EXPERIMENTATION IN SOCIAL PSYCHOLOGY

ELLIOT ARONSON, *University of California, Santa Cruz*
TIMOTHY D. WILSON, *University of Virginia*
MARILYNN B. BREWER, *The Ohio State University*

PROLOGUE

This chapter is aimed primarily at helping the reader to learn to think like an experimental social psychologist. It is both a blessing and (in a sense) a curse to think like an experimental social psychologist. Let us give you one of many possible examples of what we mean by that statement. While we were working on this chapter, we happened to pick up a copy of *The New Yorker* magazine where we read an excellent, highly informative essay by James Kunen about college-level educational programs in our prisons. Kunen (1995) wrote passionately about the effectiveness of these programs and how, in an amendment to the crime bill, a generally punitive, "penny-wise/pound foolish" congressional majority was eliminating these programs after characterizing them as wasteful, and as tending to coddle criminals. The essay contains a few vivid case histories of prisoners who completed a college program and went on to lead productive lives. Any systematic data? You bet. Kunen reports one study released in 1991 by the New York State Department of Correctional Services, which "found that male inmates who completed one or more years of higher education in prison had a recidivism rate, four years after their release, more than twenty percent lower than the average for all male inmates" (Kunen, 1995, p. 36).

The liberal/humanist in us wanted to get excited by the results of this study; it would be terrific to have convincing data proving that prison education really pays off. But alas, the experimental social psychologist in us was far more skeptical than Mr. Kunen. Yes, it *would* be wonderful to have convincing data on this issue, but, unfortunately, these data fall far short of the mark. We must raise at least one vital methodological question: Are the prisoners who were "assigned" to the control condition the same kind of people as those "assigned" to the experimental condition? That is, might it not be the case that the prisoners who signed up for the course of study and completed a year of it were different *to begin with* (say, in motivation, ability, intelligence, prior education, mental health, or what have you) from those who did not sign up for the course—or those who signed up but dropped out early? As you will see, this is not simply nit-picking; if they were different from the general run of prisoners, then it is likely (or, at least, possible) that they would have had a lower rate of recidivism even without having taken the course of study. If so, then it wasn't the prison courses that caused the lower recidivism.

The *curse* of thinking like an experimental social psychologist is that it keeps us from rejoicing. Part of the

This essay is a major revision of a chapter by Elliot Aronson and J. Merrill Carlsmith which first appeared in the 1968 edition of Lindzey and Aronson's Handbook of Social Psychology. *To the best of our knowledge that was the first time an experimental social psychologist attempted a formal and thorough presentation of the so-called tacit knowledge of the experimental method. By "tacit knowledge" we refer to the countless, mundane—but vitally important—details involved in designing and conducting an experiment in this field. Much of this knowledge is considered to be intuitive in that it is difficult to articulate because it is something that experienced experimentalists simply "know"; i.e., this knowledge is almost invariably gleaned from day-to-day experience in the laboratory rather than the classroom.*

It is with a deep sense of gratitude that we acknowledge our indebtedness to Merrill Carlsmith (a very intuitive experimentalist) for the important contributions he made to the original chapter as well as to our current thinking about methodology. His premature death in 1984, at the age of 48, was an irreparable loss to his friends and to the discipline.

blessing is that it keeps us from rejoicing—over potentially meaningless data like those described above. Moreover, as you shall see over and over again throughout this chapter, another part of the blessing is that experimental social psychologists are able to use their knowledge and skill to perform the appropriate research to test hypotheses like these in a solid and more convincing manner. For example, there are some simple but elegant solutions to the methodological shortcomings of the prison data, solutions we will discuss in this chapter.

INTRODUCTION

There are a great many ways of gathering information about social behavior. We can simply observe people; we can interview them about their attitudes, beliefs, intentions, and motivation; we can ask them to fill out questionnaires and rating scales. The list of techniques and variations on these techniques is filled with interesting possibilities. These techniques have provided us with some of our richest and most fascinating data about social phenomena, as described in other chapters in this *Handbook*. In this chapter we hope to convey something about the approach that has been the workhorse of social psychological research, the experimental method. With this method the researcher randomly assigns people to different conditions and ensures that these conditions are identical except for the independent variable (the one believed to have a causal effect on people's responses). Although there is merit in all methods of investigating social behavior, the experiment has been the chief method of choice for social psychology.

We have two main missions in this chapter. First, it is important to discuss why the experiment is 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 experimental method 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 much more difficult to do experiments, such as the study of close relationships and culture. Another reason is that promising new 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 these new topics and techniques. In the first part of the chapter we will discuss the advantages of experiments in general terms, and then return to these issues at the end of the chapter in a discussion of validity and realism in experiments.

The middle 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 middle part, we hope to provide detailed instructions in "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 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 EXPERIMENTS?

Let's begin with an example of a research problem and a discussion of different ways this problem could be addressed empirically. For this purpose, we will use a problem that is dear to the hearts of most social psychologists, including ourselves: prejudice and stereotyping. Perhaps no other social problem has captured the attention of social psychologists as much as this one, from early research by Allport (1954) and Sherif et al. (1961) to modern research on in-group favoritism and the cognitive bases of stereotyping (Abrams & Hogg, 1990; Brewer & Brown, 1998, in this *Handbook*; Fiske et al., 1998, in this *Handbook*; Hamilton & Trolier, 1986; Tajfel, Billig, Bundy, & Flament, 1971; von Hippel, Sekaquaptewa, & Vargas, 1995). One reason for this fascination is that few problems are as prevalent and seemingly ingrained in human nature as discord between social groups. Centuries-old conflicts persist in many parts of the world, such as disputes between Israelis and Arabs, Irish Protestants and Irish Catholics, and Serbians and Croatians. In the United States racism persists decades after the Civil Rights movement of the 1950s and 1960s.

The causes of prejudice and stereotyping have been debated at great length by philosophers, social scientists, politicians, and pundits of all kinds. Social psychology offers a unique perspective by studying prejudice experimentally. Indeed, what sets social psychology apart from most other disciplines is the claim that it can discover the causes of human behavior by conducting scientific research. What sort of research might be done to understand the causes of prejudice?

We often ask our students this question at the beginning of our courses. We find a definite preference for certain types of studies, and these preferences, we believe, are highly instructive, because they reveal a good deal about most people's understanding (or lack thereof) of the experimental method. Consider this response, which is pretty typical of what our students suggest:

The best way to study prejudice, it seems to me, is to go to places where it is most likely to be seen and study what happens. It would be interesting, for example, to hang out at a car dealership and watch how the salespeople treat white versus African American customers. Or maybe you could go to the rental office of an apartment complex in a white neighborhood, and see what happens when an African American comes in and inquires about an apartment. If we see any signs of prejudice we could interview people to see why they acted the way they did.

As we tell our students, valuable insights can be gained by such careful observations of everyday behavior. One might even take this simple observational study a half step further and compare the experience of African Americans with the experience of Caucasians in a variety of situations. For example, the television program *Primetime Live* conducted a study much like the one described above and uncovered some disturbing examples of prejudicial behavior (Lucasiewicz & Sawyer, 1991). Two college-aged middle-class males—one African American, one white—were filmed with a hidden camera as they encountered several everyday situations, including the attempt to buy a car and rent an apartment. In viewing the show, it was clear that the differences in the way the two men were treated were striking. The white man, for example, was offered a better price on the same car by the same salesman than was the African American.

It does not take our students long, however, to recognize that this type of study has many limitations. When journalists report the results of their observations it is difficult to gauge the typicality of what they report. In the *Primetime Live* segment, for example, the reporter (Diane Sawyer) casually mentioned that the two men were not *always* treated differently, but the viewer has no way of determining how often this was the case. The examples of blatant prejudice were undoubtedly more "newsworthy," and thus more likely to be broadcast. All the segments showing the African Americans being treated fairly were left on the cutting-room floor, so to speak. In addition, there is no way of knowing to what extent the different treatment the men received was due to the fact that one was African American and one was white. Undoubtedly the producers attempted to make sure that the two men were similar in some respects; the men dressed similarly, were of similar age, and so on. Who is to say, though, whether other differences in appearance, demeanor, facial expression, or personality contributed to the way they were treated? It is highly likely that race played a role—and perhaps even a major role—but the point is that it is not possible to disentangle the effects of race with other uncontrolled factors.

If social scientists were conducting the study, instead of journalists, these problems would be easy enough to correct. Nothing would have been left on the cutting-room floor; scientists would report precisely how often prejudice was observed and how often it wasn't, rather than presenting only the newsworthy cases. A larger sample of interactions would be observed between many different kinds of people, to make sure that the differences were not skewed by the specific characteristics of one or two individuals. Even so, there would be some drawbacks to such an observational study. Chief among them is the difficulty of discovering the *causes* of prejudice. We might gain insights as to how often prejudiced behaviors occur, and obtain clues to some of its causes by observing when it occurs and when it does not. We could not, however, make any definitive statements about causality.

One solution to this problem would be to question the people we observe. In the *Primetime Live* segment filmed at a car dealership, for example, Diane Sawyer asked the car salesman, who had just been observed treating the African American man more negatively than the white man, to explain his actions. As you can imagine, the salesman did not say, "Well, Ms. Sawyer, let me explain to you the deep-seated causes of my prejudice. It all began when I was three and my father wouldn't let me play with an African American kid down the street. . . ." Instead he was quite defensive and refused to admit that he was biased in any way. Well, you might say, this is because he knew his answers would be broadcast on national television. And you might be right. As we will see, however, there are other deeper reasons to be wary of the answers people give about the causes of their behavior. They often do not want to tell scientists the real causes, out of embarrassment or defensiveness. More fundamentally, they might not even know the causes of their own behavior.

Observational versus Correlational versus Experimental Studies

The observational method is one in which naturalistic behavior is systematically observed and recorded. This was the method suggested by our student, in which people's behavior toward African Americans would be observed in natural settings. As we have noted, this method is often valuable for generating hypotheses about the causes of social behavior, but it is a poor technique for testing causal hypotheses.

As a step closer to understanding the causes of a phenomenon, we could try to uncover variables that predict its occurrence. Questions about prediction are often addressed with the *correlational method,* in which two or more variables are systematically measured, and the relationship between these variables is assessed. If there is a correlation between the variables, then we can predict one from the other (within a given margin of error). The *Primetime Live* segment was more of a correlational study, in that the one variable (the race of the college student) was correlated

with another (how the student was treated by people at car dealerships, rental offices, etc.). Correlational studies go beyond observational studies by making systematic observations of at least two variables and correlating these variables with each other, allowing the researcher to estimate the extent to which people's standing on one variable predicts their standing on the other variable. For example, Tumin, Barton, and Burrus (1958) assessed people's level of education and amount of prejudice toward African Americans and found a negative correlation: the more education people had, the lower their prejudice.

