Research Questions and Hypotheses

2

The Research Question

In this chapter, I examine the principal types of research questions, give examples of each, and review the design expectations that each type of question establishes. I then address the complex and controversial matter of causation and raise issues that must be resolved whenever the research question is of a causal nature. Finally, I discuss the importance of providing a definition of terms and formulating clear hypotheses.

Research can be regarded as a process of asking a question (or a related series of questions) and then initiating a systematic process to obtain valid answers to that question. In reporting the research, the question should be made clear to the reader. If it is not explicitly stated, the reader ought to be able to infer what the question is from reading the introductory material in the text. It is disconcerting to read research reports that reflect the author's apparent confusion about the precise question or questions that the research is meant to address. This sometimes leads to hypotheses that do not seem to be consistent with the question, is followed by procedures that do not test the hypotheses, and culminates in conclusions that may not be too closely connected to the original question. The question can be directly posed, or it can be incorporated in a statement of the problem. The statement of the problem

tells the reader the intent of the study and sets the stage for what is to follow.

There are several types of research questions. It is advantageous to understand the characteristics of these types and to be able to identify the type of question that is being asked in the study at hand. Different types of questions call for different approaches to seeking answers. To a large measure the type of question dictates the formal characteristics that are required of the research design. When an accomplished reader receives information about the type of question that is being addressed, it sets up expectations that either are met or are not met in the study.

Having identified the type of question, the astute reader soon knows whether the design is appropriate to a question of that type. The reader also can tell when a design is simply incapable of providing answers to the type of question that is asked in the research. When that happens, the most telling aspect of the critique will have already been written. An example of this kind of disparity would be a study whose question is formulated in causal terms, but whose design is organized to obtain a correlational relationship. The reader who perceives the discrepancy between the question and the design is on the alert for any erroneous conclusions about causes and effects.

No particular meaning should be attributed to the order of presentation of the questions in the following categorization of types of research questions.

TYPES OF RESEARCH QUESTIONS

Existence Questions

"Does *x* exist?" (where *x* is a thing, an attribute, a phenomenon, a behavior, an ability, a condition, a state of affairs, etc.)

Examples:

Are there radio transmissions from outer space?
Can neonates perceive color?
Is there such a thing as extrasensory perception (ESP)?

Answers to existence questions are important when the existence or nonexistence of something is controversial. It becomes particularly important when some theory rests on it. The following are questions that at one time or another have intrigued psychologists: Is there such a thing as the unconscious? Can animals use tools to solve problems? Can chimpanzees communicate by means of symbols? It is not necessary to show that the existence of something is generalized. Merely to prove that it is there would be sufficient. Thus, we would have to pro-

duce only a single horse who is clever enough to do arithmetic to force a major revision of current beliefs. Just this sort of an instance was the famous case of Clever Hans (Pfungst, 1911), who occupies a hallowed niche in the history of psychology. An impressive number of illustrious witnesses were deceived into believing in the horse's arithmetical prowess as they observed his ability to tap out correct answers to arithmetic problems with his hoof. The introduction of simple control procedures shattered the illusion and forced Clever Hans to revert in status from mathematical celebrity to beast of burden. All those who believe that this loss of status affected his self-esteem are probably also convinced that he could do arithmetic after all.

Answers to existence questions usually require careful scientific work and the application of scientific methods to the study of the evidence, whether that evidence consists of a single case ($N = 1$, as was the case with the inquiry into the abilities of Clever Hans) or a large N. The researcher must design the study in a way that systematically rules out rival explanations. Evaluation of the research rests on how comprehensively and effectively this has been accomplished.

In some other sciences more than in psychology, acts of discovery, rather than controlled experiments, serve to demonstrate the existence of things that were not known to exist before. Sometimes such discoveries are accidental, but usually they are the result of planned and informed searches by scientists who have a good idea of what they are looking for and where to look. Examples of such searches include astronomers who scan the sky for hitherto unknown heavenly bodies, archeologists and paleontologists who dig in likely places for particular things that they hope to find, and naturalists who search for discoveries where few informed individuals have searched before.

Questions of Description and Classification

Having established or accepted the affirmative answer to the question about whether *x* exists, the ensuing questions ask about the description and classification of *x*. "What is *x* like? To what extent does it exist? Is it variable or invariant? What are its characteristics? What are its limits? Is it unique or does it belong to a known class (taxonomy)? Is the description distinctive for this particular subclass?" Before the expedition to the moon, the composition of its surface was actively debated. Samples of rocks that were brought back proved that rocks, among other things, existed there. The rocks were then extensively and scientifically studied in an effort to describe and to classify them. This answered questions about whether they shared characteristics with earth rocks and whether they possessed any unique characteristics.

Examples:

What are the personality characteristics of adolescent anorexic girls?

What are the child-rearing practices of drug-addicted mothers?

