but also provides important insights and helps reduce feelings of guilt or discomfort generated by such self-discovery (see Holmes, 1976a, 1976b; Ross, Lepper, & Hubbard, 1975).

CONCLUDING COMMENTS

We hope that this chapter has helped explain why laboratory experiments are often the method of choice for social psychologists and has provided useful tips about how to conduct experiments. We want to emphasize, however, that social psychology cannot live by lab experimentation alone and that we must use multiple methods if we are to

advance theory and find solutions to important problems. Basic, process-oriented experimental research may isolate important causal processes, but convincing demonstrations that those processes operate in applied settings are essential before theory can be converted into practice.

The research literature on self-affirmation and stereotype threat provides a particularly good example of how a synthesis between field and laboratory experiments can work at its best. Research in these areas began with laboratory experiments conducted with college student participants, showing that (a) people can deal with threats in one domain by affirming themselves in another (Steele, 1988), and that (b) targets of prejudice perform poorly under conditions of stereotype threat, in which they are concerned that their performance will confirm a negative stereotype of their group (Steele & Aronson, 1995). Based on these ideas, Cohen, Garcia, Apfel, and Master (2006) developed an intervention to improve the academic performance of African American middle school students. Some students were randomly assigned to a self-affirmation condition in which they chose values that were important to them and wrote about these values for 15 minutes. Students in the control condition wrote about why the values might be important for someone else. This simple intervention had remarkable effects: Thoughts about race became less accessible to the students in the self-affirmation condition and the students achieved higher grades during the remainder of the academic term. We dare say that the idea that such a "miniscule" intervention could have such dramatic effects would never have occurred to researchers without the prior laboratory experiments on self-affirmation and stereotype threat.

Another good example of the creative interplay between laboratory and field experimentation is the work of Aronson and his colleagues on the effects of cooperative learning (Aronson & Bridgeman, 1979; Aronson & Osherow, 1980; Aronson, Stephan, Sikes, Blaney, & Snapp, 1978). The research began as an experimental intervention in response to a crisis in the Austin (Texas) school system following its desegregation. Aronson and his colleagues observed the dynamics of the classroom and diagnosed that a major cause of the existing tension was the competitive atmosphere that exacerbated the usual problems brought about by desegregation. They then changed the atmosphere of existing classrooms by restructuring the learning environment so that some students were teaching one another in small, interdependent "jigsaw" groups, while others continued to study in more traditional classrooms.

The results of this and subsequent field experiments showed that the cooperative classroom atmosphere decreased negative stereotyping, increased cross-ethnic liking, increased self-esteem, improved classroom performance, and increased empathic role taking. At the same time, Aronson and his colleagues were able to enhance their understanding of the underlying dynamic of this cooperative behavior by closer scrutiny under controlled laboratory conditions. For example, in one such laboratory experiment, they showed that, in a competitive situation, individuals make situational self-attributions for failure and dispositional self-attributions for success, while making the reverse attributions to their opponent. However, in a cooperative structure, individuals gave their partners the same benefit of the doubt that they gave to themselves, that is, dispositional attributions for success and situational attributions for failure (Stephan, Presser, Kennedy, & Aronson, 1978).

Field experimentation in applied settings often provides an opportunity for impact and involvement of research participants that vastly exceeds any ever achieved in the laboratory. However, the focus of such research also tends to be more limited than the general tests of theory underlying most laboratory research efforts, because they are forced to deal only with variables found in the particular applied context under study. If the distinctive contribution of experimental social psychology to the general body of knowledge is ever to be realized, an optimal integration of theory-oriented laboratory research with applied field experimentation will be required.

At present we are concerned because the alternative research modes in social psychology seem, for the most part, to be functioning in isolation from each other. What is needed now is a new attempt at synthesis; that is, to construct a more limited (and perhaps closer to the original) version of the Lewinian model of the interplay between laboratory and field research. Such a synthesis will require a concern with discovering more specifiable relationships rather than with attempts to find sweeping general theories of human social behavior. It will require an emphasis on assessing the relative importance of several variables, which all influence an aspect of multiply-determined behavior, rather than on testing to see if a particular variable has a "significant" impact. And it will require a sensitivity to the interaction between research design and research setting and the benefits of multiple methodologies. We have great faith in our fellow social psychologists' ability to meet these challenges. Indeed, many are already deeply immersed in research programs that are increasing our understanding of basic social psychological processes and having an impact on real world problems. We hope this chapter inspires a new generation of social psychologists to do the same.