Correlational designs are still inadequate, however, in specifying cause and effect. This is nothing more than an elaboration of the old phrase "correlation does not prove causation." Whenever we observe that Variable X is correlated with Variable Y, there are three possible causal relationships: X could be causing Y, Y could be causing X, or some third variable could be causing both X and Y. Consider the correlation between lack of education and prejudice. It could be that there is something about getting an education that causes a reduction in prejudice. Although it may seem implausible, it is equally possible that there is something about prejudice that causes a lack of education (perhaps prejudiced people find it difficult to work with others and this impedes their educational progress). It is also possible that there is absolutely no causal link between education and prejudice. Some third variable, such as intelligence or social class, might cause people both to get more education and to not be prejudiced.

There are many examples of people making unwarranted causal inferences from correlational data with potentially serious consequences. Consider the results of a study in which the type of birth control women used was correlated with whether they had sexually transmitted diseases (STDs). Surprisingly, the researchers found that women whose partners used condoms had more STDs than women who used diaphragms or contraceptive sponges (Rosenberg et al., 1992). This finding was reported widely in the popular press, often with the conclusion that diaphragms and sponges prevented STDs better than did condoms. This conclusion, of course, makes a causal assumption from correlational data—the assumption that the types of birth control (Variable X) had a causal effect on STDs (Variable Y). The study, however, showed nothing of the kind. Perhaps women who had STDs (Variable Y) were more likely to insist that their partners use condoms (Variable X). It is equally possible that some third variable contributed both to the type of contraception women used and their likelihood of getting STDs. The women who relied on condoms might have differed from the other women in any number of ways that contributed to their getting more STDs; perhaps their sexual partners were more likely to be infected, perhaps they had sex more often, perhaps they had more partners (in fact, these women reported having had sex with more partners in the previous month than did the women who used diaphragms and sponges). The conclusion that condoms protect people less is completely unfounded and could lead to dangerous behavior on the part of women and men who draw this conclusion from the correlational design.

The great advantage of the experimental method is that the causal relationship between variables can be determined with much greater certainty. This is done in two ways: by controlling all factors except the independent variable and by randomly assigning people to condition. We will elaborate on how and why this is done shortly. For now, consider the following laboratory experiment on stereotyping: Gilbert and Hixon (1991) were interested in the conditions under which a stereotype about a social group is activated when a member of that group is encountered. Some theories assume that stereotypes are activated automatically, without any conscious intent on the part of the perceiver. According to this view, when we encounter a member of a stereotyped group our stereotype comes to mind, even if we don't want it to. In contrast, Gilbert and Hixon (1991) argued that, at least under some circumstances, it takes cognitive effort to bring a stereotype to mind. If people are distracted or preoccupied, the stereotype will not be activated.

Gilbert and Hixon (1991) tested this hypothesis as follows: white college students were asked to watch 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 fifteen seconds, such as POST or PAST. 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,* which is a variable that the researcher varies to see if it has an effect on some other variable of interest (the *dependent variable*). The other independent variable in this study 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. Gilbert and Hixon (1991) predicted that people who had to remember the eight-digit number would not have the cognitive resources to activate their stereotype of Asians, and thus should judge the Asian woman no differently than the Caucasian women. Not busy participants, however, should have the resources to call to mind their stereotype, and thus should judge the Asian woman differently than the Caucasian woman.

The way in which the activation of stereotypes was measured was as follows: 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 (as established

in pretesting). 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: 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. In a second study, Gilbert and Hixon (1991) 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).

The Gilbert and Hixon (1991) experiment differs in many ways from the other studies of prejudice we have considered (the observational study proposed by our student, the *Primetime Live* "study," and the correlational study by Tumin et al. (1958) on level of education and amount of prejudice). We suspect that some readers will find it unsatisfactory: How can stereotyping be studied in the confines of the laboratory, on such artificial tasks as asking people to complete word fragments? What can be learned about prejudice in a study in which people never actually interacted with anyone but only watched videotapes?

There *are* drawbacks to laboratory experiments that we will consider at some length. For now we point to the main advantage of experiments: their ability to determine causality. To illustrate this point, imagine the following hypothetical study that tested Gilbert and Hixon's hypotheses about stereotype activation with a correlational design. At a large state university, the researchers 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. The researchers 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 drills intrudes into the classrooms. Students in these rooms are assumed to be 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 the researchers 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, we suspect, this study 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. On the other hand, the limitations of this study should by now be clear: 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 causal relationship between these variables. For example, there is no way of knowing how students who took classes in the noisy building differed from students who took 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. Second, 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 control variation to insure 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 similar to 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.[1] 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 0.30 (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 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 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 the controls. 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 can we 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 may be that the flaw in the experiment is an act of commission (the researchers erred by introducing a confound), whereas the flaw in the correlational study was more an act of omission (the researchers erred by failing to consider the natural confounding of breakfast eating with any number of other variables). The correlational study was done poorly, 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. For example, suppose a researcher measured three variables: teachers' expectations about how well their students will perform, the children's academic achievement, and the educational level of the children's parents. Do the teachers' expectations predict the children's academic achievement independently of the third variable? Or, does the educational level of the parents predict both teacher expectations and children's academic achievement, with no relation between these latter two variables? Several

1. We have adapted this example from Mook, Wilson, and DePaulo (1995).

different relationships between these variables might exist and structural equation modeling is a useful technique for distinguishing between competing models.

We do not have the space to review all the pros and cons of structural equation modeling (for excellent reviews see Kenny, Kashy, & Bolger, 1998, in this *Handbook*; 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 of the main reasons 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 between two variables in a structural model (e.g., between teacher expectations and children's academic performance), 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 (e.g., perhaps the children's performance on standardized tests influences both academic performance and teachers' expectations). 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 (as Rosenthal, 1994, did in his seminal, experimental work on expectancy effects).

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 often 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? We will consider this question at length when we discuss external validity later in the chapter. To set the stage for this discussion it is useful to discuss some basic distinctions between different types of research.

Field versus Laboratory Research

Many observational and correlational studies are conducted in the field, such that people are unaware that they are being studied or observed. Laboratory research is true to its name: it is conducted in a laboratory, usually in such a way that people know they are being observed or that they are in a scientific investigation. When we ask students how they would study prejudice, they almost always propose field over laboratory studies. Their (quite reasonable) assumption is that behavior is best understood if it is observed in the context in which it naturally occurs. It is more difficult to generalize when people are observed in an artificial setting that they usually do not encounter in everyday life, such as in a laboratory in the psychology department of a university. It is unlikely, for example, that many people are asked to observe videotapes of people holding up cards with word fragments, and to generate words from these fragments.

Why, then, did Gilbert and Hixon (1991) conduct their study in the laboratory? We discussed the answer earlier: to gain enough control over the situation to be able to make causal inferences. In the hypothetical correlational study we discussed, in which students rated their instructors on the first day of class, there was no way to ensure that students in the different buildings had instructors with similar personalities. By showing people the same videotape of an Asian or Caucasian woman, Gilbert and Hixon (1991) solved this problem. Differences in ratings of the woman were likely to be due to the independent variable (cognitive busyness), and not to differences in who people were rating. A critical feature of experimental designs is that the situation is identical for all people except for the independent variables of interest. This is more difficult in naturalistic settings, in which the experimenter is often unable to control extraneous variables that can contaminate the results.

The distinction between field and laboratory research, however, is not always so clear-cut. One common misconception is to confuse it with the difference between experimental and correlational designs. True enough, most correlational studies are conducted in the field, and most experiments are conducted in the laboratory. As seen in Fig. 1, however, the two distinctions are not identical. It is sometimes possible to conduct experiments in naturalistic settings; we will discuss some classic field experiments later in this chapter. It is also possible to do correlational studies in the laboratory. The most common example of correlation laboratory studies is the attempt to develop a new measure of personality, in which people are given a battery of personality tests in a laboratory session, to see how much the new measure correlates with existing measures. Nonetheless, it is true that social psychological studies are not evenly distributed across the four cells in Fig. 1. The majority of studies are laboratory experiments. Increasingly, however, social psychologists are conducting research in naturalistic settings, and we will discuss both laboratory and field experiments later in the chapter.

In general, the laboratory makes it easier to accomplish the random assignment of people to conditions. In addition, a laboratory setting permits the researcher to manipulate independent variables more precisely and to eliminate or minimize the intrusiveness of "extraneous" variables. Advocates of laboratory experiments believe that the world is a complex place consisting of a great many noisy vari-

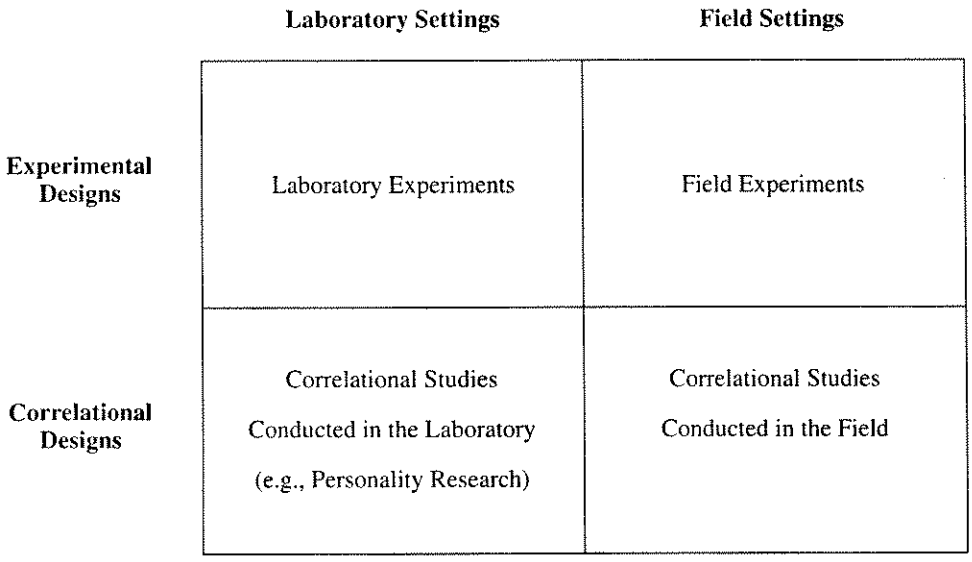

FIGURE 1 Experimental versus Correlational Designs in Field versus Laboratory Settings.

ables, a condition that impedes the chances of obtaining a pure indication of the effect of one variable upon another. If the experimenter wants to discover the effects of an event on the behavior, attitudes, or feelings of participants, the laboratory provides the sterility that enables observation of those effects unencumbered by extraneous variables that could confound interpretation. Conversely, the field is generally regarded as being more "real." In the real world the event in question always occurs in context; it is that very context that might have important but extraneous effects upon the behavior, feelings, or attitudes of the individual. Critics of the laboratory setting have suggested that it is silly to eliminate contextual variables in the interest of precision if those variables are always present in the world. We will elaborate on this distinction later in the chapter. For now we focus on another distinction concerning the focus of an experiment.