Research answers to this kind of question usually call for more than simple description. They require (a) statements about the generality of the description to the subclass that the research sample represents and (b) statements about the uniqueness of the description of that subclass. When such statements are provided, description questions turn into descriptive–comparative questions. When such statements are not provided, it is impossible to know whether the description is distinctive to the subclass under study. Thus, the reader of a study designed to answer a question about the child-rearing practices of drug-addicted mothers would expect the researcher to show that (a) the sample from which the description is generated is truly representative of drug-addicted mothers, and (b) the observed method of child-rearing is unique or distinctive for this subclass of addicted mothers and does not also describe the child-rearing practices of mothers with other kinds of disorders (or, indeed, of mothers in general). Lacking statements about both commonality and distinctiveness, the description would be incomplete from a research point of view and potentially misleading. The information on which to base the statements would have to evolve from a research design that included these additional sources of data.

Survey research primarily describes and classifies. Much work with surveys is designed for the purpose of evaluating a designated program or answering an ad hoc question rather than for the purpose of producing generalizable knowledge. When surveys are designed to yield conclusive and generalizable knowledge, the reader has reason to expect a research design that assures that the description produced by the survey extends beyond the study sample to the subclass from which the sample was drawn and has reason to anticipate that the study offers proof that the description is distinctively characteristic of that subclass.

Questions of Composition

"What are the components that make up *x*?"

Examples:

What are the factors that make up intelligence?

What are the principal components of personality?

What are the main factors that make up self-esteem?

Answers to questions of composition call for an analysis or breakdown of a whole into its component parts. Conversely, the researcher may begin with a large number of small components, determine which ones hang together to make up an identifiable factor, and ascertain whether the various factors combine to form a larger construct such as in the above examples. This type of work is epitomized by such pioneers of factor analysis as Spearman (1927), Thurstone (1931), and Cattell (1952). Because factor analysis is a mathematical procedure, the reader expects that care has been exercised to assure the accuracy of an invariably large number of computations, that the sample will be large enough and representative enough for the procedures to be valid, that experimenter bias will play no part in the identification and naming of factors, and that the individual items in the pool are well-constructed and are comprehensive in their representation of the various aspects of the construct under investigation.

Relationship Questions

"Is there an association or relationship between *x* and *y*?"

Examples:

Is honesty related to socioeconomic status?
Are Rorschach human movement responses related to IQ?
Is there an association between college grades and study time?

In relationship questions, a second variable (*y*) is introduced. More complex questions of interrelations among several variables can be asked. Using multiple regression techniques, one can ask questions about whether several variables collectively predict some outcome and what the relative contribution of each is. The question may concern the explanation of the interrelationships or may ask whether the patterns of intercorrelations fit theoretical models. In these cases, one expects that the researcher has used valid and reliable measures, that the sample is representative and of sufficient size for the number of variables under investigation, that the computations are accurate, and that interpretations do not go beyond the data by making insupportable statements about causality.

Descriptive–Comparative Questions

"Is Group *x* different from Group *y*?"

Examples:

Are women more sensitive than men?
Are men more aggressive than women?
Do younger people have better memory than older people?

The descriptive–comparative question is an elaboration of the simple descriptive question. The reader can see right away that the researcher intends to compare two or more preexisting groups. The defining characteristic of the groups may be some organismic variable such as sex, age, or weight, or an attribute variable such as socioeconomic status or educational level. These are characteristics that the researcher can identify and describe but cannot influence in the ways that are possible with experimental variables. When examining the effort of the researcher to determine whether women are more sensitive than men and to assure that their sex accounts for the difference, the reader anticipates that the women and men under comparison are equivalent in as many ways as possible other than their sex. The reader also expects that if any conclusions are drawn about women as distinct from men, the sample of men and women studied represent the general population. Expectations about the validity and reliability of the criteria measures obviously hold for this as for the other types of questions.

Causality Questions

"Does *x* cause, lead to, or prevent changes in *y*?"

Examples:

Does psychotherapy change behavior?
Does watching violent TV make children more aggressive?
Does smoking marijuana cause underachievement?

Variants of causality questions may be left open at either end: "What is/are the consequence(s) of *x*?" or "What is/are the cause(s) of *y*?" Sometimes research of this type is exploratory, but usually the investigator makes informed guesses, which focus the research and become the hypotheses.

Causality questions call for experimental research in which the experimenter manipulates the independent variable to provide the hypothesized cause or uses one that has been manipulated by nature or circumstances; the experimenter then contrasts the consequences to those observed under a no-treatment condition. Seeing that a causality question has been asked, the reader anticipates that the experimenter meticulously assigned individuals (usually randomly) to the treatment or no-treatment condition, controlled as many extraneous variables as possible so as to rule out anything else that might affect the results, applied valid treatments, controlled experimenter bias, used valid and reliable criterion measures, and analyzed the data accurately.

When causality studies are done in the form of single-case designs, the reader should expect to find care in the application and timing of

the experimental conditions, accuracy in measuring or judging the individual's behavior at baseline and when experiencing the experimental treatment, freedom from bias when analyzing and interpreting the results, and (one hopes) either enough replications to warrant generalizations or disavowal of any claims about generality.

Causality–Comparative Questions

"Does *x* cause more change in *y* than does *z*?"

Examples:

Is counseling better than group activity at preventing delinquency?

Are antidepressant drugs more effective than psychotherapy or a placebo in decreasing depression?

Is behavior therapy more effective than client-centered therapy in eliminating phobias?