REFERENCES

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct.* Retrieved July 30, 2009 from: http://www.apa.org/ethics/.

Andrews, F. M., & Robinson, J. P. (1991). Measures of subjective well-being. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (Vol. 1, pp. 61–114). San Diego, CA: Academic Press.

Aronson, E. (1966). Avoidance of inter-subject communication. *Psychological Reports, 19,* 238.

Aronson, E., & Bridgeman, D. (1979). Jigsaw groups and the desegregated classroom: in pursuit of common goals. *Personality and Social Psychology Bulletin, 5,* 438–446.

Aronson, E., & Carlsmith, J. M. (1963). Effect of the severity of threat on the devaluation of forbidden behavior. *Journal of Abnormal and Social Psychology, 66,* 583–588.

Aronson, E., & Carlsmith, J. M. (1968). Experimentation in social psychology. In G. Lindzey and E. Aronson (Eds.), *The handbook of social psychology* (Vol. 2, pp. 1–79). Reading, MA: Addison-Wesley.

Aronson, E., Fried, C. B., & Stone, J. (1991). Overcoming denial and increasing the intention to use condoms through the induction of hypocrisy. *American Journal of Public Health, 81,* 1636–1637.

Aronson, E., & Mettee, D. (1968). Dishonest behavior as a function of differential levels of induced self-esteem. *Journal of Personality and Social Psychology, 9,* 121–127.

Aronson, E., & Mills, J. (1959). The effect of severity of initiation on liking for a group. *Journal of Abnormal and Social Psychology, 59,* 177–181.

Aronson, E., & Osherow, N. (1980). Cooperation, prosocial behavior, and academic performance: experiments in the desegregated classroom. *Applied Social Psychology Annual, 1, 163*–196.

Aronson, E., Stephan, C., Sikes, J., Blaney, N., & Snapp, M. (1978). *The jigsaw classroom.* Beverly Hills, CA: Sage Publications.

Aronson, E., Willerman, B., & Floyd, J. (1966). The effect of a pratfall on increasing interpersonal attractiveness. *Psychonomic Science, 4,* 227–228.

Aronson, E., Wilson, T. D., & Akert, R. M. (1994). *Social psychology: The heart and the mind.* New York: HarperCollins.

Asch, S. (1951). Effects of group pressure upon the modification and distortion of judgment. In H. Guetzkow (Ed.), *Groups, leadership, and men* (pp. 177–190). Pittsburgh: Carnegie Press.

Bargh, J. A. (1989). Conditional automaticity: Varieties of automatic influence in social perception and cognition. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 3–51). New York: Guilford.

Bargh, J. A. (1990). Auto-motives: Preconscious determinants of social interaction. In E. T. Higgins & R. M. Sorrentino (Eds.), *Handbook of motivation and cognition* (Vol. 2, pp. 93–130). New York: Guilford.

Baumeister, R. F. (1982). A self-presentational view of social phenomena. *Psychological Bulletin, 91,* 3–26.

Baumeister, R., Vohs, K., & Funder, D. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior?. *Perspectives on Psychological Science, 2,* 396–403.

Blascovich, J., Loomis, J., Beall, A., Swinth, K., Hoyt, C., & Bailenson, J. (2002). Immersive virtual environment technology as a research tool for social psychology. *Psychological Inquiry, 13,* 103–125.

Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull & R. S. Wyer, Jr. (Eds.), *Advances in social cognition* (Vol. 1, pp. 1–36). Hillsdale, NJ: Erlbaum.

Campbell, D. T. (1957). Factors relevant to validity of experiments in social settings. *Psychology Bulletin, 54,* 297–312.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand McNally.

Carlsmith, K. M., Wilson, T. D., & Gilbert, D. T (2008). The paradoxical consequences of revenge. *Journal of Personality and Social Psychology, 95,* 1316–1324.