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

In some experiments the researcher is mainly interested in studying a phenomenon that he or she wants to understand and possibly change, such as prejudice. In others, the researcher is interested in 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, for example, without understanding the psychological mechanisms that produce it? 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 versus 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).

Our position is that to understand the causes of social psychological phenomena, it is often necessary to conduct experimental rather than correlational or observational studies, with process-oriented rather than problem-oriented approaches. Again, we do not intend to demean or devalue other approaches; any problem is best understood with a variety of techniques, and we have used other methods ourselves (e.g., correlational designs). To really get at the heart of a problem, however—namely, to understand its causes—experimental, process-oriented studies are often the method of choice, usually conducted in the laboratory instead of the field.

Interestingly, this assertion runs against the grain of the layperson's view of how research should be conducted. Think back to our early example of the student's suggestion about how to study prejudice: it was an observational study, it was conducted in the field, and it was problem-oriented rather than process-oriented. In general, we find that many people new to the field of social psychology believe that a problem should be studied in as natural a context as possible (i.e., in the field), often using correlational or observational techniques. Further, there is a definite preference for problem-oriented research. If the goal is to understand prejudice then the topic of the study should be prejudice. We have never heard a student new to the field say, "Prejudice should be studied by doing laboratory studies on automaticity or cognitive dissonance theory!"

Because of this (understandable) bias, students are sometimes dismayed to encounter so many studies that are just the opposite: experimental, laboratory, process-oriented studies. We hope the reader is at least somewhat convinced that in addition to studying a phenomenon such as prejudice in naturalistic settings, it is critical to study its underlying mechanisms in the laboratory, so that the causal relationships between intervening variables can be established more definitely. Sometimes, as we will see, this involves studying fundamental psychological processes, such as cognitive dissonance, attitude change, and social cognition in the laboratory, even if such studies do not directly investigate prejudice.

We will return to a discussion of the limits of laboratory research later in the chapter, when we discuss such issues as how the results obtained in the laboratory can be generalized to everyday life. Given the clear advantages of laboratory experiments—namely, the ability to determine causal relationships between variables—we turn now to a detailed discussion of how to do a laboratory experiment. In discussing the nuts and bolts of experimentation we will not lose sight of these important questions about the advantages or disadvantages of experiments and will in fact 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 for its purity as a model of experimental efficiency. Second, we discuss it as an example of basic, process-oriented research that is applicable to many different phenomena, including the one already mentioned—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 which 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 nearly forty 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 four decades ago—because it illuminates basic psychological processes does have a lot to say about a variety of current real-world phenomena, including prejudice.

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 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 or not 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 the ability 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. Though 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. While this is an important individual decision for the novice, 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, and will discuss in detail later, each approach has its advantages and disadvantages. There is no logical reason, however, for starting in one domain or the other nor is there any reason for assuming that particular hypotheses lend themselves more easily to the laboratory or the field. The decision is frequently dictated by such factors as the momentary availability of resources, idiosyncratic preferences of the experimenter, and so on.

Suppose you decide to bring the experiment into the laboratory. The next decision you must make is 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 these 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. Thus, the crucial distinction between an impact experiment and a judgment experiment is whether or not the event in question is happening to the participant. In the Aronson and Mills experiment, for example, the embarrassment produced by reciting obscene words was happening to the participants themselves. It is the effect of that embarrassment that is the major interest of the experimenter. In a judgment study the event might be important and dramatic, but it is not happening to the participant. For example, I (the participant) might read about or witness (via film) an aggressive or violent act (which might sicken or outrage me), but the violent act is not happening to me.

There are ideas that can be tested by either technique; for example, the investigation of equity. In some of these experiments, people are simply handed a description of the effort expended and product produced by individuals, given a distribution of the relative rewards or payments to the individual, and asked to evaluate the equity of the distribution. In other experiments, a person's own effort or output is rewarded in a more or less equitable way, and he or she is allowed to respond.

Some hypotheses, however, can only be tested with one type of experiment. 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.

The Four Stages of Laboratory Experimentation

The process of planning a laboratory experiment consists of four basic stages: (1) setting the stage for the experiment, (2) constructing the independent variable, (3) measuring the dependent variable, and (4) planning the postexperimental 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 much more complex and involves a wider scope of planning than does the judgment experiment. In effect, the judgment experiment is a "bare bones" operation. Although the design of both types requires attention to similar issues (e.g., random assignment, the order of presentation of the stimulus materials, and the context in which these materials are presented), the impact experiment entails a more elaborate scenario. 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: (1) 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, and (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 a 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 participants 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 do 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 one wants to make people fearful. One 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 here 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 sixty-four 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 The independent variable is the experimental manipulation. It is, ideally, a variable that is independent of all sources of variation except those specifically under the control of the experimenter. 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.

We will return to this issue later in the chapter, when we discuss different kinds of validity of experiments. 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. Although such characteristics as prejudice, intelligence, self-esteem, and socioeconomic class can be measured and taken into account or ignored, they should not be regarded as independent variables of an 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. Since people who become anxious easily may be different from those who do not, we

are dealing with an implicit personality 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, in one study, Wallace and Sadalla (1966) placed participants in a room with a complex machine and had a confederate tempt them to press a conspicuous button on the front of the machine. When a participant pressed the button, the machine exploded. Unfortunately, whether or not a participant chose to press the button was determined by the participant and not by the experimenter. Since there may be important differences between those who choose to press and those who do not, the experimenters in this kind of situation relinquish control to the participant and are left with a nonexperimental study.

The problem of free choice is a particularly sticky one because, if the hypothesis involves the effect 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. One solution to this problem is to conduct a pilot test of the variable until a level is found for it that is just sufficient enough to inhibit participants from actually choosing the "wrong" behavior. For example, in an experiment by Aronson and Carlsmith (1963), children were given either a mild or severe threat to prevent them from playing with a desirable toy. In order for this experiment to work, it was critical to make the mild threat strong enough to ensure compliance. On the other hand, it could not be too strong, for the experimental hypothesis hinged upon the child's not having a terribly good reason for declining to play with the toy. The situation had to be one in which the child was making a choice whether to play or not to play with the specific toy and was bothered by the lack of a good reason to avoid playing with that toy. It is sometimes possible to find such a level by elaborate pretesting. As an alternative, in some experimental situations a solution can be effected through the use of instructions that give a strong perception of choice, although little choice is actually present.

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-subject 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-subject 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) has used a between-subject 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-subject 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-subject 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 are averaged 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 indepen-

dent variable, and a large number of participants often have to be run to detect the "signal." In a within-subject 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 nonlies, thereby "subtracting 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-subject 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). Demand characteristics refers to features introduced into a research setting by virtue of the facts 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 are likely to make participants highly 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." This problem can present itself in both impact and judgment experiments, particularly those in which each participant is exposed to more than one variation of the stimulus. Such a procedure, by its very nature, increases the probability that the participant will begin to guess which aspects of the experiment are being systematically varied by the experimenter. This is less of a problem in most impact experiments where participants are presented with only one variation of a given independent variable. But, of course, manipulations with high impact may also create problems of participant awareness. It is for this reason that experimenters frequently employ deception, elaborate cover stories, and the like.

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 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 & Lawson, 1964; Rosenthal, 1994).

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 un-

aware 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. 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 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 questionnaire and asked them to read the list 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 participants 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 participants 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 premanipulation 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 assigned a different experimenter (unaware of the participant's experimental condition) to collect 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 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 awhile, 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 (1982) study mentioned earlier, the independent variable (whether people were constrained from playing with an unattractive 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.

Of course, if one is primarily concerned with eliminating participant awareness, the ultimate strategy is to induce the independent variable in such a manner that the participants are oblivious to the fact that any experiment is taking place at all. This strategy is best implemented in field settings, which we will discuss shortly.

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.

The experimenter also must ensure that the participants attend throughout the course of the experiment to the relevant stimulus conditions that constitute the independent variable. In judgment research this aspect of impact is particularly critical. All the care and effort devoted to careful and systematic stimulus control are wasted if the participant because of boredom or inattention fails to perceive the critical variations in the stimuli presented. Increasingly, stimuli in social psychological experiments are being presented in carefully controlled ways on computers, and it is interesting to note that this can be a two-edged sword with respect to keeping people's attention. Some participants find working on a computer more novel and interesting than working with the usual paper-and-pencil instruments, and thus pay more attention to the information that is presented. Other participants, however, find computer presentations to be a less engaging social exchange than a real-life interaction, and quickly become bored. The differences between communication over computers and direct communication between people is an interesting topic of research in its own right (e.g., Kiesler & Sproull, 1987). For present purposes the point is that researchers should be careful that their task is engaging to keep people's attention, whether it is presented over a computer or in the "old-fashioned way."

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 produces 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, Hodges, and LaFleur (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 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. If the slide projector malfunction did not influence people's memory, the hypotheses could not be tested. 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 conducted in the area 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 of thumb 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 (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 three 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 behave 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 ob-

taining 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; Stone et al., 1994). 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 a huge 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 recounted 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 the 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 and 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.

As another example, consider the Aronson and Mills (1959) study on initiation into a group. The major goal of this study was to investigate some of the conditions and consequences of cognitive dissonance. The main dependent variable was people's ratings of the attractiveness of the group they listened to. Again, there were no measures of actual behavior. The assumption was that people's questionnaire ratings of the group were a good proxy of actual behavior toward a group, an assumption that has been borne out in other studies (see, for example, Aronson & Osherow, 1980). Considered even more broadly, however, the Aronson and Mills study can also be viewed as an investigation of some of the processes responsible for prejudice: developing favoritism for one's own group due to self-justification needs, and, possibly, derogating other groups as a result. Superficially, it seems odd to suggest that a laboratory study in which people listened to a boring group discussion about sexual behavior in animals might have anything to do with prejudice. True, the researchers did not investigate the end product of prejudice: negative manner toward an individual because of his or her group membership. Nevertheless, it seems perfectly reasonable to view this experiment as a study of some of the psychological processes *responsible* for prejudice.