Causality questions are comparative in the sense that the effects of *x* must be compared with non-*x* (the absence of *x*). In causality–comparative research, the effects of *x* are compared with a rival treatment and not simply with the absence of the experimental treatment. All of the things that are expected of research on causality questions apply to causality–comparative questions, but with the additional provision that the rival treatment is valid and is given in an unbiased manner.

Causality–Comparative Interaction Questions

"Does *x* cause more change in *y* than does *z* under certain conditions but not under other conditions?"

Examples:

Does counseling prevent delinquency more than do group activities in girls but not in boys?

Is behavior therapy more effective in eliminating phobias in adolescents than is client-centered therapy, but less effective with adults?

Is a certain medication more effective than psychotherapy in treating endogenous depression, but less effective in treating reactive depression?

As can be seen, causality–comparative interaction questions are just elaborations of causality–comparative questions. The addition of one or more independent variables enables the researcher to determine whether these variables interact with the first independent variable and with each other. The variables can be preexisting characteristics of the

participants, environmental conditions, time or order factors, and so on. The reader expects the same things as for causality–comparative research, with additional attention paid to the careful application of the added independent variables.

CAUSATION

Having phrased some of the research questions in terms of cause and effect, yet mindful of the debate over causation by philosophers, I must clarify how this term is used here and throughout the text. The philosopher Richard Taylor (1967) traced the history of the debate from Aristotle's concept of formal, material, efficient, and final cause through the views of Hume, Kant, and John Stuart Mill and up to the philosophers of the modern era such as Bertrand Russell.

Taylor noted that some philosophers consider the concept of cause to be worthless, anthropomorphic, and "replaceable by such less esoteric concepts as concomitant variation, invariable sequence, and so on" (p. 57). Russell saw no difference between cause and effect and called it a "relic of a bygone age" (p. 57). Defending the use of the concept of cause, Taylor stated:

> nevertheless, it is hardly disputable that the idea of causation is not only indispensable in the common affairs of life but in all applied science as well. . . . It is true that the concept of causation is a theoretically difficult one, beset with many problems and the source of much metaphysical controversy, but the suggestion that it can be dispensed with is extreme. (p. 57)

One of the unresolved issues is the concept of *power,* or causal efficacy, and whether it is essential. David Hume argued that instead of trying "to explain changes in terms of causes having the power to produce them," changes should instead be regarded as "invariably conjoined with others" (p. 58). Other philosophers stressed the importance for causation of voluntary actions by an agent. Factors with the power to bring about or prevent effects are referred to as *levers.*

Another unresolved issue is that of necessity of causes as against invariable sequence. Laws of nature are seen as necessities in one view but only as uniformities in another view. John Stuart Mill argued that unconditionality of connection is required for the connection to be considered causal. The conceptual requirement that connection had to hold in all imaginable circumstances severely limits the idea of causation. Some philosophers, according to Taylor, "reserve the expression 'the cause' for some causal condition of an event that is conspicuous or novel or, particularly, one that is within someone's control" (p. 63).

On the issue of a cause as a sufficient condition, Taylor stated that it is widely held that

> *a causal condition* of an event is any *sine qua non* condition under which that event occurred or any condition which was such that, had the condition in question not obtained, that event (its effect) would not have occurred, and *the cause* of the event is the totality of those conditions. (Taylor, 1967, p. 63)

Inasmuch as it is generally accepted that causes should always occur before their effects, a clause about the temporal sequence should be added to this definition.

In view of what is generally regarded as the plurality of causes, there is a distinction between talking about *the* cause and *a* cause. Mill maintained that "the cause of an event is a whole set of conditions, as we have 'no right to give the name cause to one of them, exclusively of the others'" (p. 63).

Taylor concluded his review by acknowledging, "From the foregoing considerations it is apparent that some of the main philosophical problems of causation do not yield to any easy solution. . . . Here, then, as in so many areas of philosophy, our advances over our predecessors appear more illusory than real" (p. 66).

Psychological theorists have not stayed completely away from the issue. Egon Brunswik (1943/1951) stated, "If we are not to forget the teachings of Hume and John Stuart Mill, we must realize that there is nothing observed but concomitant variation—of greater or lesser relative frequency—and that all analysis of causal textures rests upon this foundation" (p. 202). However, Clark Hull (1943/1951) argued that "the outcome of a dynamic situation depends upon 1) a set of antecedent conditions and 2) one or more rules of laws according to which, given a certain period of time, these conditions evolve into different conditions or events " (p. 204).

In common usage, the second edition of *Webster's* (Neilson, 1954) defines *cause* as "that which produces an effect; it is that without which the result would not have been" (p. 427). Here we have a definition in terms of necessary and sufficient conditions. The distinction between necessary and sufficient conditions can be illustrated by a Sufi parable cited by Shulevitz (1996) in her review of a book by Adam Phillips: "A man is standing in his yard throwing corn. A passer-by asks him why, and he replies, 'Because it keeps the tigers away.' 'But there aren't any tigers here,' the passer-by protests. 'Well, it works then, doesn't it?'" (p. 10).

Webster's also differentiates between a *cause* and a *reason*. A reason is defined as "that which explains or justifies a result" (p. 427). The example is given of the cause of a railroad accident being failure of air

brakes, whereas the reason is carelessness in inspecting the apparatus. The text acknowledges that what is a cause for one person may be considered a reason by someone else.