Chaiken, S. (1987). The heuristic model of persuasion. In M. P. Zanna, J. M. Olson, & C. P. Herman (Eds.), *Social influence: The Ontario Symposium* (Vol. 5, pp. 3–39). Hillsdale, NJ: Erlbaum.

Christensen, L. (1988). Deception in psychological research: When is its use justified? *Personality and Social Psychology Bulletin, 14,* 664–675.

Cohen, D. (1977). *Psychologists on psychology.* New York: Taplinger.

Cohen, G., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: a social-psychological intervention. *Science, 313,* 1307–1310.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experiments: design and analysis issues for field settings.* Shokie, IL: Rand McNally.

Dawes, R. B., McTavish, J., & Shaklee, H. (1977). Behavior, communication, and assumptions about other people's behavior in a common dilemmas situation. *Journal of Personality and Social Psychology, 35,* 1–11.

de la Haye, A. (1991). Problems and procedures: A typology of paradigms in interpersonal cognition. *Cahiers de Psychologie Cognitive, 11,* 279–304.

de Houwer, J. (in press). What are implicit measures and why are we using them? In R. Wiers & A. Stacy (Eds.), *Handbook of implicit cognition and addiction.* New York: Sage.

DePaulo, B. M. (1992). Nonverbal behavior and self-presentation. *Psychological Bulletin, 111,* 203–243.

DePaulo, B. M., Lanier, K., & Davis, T. (1983). Detecting the deceit of the motivated liar. *JourÍnal of Personality and Social Psychology, 45,* 1096–1103.

Diener, E. (1994). Assessing subjective well-being: Progress and opportunities. *Social Indicators Research, 31,* 103–157.

Diener, E., & Biswas-Diener, R. (2008). *The science of optimal happiness.* Boston: Blackwell Publishing.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 75–109). San Diego: Academic Press.

Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford, CA: Stanford University Press.

Fordyce, M. W. (1988). A review of research on the happiness measures: A sixty second index of happiness and mental health. *Social Indicators Research, 20,* 355–381.

Gerard, H. B., & Mathewson, G. C. (1966). The effects of severity of initiation on liking for a group: A replication. *Journal of Experimental Social Psychology, 2,* 278–287.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist, 46,* 107–119.

Gilbert, D. T. (2006). *Stumbling on happiness.* New York; Knopf.

Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology, 60,* 509–517.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44,* 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293,* 2105–2108.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4–27.

Haney, C., Banks, C., & Zimbardo, P. (1973). Interpersonal dynamics in a simulated prison. *International Journal of Criminology and Penology, 1,* 69–97.

Harackiewicz, J. M., Manderlink, G., & Sansone, C. (1984). Rewarding pinball wizardry: Effects of evaluation and cue value on intrinsic interest. *Journal of Personality and Social Psychology, 47,* 287–300.

Harlow, H. F., & Zimmerman, R. R. (1958). The development of affectional responses in infant monkeys. *Proceedings of the American Philosophic Society, 102,* 501–509.

Higgins, E. T., Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology, 13,* 141–154.

Holmes, D. S. (1976a). Debriefing after psychological experiments: I. Effectiveness of postdeception dehoaxing. *American Psychologist, 31,* 858–867.

Holmes, D. S. (1976b). Debriefing after psychological experiments: II. Effectiveness of postexperimental desensitizing. *American Psychologist, 31,* 868–875.

Langer, E. J. (1989). Minding matters: The consequences of mindlessness-mindfulness. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 22, pp. 137–174). San Diego, CA: Academic Press.

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the overjustification hypothesis. *Journal of Personality and Social Psychology, 28,* 129–137.

Milgram, S. (1974). *Obedience to authority: An experimental view.* New York: Harper & Row.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38,* 379–388.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review (pp. 265–292). In J. A. Bargh (Ed.), *Automatic processes in social thinking and behavior.* New York: Psychology Press.

Orne, M. (1962). On the social psychology of the psychological experiment. *American Psychologist, 17,* 776–783.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change.* New York: Springer-Verlag.

Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology, 41,* 847–855.

Reis, H. T. (1982). An introduction to the use of structural equations: Prospects and problems. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 3, pp. 255–287). Beverly Hills, CA: Sage.