Occasionally, the researcher is interested in both problem and process simultaneously. For example, in the condom research (mentioned above), Aronson and his students were interested in persuading sexually active teenagers to use condoms as a way of safeguarding against AIDS. That's problem-oriented research. They were also interested in testing their new theory of hypocrisy induction. Specifically, they were testing the hypothesis that a change in behavior could be brought about by making people mindful of the fact that they were not practicing what they were preaching—in this case, that the participants were urging others to practice safer sex, while they, themselves, were falling far short of the mark. In this instance, the investigators felt it was important to test the process in a manner that came as close to addressing the problem as they could reasonably get. If they had been primarily interested in the process, they would have chosen a different problem—one where a meaningful dependent variable would be easier to measure (see, for example, Dickerson, Thibodeau, Aronson, & Miller (1992).

A third (and related) reason why nonbehavioral measures are often used is that, in many situations, 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.

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 & Stone, 1985).

In recent years techniques have been developed to avoid some of the problems of reports about one's own cognitive processes; for example, Ericsson and Simon (1993) advocate the use of verbal protocols, in which people "think aloud" into a tape recorder, verbalizing whatever thoughts they happen to have. This technique can be a valuable means of assessing the contents of consciousness, avoiding the kinds of interpretations and distortions that occur when people attempt to reconstruct, after the passage of time, what they had been thinking (Fiske & Ruscher, 1989). Even when thinking aloud, however, people cannot report cognitive processes that are inaccessible to consciousness. Ericsson and Simon (1993) acknowledge this problem but are of the opinion that nonconscious processing is rare. In contrast, a good deal of recent research has found evidence for widespread nonconscious processing of many types (e.g., Higgins & Bargh, 1992; Jacoby, Lindsay, & Toth, 1992; Kihlstrom, 1987; Wegner, 1994).

Assuming that verbal protocols tap all aspects of cognitive processing is a dangerous enterprise, and whenever possible, researchers should check the validity of the reports (Wilson, 1994). Further, there is some evidence that verbal protocols can be reactive, by changing the nature of people's thought processes (Schooler, Ohlsson, & Brooks, 1993; Wilson, 1994). More generally, any type of self-report instrument should not be taken on faith; its relation to the criterion variable of interest, which is usually overt behavior, should be tested.

There are several more mundane problems to be considered in making concrete decisions about what the dependent variable should be. One constantly recurring question is the extent to which the behavior of the participant should be constrained. This takes several forms. First, should one attempt to block most possible alternative behaviors so as to maximize the likelihood of observing changes in the specific variable of interest? For example, in a dissonance study, should the experimenter attempt to rule out all possible methods of reducing dissonance except the one he or she has decided to study? Clearly doing this will maximize the likelihood of observing differences in the behavior studied. This is a perfectly sound and reasonable technique. Indeed, it is part of our definition of experimental control. However, we do this only when we ask a certain kind of question, namely: "Is there dissonance in this situation and does it get reduced?" If this is the question, the experimenter should attempt to construct the experiment in order to be ready and able to measure the effects of the independent variable as powerfully as possible.

The investigator, however, may have a different question in mind. He or she may want to find out how people typically reduce dissonance. If this is the question, the preceding technique will almost certainly obscure what the participant really is likely to do in a situation of this sort and present the experimenter with an artificial relationship. The same concern arises when a researcher tries to decide whether to use open-ended questions or a rigidly constrained measure. Although the more quantitative measure may increase the likelihood of observing differences between experimental treatments, it also may obscure what the behavior of the participant would normally be. Any experimenter who has seen many participants close at hand has experienced the feeling that a given person is "really" showing many interesting effects, although the measures are too constrained to be sensitive to them. One answer to this problem is to use more qualitative methods, which are open-ended interviews with participants without a prior set of questions or hypotheses; trying to understand the world as participants see it, without imposing the researchers' world view.

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 ear-

lier 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 study by Wilson et al. (1993) 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 that 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 (DePaulo, 1992; Baumeister, 1982; 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, Rholes, and Jones, 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 seven-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 seven-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. Consider the following examples.

1. As we have seen, one of the most important topics in social psychology is prejudice and stereotyping. Measuring how prejudiced someone is is very difficult, however, for the reasons we have been discussing: people do not want to admit their prejudice. As a way around this problem, researchers are increasingly relying on indirect measures of prejudice, rather than self-report instruments (e.g., Brewer, Dull, & Lui, 1981; Crosby, Bromley, & Saxe, 1980; Dovidio & Fazio, 1992; Fazio, Jackson, & Williams, 1995). For example, in the Gilbert and Hixon (1991) study reviewed earlier, the extent to which people used an Asian stereotype was measured by giving people a word completion task, and counting the number of times they came up with words that were consistent with the stereotype (e.g., saying "polite" instead of "police" when given the fragment "poli_e"). The more indirect and uncontrollable a measure of prejudice is, the more confident we can be that it is tapping how people really feel and not the feelings they want to display publicly.

2. In recent years it has become clear that some responses occur automatically, in the sense that they are uncontrollable, occur without conscious intention or awareness, and do not require processing resources. Automaticity is not an either-or phenomena, but occurs in degrees (Bargh, 1989). Research on automaticity was initially studied by cognitive psychologists interested in motor behavior and relatively "low level" responding, but it quickly became apparent that many kinds of interesting social behaviors are also automatic—which means that people cannot easily control these responses according to self-presentational concerns or demand characteristics. A number of recent researchers have taken advantage of this fact, including measures of automaticity to explore a number of fascinating mental phenomena (e.g., Bargh, 1990; Gilbert, 1991; Greenwald & Banaji, 1995; Wegner, 1994).

Social psychologists have borrowed a number of other dependent variables from cognitive psychology as well. Attitude researchers have long regarded differential or biased attention to, and memory for, pro versus con attitudinal statements as a measure of an individual's attitudinal position. Recently, more sophisticated measures of recognition, reaction time, and accuracy of recall have been applied to other aspects of social perception. Again, assuming people exercise less control over memory than they do over verbal self-reports, such measures may reveal phenomena that otherwise may be suppressed—as in our example of prejudice.

For instance, it may no longer be socially desirable for subjects to admit that they think physicians should be men rather than women. Thus, a direct measure might fail to reveal such biased expectations even if they existed in subjects' minds. Suppose instead that a researcher presented people with a picture of a man or a woman dressed like a physician and surrounded by medical paraphernalia. The participants could then be asked to identify the person's profession as quickly as possible. If it requires more time to identify correctly the female picture than the male picture, this differential recognition may be taken as an indication of a continuing propensity to think of the "ideal" doctor as a male. Similarly, suppose people are able to recall information about a pattern that fits social stereotypes to a greater extent than information that is unrelated to the stereotype (e.g., Hamilton & Rose, 1980; Brewer, Dull, & Lui, 1981). This would provide indirect but reasonably conclusive evidence of the presence of stereotyping, where a more direct measure may have revealed none. The use of all such measures presupposes some hypothetical mechanism linking biases in information processing and recall with underlying cognitive structures or beliefs.

3. The study of nonverbal communication has included measures of behavior that are difficult to control. When someone is trying to deceive someone else, for example, he or she is actively trying to control or mask their nonverbal responses in such a way that the other person is unaware that they are lying. Some nonverbal channels (e.g., facial expressions) are easier to control than others (e.g., tone of voice), thus some channels are more likely to "leak" a person's true feelings. By including such measures, researchers can more easily bypass people's attempts to control or suppress how they really feel (DePaulo, 1992).

4. The use of physiological measures, such as galvanic skin response or heart rate, is increasing. An obvious advantage of such measures is that they are difficult for people to control. A drawback is that often there is no single, physiological response that is a precise measure of the psychological state the researcher wants to measure. There is no physiological measure, for example, that is a perfect correlate of anxiety or fear or happiness. The use of such measures is becoming increasingly sophisticated, however, and shows promise for measuring psychological states. For example, some researchers are using measures of electromyographic activity over facial muscles to assess both the valence and intensity of people's affective reactions (Cacioppo, Petty, & Tassinary, 1989). In addition, sometimes physiological responses are of interest in their own right, as indicants of people's health and responses to stress (e.g., Pennebaker, 1983). Rodin and Langer (1977) and Schultz and Hanusa (1978), for example, were interested in the effects of perceived control on the health of nursing home residents and found that changes in perceived control had effects on the residents' health, as assessed by such things as the number of medications they took and ratings by the staff. In addition, these studies included what can be considered the ultimate dependent measures: mortality rates. Interestingly, changes in the residents' perceptions of control and predictability influenced how soon residents died.

An interesting variation on the true physiology measure was devised by Jones and Sigall (1971). Their technique,

dubbed the "bogus pipeline," consists of convincing the participants that the experimenter has an accurate physiological measure of their attitudes. This is accomplished by the use of an electrical apparatus rigged so that, before the experiment, the participants receive a striking demonstration: the electrodes attached to their arm affect a needle on a dial in a manner consistent with their actual feelings on a number of issues. Actually, the dial is surreptitiously manipulated by the experimenter. Subsequently, the participants are asked to state their true attitudes while attached to the electrodes (although they themselves cannot view the dial). Since the participants believe that the experimenter can read the dial and that the dial reflects their real feelings, they are motivated to respond as accurately as possible. This device has proved particularly useful in measuring socially sensitive attitudes. For example, Sigall and Page (1971) found that white participants connected to the "bogus pipeline" expressed attitudes toward blacks that were more stereotypically negative than did participants in a "no pipeline" control condition.

Planning 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 four important goals:

1. To ensure that the participants are in a good and healthy frame of mind.
2. 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.
3. 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.
4. 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 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 or not the participant was suspicious and whether or not 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 or not 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 participant's 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, whether or not 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 post-experimental 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.

We would like to close this section by emphasizing our recommendation that a thorough explanation of the experiment should be provided *whether or not 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 are 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 nineteen dollars) 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 one dollar). 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, Banks, & Zimbardo, 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 it 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).

Moving into the Field

We have gone into considerable detail discussing the features and conduct of laboratory experiments because we believe it provides the prototypic (if not necessarily modal) case of social psychological experimentation. Certainly the four stages of research associated with the lab experiment—setting the stage, constructing the independent variable, measuring the dependent variable, and debriefing—are in some form common to all experimental research endeavors.

As we mentioned earlier, one big advantage of conducting an experiment in the laboratory is that the researcher has more control over the situation, which allows for a "cleaner" manipulation of the independent variable. In field experiments it is more likely that unforeseen, uncontrolled events will occur that will compromise the integrity of the experimental design. Nonetheless there are big advantages to field experiments as well, such as the fact that people are less likely to know that they are in an experiment and the setting will be more like ones people encounter in their everyday lives. A number of extremely important and clever field experiments have been done in social psychology, and in this section we will focus on the ways in which the conduct of field experiments is most likely to differ from that of the prototypic lab study.