Plurality of Causes in Research

Psychologists have long recognized the plurality of causes (e.g., Taylor, 1967). Research psychologists think more in terms of one of the causes instead of a single cause. Psychological researchers appear to accept the idea of contributory causes rather than unitary causes. They write about experimental conditions or treatments as causal under the condition that all other potential causes are held constant (i.e., kept equal); they concern themselves with moderating and mediating variables and unmeasured intervening variables that are affected by the treatment and that in turn bring about changes within the organism. Researchers can and do limit their definition to what actually occurs in the experiment. This is the "lever" that the researcher uses while trying to hold everything else constant. The investigator controls the conditions and introduces them in advance of the presumed effects, meeting the requirements of temporal sequence. With these conditions met, the investigator uses the concept of causation advisedly; its use is reserved for experimental studies.

Working Definition of Cause

To illustrate some of the problems in pinning causation down, I present the following allegory, which is intentionally as anthropomorphic as it can be. During a heated argument, John takes out a pistol and fatally shoots Jack. The coroner's report states the cause of death as "Gunshot wound caused by 8-mm bullet that tore into the carotid artery causing massive bleeding that deprived the heart and brain of oxygen until they ceased to function." The heart says, "That artery was the cause of it. It stopped feeding me so that I could no longer do my job." The artery says, "It wasn't my fault, that bullet tore into me so that I could no longer function." "Don't put the blame on me," says the bullet, "I was resting peacefully in my chamber when the powder behind me exploded and sent me headlong into the artery." "I'm not the cause," says the powder, "I was quietly sitting there when that firing pin slammed into the casing and caused me to explode." "I wouldn't have moved if that finger hadn't pulled the trigger which released the spring that drove me forward into the shell casing. John is the cause. His finger pulled the trigger, and he is the only one in this episode who has free will," says the firing pin. John admits that he pulled the trigger intentionally, but says, "I'm not to

blame because, as my therapist explained to me, I am a very angry person and have difficulty controlling my impulses. The cause is that my father, who was an alcoholic, left home when I was 3 years old, and my mother had to raise me all alone. She had to work as a prostitute and I did not even have my own name; all of her customers were called John too." John's father, if asked, would have his own version of the causes of the fatal shooting, and his account would also generously confuse causes with reasons. The trail of reasons could be followed backwards in infinite regress. When designing an experiment, there is no doubt about who is the causal agent. The experimenter sets up the conditions in advance with causal intent and tries to control all other conditions.

A nonphilosophical working definition of a *cause* for most scientific research, then, is as follows: "A proximal antecedent agent or agency that initiates a sequence of events that are necessary and sufficient in bringing about the observed effects." It is called "a" proximal antecedent in recognition of the fact that it may not be the only one. It is "proximal" because it is introduced at the time, and it did not occur so long ago that all of the things that happened in between could compromise an interpretation of causation. As an "antecedent" it clearly precedes the effect. The "agent or agency" is set up intentionally to be the experimenter and the treatment. The experimenter exercises the power and controls the lever as indicated by the term *initiates*. When successful, the experimental treatment is sufficient in that the effect does come about in the experimental group. It is necessary in that the effect is not seen in the absence of the treatment in the control group. When the definition is applied to the case of John, his pulling the trigger would be considered as the cause, because he was the proximal antecedent agent who initiated the chain of events that were necessary and sufficient to bring about Jack's death.

Unsubstantiated Claims of Causation

Somewhat easier than making assertive claims of causation is the ability to dismiss false claims of causation from studies that do not provide evidence for it. In critiquing research the emphasis is not so much on proof of cause as it is on false claims of causation. Research is used to inform the public as well as professionals, to guide them in the formation of their attitudes and opinions, and to influence social action. Investigators bear a social responsibility to do valid research and to report it accurately. Journalists and science writers who reconstruct and transmit the findings to the public have the responsibility of understanding research design well enough to summarize it and to interpret it accurately.

One has only to read almost daily news reports of scientific studies to see evidence of confusion between relationship and causation and the acceptance of the false notion that if one event preceded another it must have caused it to happen (*post hoc ergo propter hoc,* which translates as "after this, therefore because of this"). Reports of nonexperimental medical studies have become routine. One day the public is informed that the rate of arteriosclerosis is comparatively low in countries where there is high wine consumption and that therefore wine prevents arteriosclerosis. The next day an article explains that rice consumption prevents heart attacks in countries where rice consumption is high. A correlation is presented between broccoli consumption and low colon cancer rate. Thousands of people are influenced to protect themselves from all three ailments by gorging themselves on diets of rice and broccoli in wine sauce. Epidemiological correlational studies are of interest, but they give people a false idea about causality if they are based exclusively on population statistics.