Rosenthal, R. (1994). Interpersonal expectancy effects: A 30-year perspective. *Current Directions in Psychological Science, 3,* 176–179.

Rosenthal, R., & Lawson, R. (1964). A longitudinal study of the effects of experimenter bias on the operant learning of laboratory rats. *Journal of Psychiatric Research, 2,* 61–72.

Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology, 32,* 880–892.

Schachter, S. (1959). *The psychology of affiliation: experimental studies of the sources of gregariousness.* Stanford, CA: Stanford University Press.

Schachter, S., & Singer, J. E. (1962). Cognitive, social and physiological determinants of emotional state. *Psychological Review, 69,* 379–399.

Schlenker, B. R. (1980). *Impression management.* Monterey, CA: Brooks-Cole.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51,* 515–530.

Sharpe, D., Adair, J. G., & Roese, N. J. (1992). Twenty years of deception research: A decline in subjects' trust? *Personality and Social Psychology Bulletin, 18,* 585–590.

Sigall, H., Aronson, E., & Van Hoose, T. (1970). The cooperative subject myth or reality? *Journal of Experimental Social Psychology, 6,* 1–10.

Smith, S. S., & Richardson, D. (1983). Amelioration of deception and harm in psychological research: The important role of debriefing. *Journal of Personality and Social Psychology, 44,* 1075–1082.

Steele, C. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (Ed.), *Advances in experimental social psychology: Vol. 21. Social psychological studies of the self: Perspectives and programs* (pp. 261–302). San Diego, CA: Academic Press.

Steele, C., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797–811.

Stephan, C., Presser, N. R., Kennedy, J. C., & Aronson, E. (1978). Attributions to success and failure after cooperative or competitive interaction. *European Journal of Social Psychology, 8,* 269–274.

Uleman, J. S. (1989). A framework for thinking intentionally about unintended thoughts. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 425–449). New York: Guilford.

Walster, E., Aronson, E., & Abrahams, D. (1966). On increasing the persuasiveness of a low prestige communicator. *Journal of Experimental Social Psychology, 2,* 325–342.

Wegner, D. M. (1994). Ironic processes of mental control. *Psychological Review 101,* 34–52.

Wicker, A. W. (1969). Attitudes versus actions: The relationship between verbal and overt behavioral responses to attitude objects. *Journal of Social Issues, 25,* 41–78.

Williams, K. D., Cheung, C. K. T., & Choi, W. (2000). CyberOstracism: Effects of being ignored over the Internet. *Journal of Personality and Social Psychology, 79,* 748–762.

Williams, K. D., & Sommer, K. L. (1997). Social ostracism by one's coworkers: Does rejection lead to loafing or compensation? *Personality and Social Psychology Bulletin, 23,* 693–706.

Wilson, T. D. (2002). *Strangers to ourselves: Discovering the adaptive unconscious.* Cambridge, MA: Harvard University Press.

Wilson, T. D., Dunn, D. S., Kraft, D., & Lisle, D. J. (1989). Introspection, attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123–205). Orlando, FL: Academic Press.

Wilson, T. D., & Gilbert, D. T. (2003). Affective forecasting. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 35, pp. 345–411). San Diego, CA: Academic Press.

Wilson, T. D., Hodges, S. D., & LaFleur, S. J. (1995). Effects of introspecting about reasons: Inferring attitudes from accessible thoughts. *Journal of Personality and Social Psychology, 69,* 16–28.

Wilson, T. D., & Lassiter, D. (1982). Increasing intrinsic interest with the use of superfluous extrinsic constraints. *Journal of Personality and Social Psychology, 42,* 811–819.

Wilson, T. D., Lisle, D., Schooler, J., Hodges, S. D., Klaaren, K. J., & LaFleur, S. J. (1993). Introspecting about reasons can reduce post-choice satisfaction. *Personality and Social Psychology Bulletin, 19*, 331–339.

Zanna, M. P., & Fazio, R. H. (1982). The attitude-behavior relation: Moving toward a third generation of research. In M. P. Zanna, E. T. Higgins, & C. P. Herman (Eds.), *Consistency in social behavior: The Ontario Symposium* (Vol. 2, pp. 283–301). Hillsdale, NJ: Erlbaum.