Control over the Independent Variable The amount of control an experimenter has over the independent variable is a matter of degree. In some cases, the researcher constructs experimental situations from scratch, creating the background context as well as experimental variations. In other cases, the experimenter controls less of the setting but introduces some systematic variation into existing conditions, as in the field experiment by Piliavin, Rodin, and Piliavin (1969) in which the behavior of an experimental accomplice was varied in the largely uncontrolled context of a New York subway train in order to study bystander helping in that setting.

In other field research, the experimenter does not manipulate the stimulus conditions but instead selects among naturally occurring stimulus situations those that embody representations of the conceptual variable of interest. Here the line between experimental and correlational research becomes thin indeed, and the distinction depends largely on how standardized the selected field conditions can be across participants. One good illustration of the use of selected field sites in conjunction with laboratory research comes from the literature on mood and altruism. A variety of mood-induction manipulations have been developed in laboratory settings, such as having participants read affectively positive or negative passages (e.g., Aderman, 1972). After the mood state induction, participants are given an opportunity to exhibit generosity by donating money or helping an experimental accomplice. Results generally show that positive mood induction elevates helping behavior (Salovey, Mayer, & Rosenhan, 1991). Despite multiple replications of this effect in different laboratories with dif-

ferent investigators, the validity of these findings has been challenged both because of the artificiality of the setting in which altruism is assessed and because of the potential demand characteristics associated with the rather unusual mood-induction experience.

To counter these criticisms, researchers in the area took advantage of a natural mood-induction situation based on the emotional impact of selected motion pictures (Underwood et al., 1977). After the pilot research in which ratings were obtained from moviegoers, a double feature consisting of *Lady Sings the Blues* and *The Sterile Cuckoo* was selected for its negative affect-inducing qualities, and two other double features were selected to serve as neutral control conditions. A commonly occurring event—solicitation of donations to a nationally known charity with collection boxes set up outside the movie theater lobby—was chosen as the vehicle for a measure of the dependent variable of generosity.

Having located such naturally occurring variants of the laboratory mood-induction operation and altruism measure, the major design problem encountered by the researchers was that of participant self-selection to the alternative movie conditions. Whereas random assignment of volunteer moviegoers was a logical possibility, the procedures involved in utilizing that strategy would have created many of the elements of artificiality and reactivity that the field setting was selected to avoid. Therefore, the investigators decided to live with the phenomenon of self-selection and to alter the research design to take its effect into consideration. For this purpose, the timing of collection of donations to charity at the various theaters was randomly alternated across different nights so that it would occur either while most people were entering the theater (before seeing the movies) or leaving (after seeing both features). The rate of donations given by arriving moviegoers could then be a check on preexisting differences between the two populations apart from the mood induction. Fortunately, there proved to be no differences in initial donation rates, as a function of type of movie, whereas post-movie donations differed significantly in the direction of lowered contribution rates following the sad movies. This pattern of results, then, preserved the logic of random assignment (initial equivalence between experimental conditions) despite the considerable deviation from ideal procedures for participant assignment.

Two points should be emphasized with respect to this illustration of field research. First of all, the field version of the basic research paradigm was not—and could not be—simply a "transplanted" replication of the laboratory operations. The researchers had considerably less control in the field setting. They could not control the implementation of the stimulus conditions or extraneous sources of variation. On any one night a host of irrelevant events may have occurred during the course of the movies (e.g., a breakdown of projectors or a disturbance in the audience) that could have interfered with the mood manipulation. The researcher was not only helpless to prevent such events but would not have been aware of them if they did take place. In addition, as already mentioned, in the field setting the experimenters were unable to assign participants randomly to conditions and had to rely on luck to establish initial equivalence between groups.

The second point to be emphasized is that the results of the field experiment as a single isolated study would have been difficult to interpret without the context of conceptually related laboratory experiments. This difficulty is partly due to the ambiguities introduced by the alterations in design and partly to the constraints on measurement inherent in the field situation where manipulation checks, for example, are not possible. The convergence of results in the two settings greatly enhances our confidence in the findings from both sets of operations. Had the field experiment failed to replicate the laboratory results, however, numerous alternative explanations would have rendered interpretation very difficult.

Random Assignment in Field Settings Subject self-selection problems plague field experimentation in multiple forms. In the field experiment of mood and helping behavior cited previously, random assignment to experimental conditions was not even attempted. Instead, the effects of potential selection factors were handled in other ways that involved an element of risk taking. The premovie data collection served as a check on the assumption that people who attend sad movies are not inherently different from people who attend other movies in their propensity to give to charities. But what if that assumption had proved false and there had been an initial difference in the rate of donations between attendants at the different types of movie? Such previous differences in behavior would have made interpretation of any differences in donations after exposure to the movies hazardous at best. In this case, the researchers were taking a gamble in counting on the absence of initial population differences. The logic of their experimental design required that the premovie data collection sessions be interspersed with postmovie data collection in order to control for timing effects. As a consequence, the investigators could not know until after the experiment had been completed whether the data supported their assumption of initial equivalence. Had they been wrong, the experimental design would have been undermined, and any effort expended would have been wasted.

In other settings, too, the research may rely on the essentially haphazard distribution of naturally occurring events as equivalent to controlled experimental design. Parker, Brewer, and Spencer (1980), for instance, undertook a study on the outcomes of a natural disaster—a devastating brush fire in a southern California community—

on the premise that the pattern of destruction of private homes in the fire constituted a "natural randomization" process. Among homes in close proximity at the height of the fire, only chance factors—shifts in wind direction and velocity, location of fire fighting equipment, and traffic congestion—determined which structures were burned to the ground and which remained standing when the fire was brought under control. Thus, homeowners who were victims of the fire and those who were not victimized could be regarded as essentially equivalent before the effects of the fire, and any differences in their attitudes and perceptions following the fire could be attributed to that particular differential experience. When comparisons are made between such naturally selected groupings, the burden of proof rests on the investigator to make a convincing case that the groups are not likely to differ systematically in any relevant dimensions other than the causal event of interest.

In other field research efforts, the researcher may be able to assign participants randomly to experimental conditions. However, once assigned, some participants may fail to participate or to experience the experimental manipulation. If such self-determined "de-selection" (also known as "participant mortality") occurs differentially across treatment conditions, the experimental design is seriously compromised. One way of preserving the advantages of randomization in such cases is to include participants in their assigned experimental conditions for purposes of analysis regardless of whether they were exposed to the treatment or not (assuming, of course, that one is in a position to obtain measures on the dependent variable for these participants). This was the solution applied in the two field experiments conducted by Freedman and Fraser (1966) to test the effectiveness of the "foot-in-the-door" technique for enhancing compliance.

In these studies the dependent variable was whether individuals contacted in their homes would agree to a rather large, intrusive request from the researcher (e.g., to permit a five-person market survey team to come into the home for two hours to classify household products). Of primary interest was the rate of compliance to this large request by participants who had been contacted previously with a small request (e.g., to respond to a very brief market survey over the telephone), in comparison to that of the control participants who were contacted for the first time at the time of the large request.

The purpose of the manipulation in the Freedman and Fraser studies was to test the effect of actual *compliance* to the initial small request on response to the later request. However, the operational experimental treatment to which potential participants could be randomly assigned was exposure to the request itself. Approximately one-third of those who were given the initial small request refused to comply; hence they failed to complete the experimental manipulation. If these participants had been excluded from the study, the comparability between the remaining experimental participants and those randomly assigned to the no-initial-contact condition would have been seriously suspect. To avoid this selection problem, the researchers decided to include measures from all participants in the originally assigned treatment groups, regardless of their response to the initial request. With respect to testing treatment effects, this was a conservative decision, since the full treatment was significantly diluted among those classified in the experimental group. As it turned out, the initial compliance effect was powerful enough to generate a significant difference between treatment groups (of the order of 50 percent versus 20 percent compliance rates) despite the dilution of the experimental condition. Had the results been more equivocal, however, we would have been uncertain whether to attribute the absence of significant differences to lack of treatment effects or to failure to achieve the experimental manipulation. When the experimental treatment condition is diluted even more seriously than in the present illustration, comparisons between intact treatment groups become meaningless, and more sophisticated techniques for correcting for participant self-selection must be adopted (Brewer, 1976).

When full random assignment cannot be implemented in field settings, various forms of "quasi experiments" (cf. Cook & Campbell, 1979) can be creatively employed to preserve the logic of experimental design and control without rigid adherence to specific procedures. It should be kept in mind, however, that loss of control over stimulus conditions or participant assignment inevitably carries with it some measure of risk. Assumptions upon which the quasi-experimental design rests (such as initial equivalence of different groups) may prove untenable, or uncontrolled environmental inputs may "swamp" the stimulus conditions of interest to the researcher. In such cases, the costs in terms of wasted effort are high; thus decisions to take risks in undertaking field studies must be made sensibly. It would be foolish not to adjust the features of one's research design to the practical realities of a given field setting. But it is even more foolish to proceed with an expensive study that, from the start, has a high probability of resulting in uninterpretable outcomes.

Assessment of Dependent Variables in Field Settings

In many field contexts, the design and evaluation of dependent measures is parallel to that of laboratory experiments. In the guise of a person-on-the-street interview or a market research survey, for example, the field researcher may elicit self-reports of relevant attitudes, perceptions, or preferences. Or behavioroid measures may be designed that assess the willingness of the participants to engage in relevant acts such as signing a petition or committing themselves to some future effort. Finally, situations may be

constructed so as to elicit the type of behavior of interest to the experimenter, such as providing participants with opportunities to donate to charity (Underwood et al., 1977), to help a stranger who has collapsed (Piliavin et al., 1969), or to trade in a lottery ticket (Langer, 1975). One advantage of experimentation in field settings is the potential for assessing behaviors that are, in and of themselves, of some significance to the participant. Instead of asking participants to report on perceptions or intentions, we may observe them engaging in behaviors with real consequences. In such cases, our dependent measures are much less likely to be influenced by demand characteristics or social desirability response biases. In laboratory settings participants may check a particular point on a liking scale in order to please the experimenter or to look good; however very few people would choose someone as a roommate for the entire year unless there were more powerful reasons.

In some field settings, the kinds of dependent measures typically employed in laboratory studies would be exclusively intrusive in ways that would destroy the natural flow of events characteristic of the setting. Field experimenters have to be particularly sensitive to the issue of "reactivity" discussed by Campbell and Stanley (1963). This concept refers to the possibility that the measurement of the dependent variable reacts with the independent variable or related events in such a way that effects are found that would not have been present otherwise. For example, suppose some people have seen a movie designed to reduce prejudice. They may be completely unaffected by this movie *until* they are asked to fill out a questionnaire that clearly deals with prejudice. As a result of seeing this questionnaire, the moviegoers may realize for the first time that the movie was about prejudice and may reflect on the movie in a new way that now has an influence. In effect, the introduction of the dependent measure has served as a kind of independent variable in combination with the originally intended treatment variable. Note that this kind of effect is conceptually different from experimental artifacts generated by demand characteristics or experimenter bias effects. We are not postulating that the respondent changes the expression of prejudicial attitudes in order to please the experimenter but only that no change would have taken place without the intrusion of a very obvious measurement.