Social behavior is also affected by reports of psychological correlations. A recent newspaper opinion article (Maginnis, 1996) advocated that the Pentagon no longer authorize the sale of pornographic magazines in military stores. Cited in support of the recommended policy was a national study that reported a correlation between the reading of sexually explicit magazines like *Hustler* and *Playboy* and rape rates. Both of these magazines are sold in military stores. The clear implication is that reading this material causes soldiers to rape. No consideration was given to the possibility that young men who are inclined to commit rape might also be inclined to read pornography or that rapists read more magazines of all types than do nonrapists. Whether or not anyone likes these rival explanations, they are possibilities that have not been ruled out. The mistake is to think of correlation as definitive evidence of causation.

A frequent logical error, *post hoc ergo propter hoc* is based on the correct antecedent–consequence time sequence, but the "necessary" condition is not demonstrated. It is a fallacy to think that a happening that follows another must be the result of it. In a sociological investigation, D. P. Phillips (1983) studied all homicides in the United States between 1973 and 1978. During this period, 18 heavyweight championship prize fights were held. Any homicide that occurred within 3 weeks after a match was counted as a consequence. He concluded that fatal, aggressive behavior was stimulated by heavyweight prize fights. Although the prize fights were antecedents, there is no evidence that they were necessary antecedents of the homicides that followed. It is not even known whether the assailants knew about the fights or had anything more than a casual interest in them.

In another study, D. P. Phillips (1978) reported that a significant increase in small-plane crashes followed highly publicized murder–suicides. This led him to speculate that at least some of the crashes were intentional and were stimulated by the murder–suicides. No evidence other than *post hoc ergo propter hoc* was offered.

Distinction Between Enabling *and* Causing

An *enabler* is a state or condition that permits something to happen but is not the cause of it. A doctoral degree is one of the things that enables a person to obtain a license to practice psychology, but the degree is not the cause. Proof that a condition is an enabler comes from successfully demonstrating the event when the condition is present but being unable to demonstrate it in the absence of the condition. The condition is not an enabler if the event can be demonstrated to occur in the absence of the condition or if it can be shown that the event does not happen despite the presence of the condition. The doctoral degree is necessary, but it is not a sufficient condition for licensure. Some enablers are neither necessary nor sufficient. For example, a high IQ enables a person to earn a good living. In such cases, proof becomes probabilistic, and we have to establish that the event occurs significantly more often when the enabler is present than when it is absent. More simply put, certain conditions enable things to happen that probably would not happen otherwise.

The egg-balancing ritual is a good example of belief in an enabler gone awry. A demonstration was organized in 1983 (Gardner, 1996) to show that fresh eggs could be balanced on their broad ends at 21 minutes before midnight of March 20, the time of the vernal equinox. The rationale was that the sun is directly over the equator; the length of day and night are equal; and everything in the world is in such perfect balance, peace, and harmony that even eggs can be stood on end at this moment. In an urban park in Manhattan, hundreds of eggs were distributed to believers who had gathered for the event. At the crucial moment, people succeeded in balancing eggs on end throughout the small park. This confirmed, for participants, that it was the timing that enabled it to happen. The independent variable here is the time of the attempt. There was no control condition. As a matter of fact, eggs can be balanced on any other day of the year as well. A rough concrete surface makes it much easier, but it can be done by a steady hand on relatively smooth surfaces. The original Chinese egg-balancing celebrations of Li Chun were usually held around February 4 or 5, the date of the onset of their lunar spring. Time was coincidental in both the American and Chinese demonstrations.

Readers of research studies should have a clear notion of what is required to make a valid claim that one thing either enables or causes another to happen.

Correlation and Causation

Correlations between two variables show the relationship or association between them and do not imply that one is the cause of the other. Multiple correlations show the associations between two or more independent or predictor variables and a dependent or outcome variable. One cannot say that the former cause the latter to happen. With the development of path analysis, an application of multiple regression, one can begin to make causal inferences and to construct causal models. The inferences can be of great theoretical interest, but the caution of Tabachnick and Fidell (1989) remains cogent:

> Demonstration of causality is a logical and experimental, rather than statistical, problem. An apparently strong relationship between variables could stem from many sources, including the influence of other, currently unmeasured variables. One can make an airtight case for causal relationship among variables only by showing that manipulation of some of them is followed inexorably by change in others when all other variables are controlled. (pp. 127–128)

DEFINITION OF TERMS

To understand the research question and the details of the study, one must clearly comprehend the meaning of the terms that are used. Scientists are forever coming up with new concepts that do not even have a name. Vocabularies have to be invented so that ideas can be communicated. When new terms are used, or old terms are used in new or distinctive ways, or general terms are used in specific or restricted ways, readers rely on authors to define those terms at the outset so as to increase the intelligibility of the report and to avoid any ambiguity or misunderstanding. Sometimes the word is named after a person (*Watt, Ampere*); sometimes a word is formed from a Latin or Greek root (*dementia praecox, schizophrenia*); sometimes words are combinations of smaller descriptive words as is typical of the German language (*Unterschiedsempfindlichkeit*); sometimes they just have a catchy ring to them (*quark*); and sometimes they are just apt descriptions of meaning (*self-esteem*). When the term is new, as all of them were at one time, it has to be defined to be understood. If this study is about psychotherapy, for instance, the author should say, "*Psychotherapy* in this study refers to short-term cognitive treatment." The terms *short-term* and *cognitive* would also have to be defined. Terms

that have standard meanings and are used in standard ways need not be defined.

Terms that are being used in a specific way for a particular study can be defined operationally. As Stern and Kalof (1996) pointed out, "The first requirement for observations to have scientific value is that *abstractions be concretized*" (p. 12). Words such as *intelligence* and *weight* are very difficult to define, even though both of these words are in everyday usage. Defining *intelligence* in operational terms would inform the reader: "For the purposes of this study, *intelligence* is the score received on the full-scale Wechsler Adult Intelligence Scale administered by a qualified and experienced examiner." This is a poor conceptual definition but serves its purpose as a definition in terms of operations.

Hypotheses

The experimental hypotheses (H_1, H_2, H_3, etc.) are predictive statements about the expected outcome of the research. They call for a test and they embed a conclusion. The hypotheses dictate the method and design of the research and give the reader a fairly good idea about what the design will have to look like.

Explicit statements of experimental hypotheses are de rigeur in dissertations but are often omitted in more succinctly written journal articles. When this is the case, the summary of the theoretical basis for the study, the overview of research literature that preceded it, the synthesis of these materials, the statement of the problem, and the reason for the research to follow should leave little doubt in the reader's mind about what the author's predictions about the outcome are (assuming that there are predictions).

When comparisons are predicted, they have to be explicated. It is analogous to truth in advertising when an ad reads "20% less fat!". Does this mean 20% less fat than the last version of this product, 20% less than it used to have 10 years ago, 20% less than the average of other brands, or 20% less fat than protein? In a study that has boys and girls doing mental tasks under distraction conditions, the hypothesis states, "Girls will score higher under the distraction condition." Does this mean that girls score higher when distracted than boys? Or does it mean that girls score higher when distracted than when not distracted? The comparative prediction "higher than . . . " requires an object for it to acquire meaning. In distinction to the null hypothesis (H_0), experimental hypotheses (H_1, H_2, H_3, etc.) take a stand. They predict, for example, that groups assigned to different treatments perform differently (two-tailed prediction), and they may predict what direc-

tion the expected differences take (one-tailed prediction). This is in contrast to the statistical hypotheses, which are null hypotheses. Support of an experimental hypothesis requires the rejection of the null hypotheses at some acceptable level of confidence.

Some researchers state experimental hypotheses in the form of null hypotheses although they really expect to reject the null hypothesis. As Kerlinger (1986) observed, "Researchers sometimes unwittingly use null hypotheses as substantive hypotheses" (p. 190). This actuates a dilemma for the researcher, who, in the role of a skeptic or iconoclast, hopes to be able to falsify a hypothesis that others believe to be true. A way of handling this dilemma is for the researcher to word the hypothesis in its popular directional form but then to predict that the hypothesis will be disconfirmed.

HYPOTHESES AND THEORY

Rosenthal and Rosnow (1991) called attention to two ways in which hypotheses differ from theories:

> First, a theory is like a large-scale map, with the different areas representing general principles and the connections between them being sets of logical rules. Hypotheses, on the other hand, are like small sectional maps, which focus only on specific areas glossed over by the larger maps. Second, hypotheses (being more focused) are more directly amenable to empirical confrontation. (p. 28)

Some hypotheses spring from experiential observation; they are not offspring of any formal theory. One cannot help wondering where they came from and where they lead. Will the results of a test of such a hypothesis merely be added to a collection of isolated and homeless facts, each of which has yet to be placed in a small chamber within the home of some theory? More fortunate hypotheses are deduced from theories and benefit when their broader and more general origins are explained to the reader. Readers are well-served when authors state "conceptual hypotheses" in these more general terms. These conceptual hypotheses, the summary of the literature on which they are based, and the relevant findings from earlier research form the foundation on which the new study is to be bolted. The experimental hypothesis can then be stated in its more highly specific way within this broader context.

STATING THE HYPOTHESIS

The generic hypothesis about concomitant variation ($Y = fX$) states, "Dependent variable Y is a direct (or inverse) function of independent

variable *X*." As *X* increases, *Y* increases; as *X* decreases, *Y* decreases. The inverse refers to *Y* increasing when *X* decreases to yield a negative correlation. In noncausal studies where only relationships or associations are predicted, the hypotheses are stated in these relationship terms and do not promise more than the design can deliver.

Comparative questions call for hypotheses such as, "Other things being equal, Group A will score higher on the *Y* (the dependent variable criterion measure) than will Group B." Usually the phrase "other things being equal" is assumed rather than stated. In studies where preexisting groups are compared, the assumption is often more of a hope than a reality. For cause–effect experiments, the hypothesis for the simplest cases are, "(Other things being equal) the mean score of the Experimental Group A will be higher (or lower) on the dependent variable criterion measure than will be the mean score of the untreated Control Group B," or "the mean score of the participants under Condition A will be higher (or lower) than their mean score under Condition B." The experimental challenge is to make "other things" as equal as possible. The hypothesis would become even clearer if the treatments, the participants, and time (the who, what, when, where) were specified (e.g., "At posttest, depressed outpatients who receive 20 sessions of individual psychotherapy will score significantly lower on the Depression scale of the Minnesota Multiphasic Personality Inventory [MMPI] than those who remain untreated").