In order to prevent or minimize the occurrence of reactivity, field researchers may devise a variety of techniques to make unobtrusive measurements of the dependent variable of interest (see Webb et al., 1981). Some unobtrusive measures are based on observations of ongoing behavior, utilizing methods of observation that interfere minimally or not at all with the occurrence of the behavior. For instance, voluntary seating aggregation patterns have been used as an index of racial attitudes under varied conditions of classroom desegregation; observational studies of conformity have recorded public behaviors such as pedestrians crossing against traffic lights or turn signaling by automobile drivers, and studies of natural language often resort to eavesdropping on conversations in public places. Cialdini et al. (1976) used naturalistic observation of clothing and accessories to study what they call the "Basking in Reflected Glory" phenomenon. They recorded the wearing of T-shirts and other apparel bearing the school name or insignia by students in introductory psychology classes at seven universities each Monday during football season. The proportion of students wearing such apparel at each school proved to be significantly greater on Mondays following a victory by that school's team than on Mondays following defeat. A simple monitoring of public displays provided quantitative confirmation of the hypothesized tendency to identify with success.

Other observational techniques may rely on the use of hidden hardware for audio or video recording of events outside the experimenter's control, such as physical traces left after an event has occurred or archival records that are kept for administrative or economic purposes (police files, school absenteeism records, and sales figures). One interesting illustration of the use of unobtrusive physical trace measures is provided in the previously mentioned Langer and Rodin's (1976) field experiment testing the effects of responsibility inductions on the well-being of residents of a nursing home. The major outcome of interest in that study was the general alertness and activity level of the residents following introduction of the experimental treatment. This level was assessed not only by the traditional methods of self-report and the ratings of nurses but also by various specially designed behavioral measures. One measure involved covering the right wheels of patients' wheelchairs with two inches of white adhesive tape, which was removed after twenty-four hours and analyzed for the amount of discoloration as an index of patient-activity level. Alas, clever ideas do not always work; the amount of dirt picked up by the tape turned out to be negligible for patients in all conditions.

The results of the Langer and Rodin nursing home study serve to illustrate some of the problems of reliance on unobtrusive measures in field settings. The adhesive tape index did not produce any detectable treatment effect; other, more direct and experimenter-controlled, self-report, and behavioral measures demonstrated significant impact of the experimental treatment. Had the researchers been forced to limit their assessment of effects to the least intrusive measure, they would have missed a great deal. The validity of dependent variable measures—the extent to which they measure what they are supposed to measure—is of concern in any research endeavor. However, the farther removed the actual measure is from the variable of interest, the more reason there is for concern. For instance, consider the number of steps involved in going from the dependent variable of patient-activity level of the measurement of discoloration of white adhesive tape in the nursing home study. First patient activity had to be translated into dis-

tance traveled in the wheelchair, which in turn had to be related to the amount of dirt picked up by different sections of the tape, which in turn had to produce measurable differences in discoloration. In such a chain, many intervening processes can reduce the correspondence between the initial variable (activity) and the measured outcome—the speed with which the wheelchair traveled, how often the floors were cleaned, whether the patients' movement was self-propelled or passive, and so on. Reliance on a single measure affected by so many irrelevant factors would have been treacherous indeed.

Sometimes indirect, unobtrusive measures do prove sensitive to experimental treatments but still turn out to be measuring the wrong thing. For example, the residents of Portland, Oregon, participated in an experimental attempt to decrease automobile use by lowering bus fares for a trial period (Katzev & Backman, 1982). Bus-rider records provided evidence that the goals of the study were being met during the experimental period: bus ridership was way up. Unfortunately, however, the use of bus-rider records as an indirect (an unobtrusive) measure of reduction in automobile use proved to be misleading. The researchers kept careful odometer records of cars before and during the study and found there was no decrease in average miles driven. One possible explanation is that people felt so virtuous riding to work on the bus every day that they treated themselves to long recreational car trips on weekends! Reliance on bus ridership alone as a measure of the program's success would have led to an inappropriate conclusion.

VALIDITY AND REALISM IN EXPERIMENTS

To this point we have discussed in some detail how laboratory and field research is conducted. It is important to return to a fundamental question: Given the advantages of field experiments, why are laboratory studies conducted at all? Indeed, it seems to us that the perfect social psychological study would be one that was conducted in a naturalistic setting, in which people were randomly assigned to experimental conditions, the independent variable was one that was impactful and involving, and all extraneous variables were controlled. Sounds good, doesn't it? Unfortunately, such a study is like a Platonic ideal that can rarely be achieved. For reasons we have already discussed, experimentation almost always involves a trade-off between competing goals: control and realism. It is worth discussing this trade-off in more detail, in terms of precisely what is meant when we say an experiment is "well-controlled" or "realistic."

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.

The meaningfulness of experimental research rests first of all on *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 systemic 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 before the research operations. In the field study of mood and altruism cited earlier, had it happened that the moviegoers who attend sad films differed from attendees at other movies in their propensity to contribute to charities even before exposure to the movies themselves, that difference would have constituted a threat to the internal validity of the study. In that case, any differences obtained between groups in donations following the movies could be interpreted either as an effect of the movie mood induction or of personality differences unrelated to the treatment. Other factors that can undermine internal validity include the occurrence of events during the course of the research that are unrelated to the treatment variable and that produce different effects in the various experimental groups.

Internal validity is the sine qua non of good experimental research. The procedures for standardizing treatments, avoiding bias, and assuring comparability of participant groups discussed in our earlier section on planning and conducting laboratory experiments are all addressed to internal validity concerns. The essence of good experimental design is to control the assignment of participants to treatment groups and the conditions of treatment delivery in such a way as to rule out or minimize threats to the internal validity of the study. Thus any differences obtained on outcome measures can be traced directly to the variations in treatment introduced by the experimenter.

One major limitation on internal validity is the extent to which unmeasured individual differences may obscure the results of an experiment. The ideal of an experiment is to take two identical units (corn plants, rocks, rats, children, or fraternities) and to apply different experimental treatments to them. Although it is a philosophical truism that no two units are ever precisely identical, the experimenter must strive to make them as close to identical as possible. One can approximate this idea much more satisfactorily in most sciences, and even in most of psychology, than is

possible in social psychology. Our participants differ from each other genetically, in learned personality characteristics, in values and attitudes, in abilities, and in immediate past experiences. Any and all of these differences may have a large impact on the way in which participants respond to our experimental treatments. This is one reason why it is important in most social psychological experiments to have large sample sizes; through random assignment and large numbers, the differences between individuals are "averaged out." It is also the reason, as we saw earlier, that a within-subject manipulation of the independent variable, when feasible, is advantageous. In a within-subject design, each participant serves as his or her own control, which controls for any number of individual difference variables that would be treated as error variance in a between-subject design.

As we have seen, it is usually much easier to maintain high internal validity in a laboratory experiment, because the researcher has much 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* 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 fifty-year-old working-class immigrants. External validity refers to the extent to which a particular causal relationship is robust across populations or 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 had failed a test of creative problem-solving might 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. The external validity of a finding is challenged if the relationship between independent and dependent variables is altered if essentially the same research procedures were conducted in a different laboratory or field setting or under the influence of different experimenter characteristics. 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.

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 only 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 with numerous stimulus features. As we discussed earlier, one of the most difficult parts of experimental design is constructing a concrete independent vari-

able (e.g., reciting a list of obscene words) that is a good instantiation of the conceptual variable (cognitive dissonance produced by undergoing an embarrassing initiation to a group). 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), has 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. Our confidence in the validity of cause and effect results from a particular experiment may be enhanced if the finding is repeated on other occasions. However, the degree to which a study has internal validity is determined by characteristics intrinsic to the study itself. 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. A researcher does the best that he or she 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 broadly between 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 af-

fected 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 to those that occur in everyday life. Consider the Gilbert and Hixon (1991) study we described at the beginning of the chapter. 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 very 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.

For example, Langer (1989) has conducted a number of studies on "mindlessness," or what happens when people are in a state of reduced attention. People are hypothesized to be on "automatic pilot," following well-learned rules rather than actively attending to what is happening in their environment. To study this mode of processing, Langer, Blank, and Chanowitz (1978) sent a nonsensical memo to a sample of secretaries that asked them to return the memo immediately to the sender. If people thought about the request they probably wouldn't return the memo; after all, if the sender wanted the memo, why didn't he just keep it in the first place? In some conditions Langer et al. attempted to mimic everyday situations in which people are not processing information very carefully and thus mindlessly do rather silly things. In one condition, for example, the memo was made to look like the ones the secretaries got most of the time: it was unsigned and was phrased more in terms of a request ("I would appreciate it if you would return this paper immediately to Room 238") than a demand ("This paper is to be returned immediately to Room 238"). In this condition almost everyone returned the memo to its sender. When the memo was unusual, however—such as when it was phrased in terms of a demand—people were predicted to think about it more carefully, recognize its absurdity, and fail to return it—which is exactly what happened. For our purposes, the point is that the critical experimental condition—the one in which people were predicted to respond mindlessly—was one that was the lowest in experimental realism, but high in psychological 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 (perhaps mindlessly!) 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 research 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

armed 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 or not 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.

Mook also points out that some experiments are valuable because they answer quesitons 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. 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. This is especially true of basic research, of course; the point of applied research is to find solutions to problems that will work in everyday life. 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.

The Basic Dilemma of the Social Psychologist

It should be clear by now why our perfect experiment—one conducted in a naturalistic setting, in which people were randomly assigned to experimental conditions, the independent variable was one that was impactful and involving, and all extraneous variables were controlled—would be extremely difficult to perform. It is next to impossible to design an experiment that is high in both internal and external validity. We see this as the basic dilemma of the experimental social psychologist. On the one hand we want maximal control over the independent variable, to maintain internal validity. We want as few extraneous differences as possible between our treatments. We want a precise specification of the treatment we have administered and of its effect on the participant. These desires lead us to try to develop manipulations that are highly specifiable, in which the differences between treatments are extraordinarily simply and clear and in which all manipulations are standardized—in short, to an approximation of something like a verbal learning experiment. On the other hand, if the experiment is controlled to the point of being sterile, it may fail to involve participants, have little impact on them, and therefore may not affect their behavior to any great extent. It might be quite low in psychological realism, limiting the extent to which we can generalize the results to other settings.