Armed with this blueprint of the study, the reader knows that 20 sessions of psychotherapy are to be given to depressed outpatients and that their scores on the MMPI Depression Scale are to be compared with the scores of a group of depressed individuals who do not receive treatment. Expectations have been established. The reader can now look forward to seeing how well the house is built. At the end, when viewing the conclusions, the reader can look backwards to see how well they match the predictions that were made at the beginning.

CONSISTENCY

Consistency of the research question, the hypotheses, the design, the analysis, and the conclusions is something that the reader expects to observe in good research. If the investigator who stated the above hypothesis uses only a treated group and then correlates depression with time in treatment, the design is not consistent with the hypothesis and does not test it. Instead, the actual study addresses the question of whether there is an association between length of time in treatment and depression. This may be a worthwhile question, but it is not the one for which the reader was primed.

If there is more than one independent variable, additional main effect hypotheses (H_1, H_2, H_3, etc.) and hypotheses predicting significant interactions between independent variables are stated when anticipated. A research report is always suspect when the author concludes that the hypothesis was supported even though the hypothesis was not stated and cannot be inferred from the preliminary material. The reader's suspicions are compounded when noting that the conclusions are based on a one-tailed test of significance, which is predicated on the preexistence of a directional hypothesis.

Summary

In this chapter, I examined the different types of research questions: existence questions, questions of description and classification, questions of composition, questions of relationships, descriptive–comparative questions, causality questions, causality–comparative questions, and causality–comparative interaction questions. Particular attention has been paid to establishing causal connections and avoiding errors in logic that lead to false claims of cause and effect. Clear definitions of terms and unambiguous hypotheses give the reader an understanding of precisely what the research aims to accomplish; they set up expectations about how the study might be organized to test the predictions and answer the research question. Consistency of the question, the hypotheses, the design, the analysis, and the conclusions is critical.

Research Strategies and Variables

When readers know what the study is going to be about and what the predictions are, they should examine the selection of the independent variable and the investigator's choice of a number of important defining strategies. This chapter provides a detailed discussion of these choices and strategies. The issue of generalizability, which is partly dependent on the decisions that are made, also is discussed. The decisions involve the following:

1. Independent variable: Is the independent variable manipulated by the experimenter or naturally occurring? What levels of the independent variable were chosen?
2. Time sequencing: Is the study prospective or retrospective?
3. Actuality: Is the study real or is it simulated?
4. Setting: Is the study done in the laboratory or in the field?

Independent Variable

MANIPULATED VARIABLES

To appraise a study, one must have an understanding of the theoretical and logical basis for the research and the thesis that the author is presenting. If the "statement of

the problem" that accompanies the research question and the hypotheses are clear to the reader, identification of the independent and dependent variables easily follows. If the research takes the form of an experiment, one expects the experimenter to manipulate (i.e., to vary intentionally and systematically) the independent variable so that the effects of this manipulation on a dependent variable can be observed. In this arrangement, the manipulated independent variable is an experimental treatment that is clearly the antecedent on which the dependent variable, or consequence, depends. In a study of stress effects, for example, the experimenter introduces and varies the stress experienced by the participants and observes the consequences.

NATURALLY OCCURRING VARIABLES

The researcher may use an independent variable that is being manipulated by some real-life experience. For example, the researcher may study people who are stressed upon learning of a life-threatening illness or people who have just survived a natural disaster such as an earthquake or hurricane.

STATIC GROUP VARIABLES

The researcher may select participants from appropriate preexisting groups whose identifying characteristics constitute the independent variable. These static group variables cannot be manipulated by the experimenter, nor are the experiences naturally occurring. Instead, they are characteristics of people that can be used to identify their assorted group memberships. Included here are (a) *organismic variables* that are part of the individual's physical being such as sex, skin color, age, or weight; (b) *status variables* such as education, occupation, socioeconomic status, or marital status; and (c) *attribute variables* such as diagnosis, personality traits, or social behaviors.

These variables can be used in static group designs by selecting proper contrast groups. The static group variable becomes analogous to an experimental treatment, with the contrast group serving as the rival treatment control group. The analogy holds only if the groups are truly equivalent in all other ways. When the organismic, attribute, or status variable is the independent variable, comparison with an appropriate contrast groups allows one to determine whether there is anything distinctive about the target group. When the organismic, attribute, or status variable is set up as the dependent variable, in some cases one can make inferences about how the group acquired its characteristics. When there is a logical two-way association between the

independent and dependent variable, the thesis of the study tells which one is to be considered as the independent variable. For example, an investigator who is studying the relationship between weight and self-esteem may offer the thesis that being overweight lowers self-esteem. The independent variable would consist of an overweight group to be contrasted to a group of individuals of average weight. The dependent variable would be self-esteem. Another investigator may posit that people who have low self-esteem eat excessively and gain weight. For this study, weight is conceived to depend on self-esteem. Because self-esteem is the independent variable and weight is the dependent variable, the researcher selects a high self-esteem group and a low self-esteem group and measures their percentage overweight as the dependent variable criterion measure.