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 trade-off 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.

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 has been operationalized in a wide variety of ways in both laboratory and field studies, including having people read lists of obscene words, write counterattitudinal essays, eat unpleasant foods, and make a difficult choice between which horse to bet on at a racetrack.

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 have 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.

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 and therefore require supplementary information. For these reasons, good programmatic research involves replication with systematic differences and similarities in procedures and operations so that differences in results are potentially interpretable. In many cases, including exact replication along with conceptual variations is 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 precedure in all respects save one: he changes the operations in order to eliminate this artifact. He 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 his 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 the change in the procedure. If he failed even to replicate Jones's basic experiment, we would have to conclude that there was some important factor in the variables used in the original experiment, that the results are limited to a particular population, that either Jones or Smith (or both) had unconsciously biased their data, that Smith was simply incompetent, and so on.

In many situations it is difficult to 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 is 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.

Field Experimentation and Application

Conceptual replication highlights the advantages of combining laboratory and field experimentation for purposes of theory-building. In addition, the interplay between laboratory and field research is also critical to the development of

an effective applied social psychology. Basic process-oriented experimental research may isolate important causal processes; however, convincing demonstrations that those processes operate in applied settings are essential before theory can be converted into practice.

The research literature on psychological responsibility and control provides a particularly good example of how a synthesis between field and laboratory experiments can work at its best. It began with animal research in the laboratory (Brady, 1958), extended to field studies of stress in humans (e.g., Janis, 1958; Egbert, Battit, Tundorf, & Becker, 1963), then moved to laboratory analogues (e.g., Glass & Singer, 1972; Kanfer & Seidner, 1973), and returned to the field (e.g., Johnson & Leventhal, 1974; Langer & Rodin, 1976). Results from both settings repeatedly demonstrated the potent effect of the perception of control or responsibility on an individual's ability to cope with stressful events. Even the illusion that one has control over the onset or the consequences of potential stressors is apparently sufficient to increase tolerance for stress and reduce adverse effects. As a result of these findings, procedures developed for inducing actual or perceived personal control are applicable in medical practice and in the administration of health-care institutions. At the same time, the fact that field applications permit testing research hypotheses in the presence of severe, noxious, or potentially life-threatening situations has contributed substantially to our theoretical understanding of the role of psychological factors in physiological processes.

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 reseach. 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.

ETHICAL CONCERNS IN EXPERIMENTATION

In our discussion of the postexperimental follow-up, we touched on the topic of ethics. 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.

It should be emphasized, of course, that ethical problems arise even in the absence of either deception or extreme circumstances. We refer again to the experiment by Dawes, McTavish, and Shaklee (1977) as one of many possible examples of a benign-appearing procedure that can profoundly affect a few participants in ways that could not easily have been anticipated even by the most sensitive and caring of experimenters. 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 problematical. 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 confirm 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 et al., 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.

It could be argued that the results are good or useful for society even though the procedure may be harmful to some of the participants. Again, this does not, in and of itself, justify the procedure unless the participants themselves are in a position to weigh the societal benefits against the possibility of individual discomfort. It should also be clear that an ex post facto defense is not adequate. That is, suppose one finds after running the experiment that all participants attest that they are glad they participated and would still have agreed to participate if they had been properly informed in advance. This is not adequate because many participants who might have agreed to participate in advance might attempt to justify their participation after the fact as an ego protective device or as a way of helping the experimenter save face. Once the experiment is over, an ex post facto endorsement is ambiguous at best.

During the past several years, moral philosophers have entered the controversy and have suggested some solutions to the problem of informed consent which, while creative enough, strike us as being impractical in the extreme. One example might suffice. Sable (1978) has suggested a technique called "Prior General Consent Plus Proxy Consent." In this technique, the experimenter first obtains the general consent of the subject to participate in an experiment that may involve extreme procedures. The participant then empowers a friend to serve as a proxy; that is, to examine the details of the specific procedure in advance and to make a judgment as to whether the participant would have consented to it if given the choice. If the proxy says yes, then the experimenter may proceed. While this technique may be ethical in the most technical sense, it has some obvious flaws both ethically and methodologically. First, the participants are still agreeing to something that they cannot fully understand—the proxy can be wrong. Second, it is reasonable to assume that most proxies will probably make conservative errors; that is, they will try to protect the welfare of the participant by being more cautious than the participant would have been. If that is the case, and a substantial number of proxies say no, we may have a sample of extreme and unknown bias.

In recent years, 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. Recent APA ethical guidelines include these principles (American Psychological Association, 1992):

1. Investigators must take steps to protect the rights and welfare of their participants. Because individual researchers might not be objective judges of the ethical acceptability of their studies, they should seek ethical advice from others.
2. As much as possible, the researcher should describe the procedures to participants before they take part in a study and obtain informed consent from participants that documents their agreement to take part in the study as it was described to them. Participants must be informed that they are free to withdraw from the study at any point.
3. If the participant is legally incapable of giving informed consent (e.g., he or she is a minor), the investigator must obtain permission from a legally authorized person (e.g., the participant's parent).
4. Deception may be used only if it is "justified by the study's prospective scientific, educational, or applied value" and only if "equally effective alternative procedures that do not use deception are not feasible" (American Psychological Association, 1992, p. 1609). After the study, participants must be provided with a full description and explanation of all procedures, in a postexperimental interview. Participants in all studies should be given the opportunity to learn about the nature and results of the study.
x. All information obtained from individual participants must be held in strict confidence, unless the consent of the participant is obtained to make it public. Investigators must inform participants how their data will be used, and how it will be shared with others.

We have only summarized the guidelines here; we urge the reader to study the American Psychological Association guidelines carefully before undertaking the difficult task of ethical decision making. Further, as the guidelines state, this decision 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.

Ethical Issues in Field Research In many ways the nature of field research places even more responsibility on the researcher to weigh research needs against ethical principles. Whatever the ethical compromises of laboratory experiments may be, at least the laboratory setting assures that participants are aware in some sense that they are participating in research. Even if the consent to participate is not fully informed (or is actively *mis*informed) there is the presence of an "implicit contract" between participant and experimenter that reflects their mutual expectations about the conduct of research. The researcher's contractual obligation is partially fulfilled at the time of the debriefing and postexperimental interview, where any deceptions are unveiled, participants are informed about the goals and purposes of the research, and people's responses to the research procedures are assessed. Thus, even if the participants enter the experimental session ignorant of the researcher's intent, they do so in the expectation of being fully informed by the time it is all over. The postexperimental session also provides an opportunity for subject feedback to correct errors of judgment on the part of the experimenter. If the researcher misjudged the amount of distress or embarrassment the experimental procedure will cause people, information from the first few participants can provide a basis for altering those procedures before the research has gone too far.

When experiments are conducted in field settings where participants are *unaware* at the time that they are participating in research, the basic ethical dilemma is magnified. In such cases, there is not even an implicit contract to be adhered to, and decisions regarding ethical considerations rest solely with the experimenter. There are two different versions of participation without awareness: cases in which the participants are not informed that they had been involved in a research study until after their data have been collected, and cases in which participants are never informed. When participants are contacted and debriefed at the end of the field research, the goals and conduct of the postexperimental interview are essentially the same as those in a laboratory study, although there is one important difference, namely, the fact that the participants had no previous opportunity to decline participation. Hence not even an implicit obligation to cooperate can be assumed. Special sensitivity on the part of the researcher is required to avoid embarrassing people and to ensure that mechanisms are available for refusal to be included in the study after the fact. Again, however, postexperimental contact allows for

participant feedback and participation in judgments about legitimacy and appropriateness of the research procedures.

When participants are never told about the research, the opportunity for corrective feedback is greatly reduced. Under such circumstances, the researcher has the full obligation and responsibility of ensuring that the privacy of participants has not been violated, that they have been protected from undue embarrassment or distress, and that their lives have not been altered in any significant way by the nature of the research procedures. This obligation places the researcher in an essentially parental role that many find uncomfortable. To avoid this stance, some have suggested that ethical principles should proscribe the observation or recording of behavior for research purposes unless every person included in the study can be fully informed. This position strikes us as misguided. There are clearly cases of innocuous observation in public places (or the use of public records for research purposes) where the necessity to contact individual participants would destroy the anonymity that makes the procedures innocuous in the first place. Where data are recorded with no possibility of identifying information being available on the persons observed, postexperimental debriefing could produce more participant embarrassment than any effects associated with the observation itself. In such cases it is probably better that the research is conducted in ways that maximize anonymity rather than informed consent.

Of course, the mere fact that the participants (or the experimenter) may be embarrassed by disclosure of the research purpose does not by itself justify a decision to forego informing participants of their research participation. Such decisions should be closely restricted to situations involving high frequency behaviors in public places where only aggregate data are needed for research purposes. The decision becomes especially delicate when the researcher has intervened in the setting in any direct way that alters the situation beyond the normal range of events. In such cases, the experimenter must be sensitive to any possibility that participants may have been affected by the intervention in ways that warrant disclosure of the experimental setup. Finally the conscientious researcher should not rely on his or her judgment alone to make such determinations but should consult widely among people who may bring different perspectives to bear on the decisions to be made.

CONCLUDING COMMENTS

We began this chapter with a defense of the experimental method. For most social psychologists, this defense is, doubtless, unnecessary. For, although those who have been in the discipline for a while have a reasonably clear understanding of the difficulties and drawbacks of this approach, most also have developed an experience-based appreciation of its enormous advantages. For beginning students of social psychology, however, we are convinced that it is vital to take a very close look at precisely why a tightly controlled experiment, with random assignment of subjects to conditions, remains the method of choice. We hope this chapter has succeeded in doing just that, and that the student is now as convinced as we are as to the viability of the experimental method.

At the same time, it is of great importance that experimentalists (both veterans and novices) avoid complacency. As we have indicated earlier, research methodologies are not static but are continually evolving. As we continue to fine-tune the experimental method, we must meet a myriad of challenges by, for example, remaining sensitive to the situation—to determine precisely how our research design interacts with the research setting that interests us. We must also maintain a continued deep concern with ethics and, at the same time, attempt to ensure that this concern does not strangle creativity. We should continue to address serious and important social psychological questions in unique ways with equally serious and important experimental interventions both inside and outside the laboratory.

REFERENCES

Abrams, D., & Hogg, M. A. (Eds.) (1990). *Social identity theory: Constructive and critical advances.* London: Harvester Wheatsheaf.

Aderman, D. (1972). Elation, depression, and helping behavior. *Journal of Personality and Social Psychology, 24,* 91–101.

Allport, G. (1954). *The nature of prejudice.* Reading, MA: Addison-Wesley.

American Psychological Association. (1992). Ethical principles of psychologists and code of conduct. *American Psychologist, 47,* 1597–1611.

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: 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.

Back, K. W. (1951). Influences through social communication. *Journal of Abnormal and Social Psychology, 46,* 9–23.

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 Press.

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.

Brady, J. (1958). Ulcers in "executive monkeys." *Scientific American, 199,* 95–100.

Brehm, J. W., & Cole, A. H. (1966). Effect of a favor which reduces freedom. *Journal of Personality and Social Psychology, 3,* 420–426.

Brewer, M. B. (1976). Randomized invitations: One solution to the problem of voluntary treatment selection in program evaluation research. *Social Science and Research, 5,* 315–323.

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.

Brewer, M. B., & Brown, R. (1998). Intergroup relations. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol 2, pp. 554–629). New York: McGraw-Hill.

Brewer, M. B., Dull, V., & Lui, L. (1981). Perceptions of the elderly: stereotypes as prototypes. *Journal of Personality and Social Psychology, 41,* 656–670.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley: Univ. of California Press.

Cacioppo, J. T., Petty, R. E., & Tassinary, L. G. (1989). Social psychophysiology: A new look. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 22, pp. 39–91). San Diego: Academic Press.

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.

Chaiken, S. (1987). The heuristic model of persuasion. In M. P. Hanna, 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.

Cialdini, R. B., Borden, R. J., Thorne, A., Walker, M. R., Freeman, S., & Sloan, L. R. (1976). Basking in reflected glory: three (football) field studies. *Journal of Personality and Social Psychology, 34,* 366–375.

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

Crosby, F., Bromley, S., & Saxe, L. (1980). Recent unobtrusive studies of black and white discrimination and prejudice. A literature review. *Psychological Bulletin, 87,* 546–563.

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.

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. *Journal of Personality and Social Psychology, 45,* 1096–1103.

Dickerson, C., Thibodeau, R., Aronson, E., & Miller, D. (1992). Using cognitive dissonance to encourage water conservation. *Journal of Applied Social Psychology, 22,* 841–854.

Dovidio, J. F., & Fazio, R. H. (1992). New technologies for the direct and indirect assessment of attitudes. In J. M. Tanur (Ed.), *Questions about questions: Inquiries into the cognitive bases of surveys* (pp. 204–237). New York: Russel Sage Foundation.

Egbert, L. D., Battit, G. E., Tundorf, H., & Becker, H. K. (1963). The value of the preoperative visit by an anesthetist. *Journal of the American Medical Association, 185,* 553–555.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

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.

Fazio, R. H., Jackson, J. R., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69,* 1013–1027.

Fiske, A. P., Kitayama, S., Markus, H. R., & Nisbett, R. E. (1998). The cultural matrix of social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol 2, pp. 915–981). New York: McGraw-Hill.

Fiske, S. T., & Ruscher, J. B. (1989). On-line processes in category-based and individuating impressions: Some basic principles and methodological reflections. In J. N. Bassili (Ed.), *On-line cognition in person perception* (pp. 141–173). Hillsdale, NJ: Erlbaum.

Freedman, J. L., & Fraser, S. C. (1966). Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology, 4,* 195–202.

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.

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology, 26,* 309–320.

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

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.

Glass, D., & Singer, J. (1972). *Urban stress.* New York: Academic Press.

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

Gross, A. E., & Fleming, I. (1982). 20 years of deception in social psychology. *Personality and Social Psychology Bulletin, 8,* 402–408.

Hamilton, D., & Rose, T. L. (1980). Illusory correlation and the maintenance of stereotypic beliefs. *Journal of Personality and Social Psychology, 39,* 832–845.

Hamilton, D. L., & Trolier, T. K. (1986). Stereotypes and stereotyping: An overview of the cognitive approach. In J. Dovodio & S. Gaertner (Eds.), *Prejudice, discrimination and racism* (pp. 127–164). New York: Academic Press.

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., & Bargh, J. A. (1992). Unconscious sources of subjectivity and suffering: Is consciousness the solution? In L. L. Martin & A. Tesser (Eds.), *The construction of social judgments* (pp. 67–103). Hillsdale, NJ: Erlbaum.

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.

Jacoby, L. L., Lindsay, S. D., & Toth, J. P. (1992). Unconscious influences revealed: Attention, awareness, and control. *American Psychologist, 47,* 802–809.

Janis, I. L (1958). *Psychological stress.* New York: Wiley.

Johnson, J. E., & Leventhal, H. (1974). Effects of accurate expectations and behavioral instructions on reactions during a noxious medical examination. *Journal of Personality and Social Psychology, 29,* 710–718.

Jones, E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin, 76,* 349–364.

Kanfer, F. H., & Seidner, M. L. (1973). Self control: Factors enhancing tolerance of noxious stimulation. *Journal of Personality and Social Psychology, 25,* 281–389.

Katzev, R., & Bachman, W. (1982). Effects of deferred payment and fare rate manipulations on urban bus ridership. *Journal of Applied Psychology, 67,* 83–88.

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol 1, pp. 233–265). New York: McGraw-Hill.

Kiesler, S. B., & Sproull, L. S. (1987). (Eds.). *Computing and change on campus.* New York: Cambridge.

Kihlstrom, J. F. (1987). The cognitive unconscious. *Science, 237,* 1445–1452.

Kunen, J. S. (1995, July 10). Teaching prisoners a lesson. *The New Yorker,* 34–39.

Landy, D., & Aronson, E. (1968). Liking for an evaluator as a function of his discernment. *Journal of Personality and Social Psychology, 9,* 133–141.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology, 32,* 311–328.

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.

Langer, E. J., & Rodin, J. (1976). The effects of choice and enhanced personal responsibility for the aged: A field ex-

periment in an institutional setting. *Journal of Personality and Social Psychology, 34,* 191–198.

Langer, E. J., Blank, A., & Chanowitz, B. (1978). The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *Journal of Personality and Social Psychology, 36,* 635–642.

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.

Lucasiewicz, M., & Sawyer, D. (1991, Sept. 26). *True Colors: Primetime Live.* New York: American Broadcasting Company. (Available from MTI Film & Video.)

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.

Mook, D. G., Wilson, T. D., & DePaulo, B. M. (1995). *Shortcomings of research on important social problems.* Unpublished manuscript, University of Virginia, Charlottesville.

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

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

Parker, S. D., Brewer, M. B., & Spencer, J. R. (1980). Natural disaster, perceived control, and attributions to fate. *Personality and Social Psychology Bulletin, 6,* 454–459.

Pennebaker, J. (1983). Physical symptoms and sensations: Psychological causes and correlates. In J. T. Cacioppo & R. E. Petty (Eds.), *Social psychophysiology: A sourcebook* (pp. 543–564). New York: Guilford.

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.

Piliavin, I. M., Rodin, J., & Piliavin, J. A. (1969). Good samaritanism: An underground phenomenon? *Journal of Personality and Social Psychology, 13,* 289–299.

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.

Rodin, J., & Langer, E. J. (1977). Long-term effects of a control-relevant intervention with the institutional aged. *Journal of Personality and Social Psychology, 35,* 897–902.

Rosenberg, M. J., Davidson, A. J., Chen, J., Judson, F. N., & Douglas, J. M. (1992). Barrier contraceptives and sexually transmitted diseases in women: A comparison of female-dependent methods and condoms. *American Journal of Public Health, 82,* 669–674.

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.

Sable, A. (1978). Deception in social science research: Is informed consent possible? *Hastings Center Report, 8,* 40–46.

Salovey, P., Mayer, J. D., & Rosenhan, D. L. (1991). Mood and helping: Mood as a motivator of helping and helping as a regulator of mood. In M. S. Clark (Ed.), *Prosocial Behavior: Review of Personality and Social Psychology* (Vol. 12, pp. 215–237). Newbury Park, CA: Sage.

Schachter, S. (1959). *The psychology of affiliation: Experimental studies of the sources of gregariousness.* Stanford: 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.

Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General, 122,* 166–183.

Schulz, R., & Hanusa, B. H. (1978). Long-term effects of control and predictability-enhancing interventions: Findings and ethical issues. *Journal of Personality and Social Psychology, 36,* 1202–1212.

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? *Personallity and Social Psychology Bulletin, 18,* 585–590.

Sherif, M., Harvey, O. J., White, J., Hood, W., & Sherif, C. (1961). *Intergroup conflict and cooperation: The robber's cave experiment.* Norman: University of Oklahoma, Institute of Intergroup Relations.

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

Sigall, H., Page, R. (1971). Current stereotypes: A little fading, a little faking. *Journal of Personality and Social Psychology, 18,* 247–255.

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.

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.

Stone, J., Aronson, E., Crain, A. L., Winslow, M. P., & Fried, C. B. (1994). Inducing hypocrisy as a means for encouraging young adults to use condoms. *Personality and Social Psychology Bulletin, 20,* 116–128.

Tajfel, H., Billig, M., Bundy, R., & Flament, C. (1971). Social categorization and intergroup behavior. *European Journal of Social Psychology, 1,* 149–178.

Tumin, M., Barton, P., & Burrus, B. (1958). Education, prejudice, and discrimination: A study in readiness for desegregation. *American Sociological Review, 23,* 41–49.

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 Press.

Underwood, B., Berenson, J., Berenson, R., Cheng, K., Wilson, D., Kulik, J., Moore, B., & Wenzel, G. (1977). Attention, negative affect, and altruism: An ecological validation. *Personality and Social Psychology Bulletin, 3,* 54–58.

von Hippel, W., Sekaquaptewa, D., & Vargas, P. (1995). On the role of encoding processes in stereotype maintenance. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 27, pp. 177–254). San Diego: Academic Press.

Wallace, J., & Sadalla, E. (1966). Behavioral consequences of transgression: I. The effects of social recognition. *Journal of Experimental Research in Personality, 1,* 187–194.

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

Webb, E. S., Campbell, D. T., Schwartz, R. D., Sechrest, L., & Grove, J. (1981). *Nonreactive measures in the social sciences.* Boston: Houghton Mifflin.

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.

Wilson, T. D. (1994). The proper protocol: Validity and completeness of verbal reports. *Psychological Science, 5,* 249–252.

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., Hodges, S. D., & LaFleur, S. J. (1995). Effects of introspecting about reasons: Inferring attitudes from accessible thoughts. *Journal of Personality and Social Psychology, 68,* 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.

Wilson, T. D., & Stone, J. I. (1985). Limitations of self-knowledge: More on telling more than we can know. In P. Shaver (Ed.), *Review of personality and social psychology* (Vol. 6, pp. 167–183). Beverly Hills, CA: Sage.

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.