Risks of Causal Inferences

When preexisting static groups are used, causal inferences, though tempting, may be risky. Dependency of one variable on another cannot be verified. Under these circumstances the use of the term *dependent variable* is more of a convenience than it is an accurate descriptor. The best that one can say is that there is an association, a connection, a relationship, a correlation between two variables. From a mathematical point of view, the term on the left-hand side of an equation is the independent variable, and the one on the right-hand side is the dependent variable. Where there are hypotheses, and particularly ones with a temporal sequence that predict some outcome, the independent variable becomes a predictor variable that logically belongs on the left.

Concerned about causation in the weight/self-esteem problem cited above, a third researcher decides to make it into an experiment instead of a static group design. The first plan is to enlist a group of participants of average weight and, over a period of time, to fatten up a random half of them by means of a high-calorie diet. The design calls for measuring the self-esteem of both groups before and after this treatment. Upon further reflection and the influence of collegial counsel, the investigator realizes that there may be an ethical problem in doing something that could have negative consequences for some of the participants. The situation is therefore reversed, and a study is designed in which one group of overweight participants is placed on a low-calorie diet to lose weight. Participants in a randomly assigned control group of equal initial weight continue with their regular eating habits. Self-esteem is measured before and after treatment (as in the original plan). The hypothesis is that the self-esteem of the diet group will increase from pretest to posttest, whereas the self-esteem of the untreated group will not.

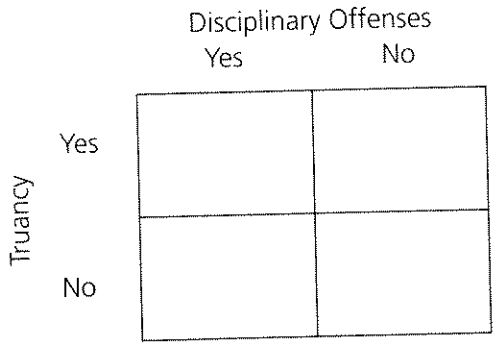
Still another investigator decides to approach the problem by making self-esteem into the independent variable and manipulating it. An experiment is designed in which the self-esteem of half of a sample of overweight participants is increased by giving them a series of success experiences. The remainder of the participants have no such experiences. The hypothesis is that the group whose self-esteem is raised will lose weight, whereas the other group will not.

These last two investigators each started out with the proposition that being overweight lowers self-esteem. They both succeeded in designing cause and effect experiments, but were these experiments consistent with their thesis? The first one wanted to show that increase in weight (independent variable) lowers self-esteem (dependent variable). Demonstrating that losing weight raises self-esteem does not prove the converse (i.e., that gaining weight lowers self-esteem). The second investigator reversed the independent and dependent variables. Demonstrating that raising self-esteem lowers weight does not prove that gaining weight lowers self-esteem. Logical errors of this sort can be found in examining the chain of reasoning used as the research proceeds from thesis, to hypothesis, to independent and dependent variable selection, to design, and to conclusions.

Unidirectional Paths

Contrast a thesis about weight and self-esteem with one about height and self-esteem. The investigator proposes that height leads to self-esteem and selects height as the independent variable and self-esteem as the dependent variable. The self-esteem of a group of tall people is contrasted with that of a group of short people. In this example, switching the independent and dependent variables, making height into the dependent variable, would be illogical. Few would believe that increasing people's self-esteem would actually make them grow significantly taller (as distinct, perhaps, from just standing taller).

One-way unidirectional paths are fixed by the logic of antecedents and consequences. It is reasonable to think that early childhood experiences could have a bearing on adult adjustment. It would be conceptually backwards to begin with a group of well-adjusted adults and a group of poorly adjusted adults as two levels of the independent variable and to make early childhood experiences into the dependent variable. Readers encountering this have to be puzzled unless the author clearly acknowledges that it is a retrospective study featuring postdiction instead of prediction. If so, the reader would be anticipating a discriminant analysis or logistic regression to test the hypothesis that early childhood experiences discriminate between well-adjusted and poorly adjusted adults.



Disciplinary Offenses and Truancy

Some studies have two or more contemporary organismic–attribute–status variables and one or more antecedent conditions or events. Take as an example the study conducted by a researcher who is interested in the association between family income during the first 12 years of children's life on school truancy and disciplinary problems in high school. The thinking is that low family income contributes to and predicts difficulties in school in the years that follow. The investigator sets up disciplinary offenses and truancy as the independent variables in the 2 × 2 analysis of variance design shown in Figure 1.

The dependent variable is family income during childhood, even though this is an antecedent and not a consequence. The concept of truancy and disciplinary problems interacting in high school to affect past family income is patently absurd. Here, too, the problem could be reframed: The investigator may ask whether past family income is a variable that discriminates between groups of high school students who display truancy and disciplinary problems and those who do not.

One-Way, Noncausal Enabling Relationships

Some studies use two attribute variables, neither of which can be manipulated and neither of which can be viewed as causal of the other, but where there is only one logical enabling path. IQ and income would be an example. One can set up IQ as the independent variable and predict income as an adult, but it would be illogical to conceptualize adult income as the independent variable and predict IQ from it. IQ can be conceived of as an enabler of income rather than a cause, but the idea that one's income as an adult could enable IQ makes little sense.

Two-Way Sequential Causation

Some variables can affect each other in either direction and do so sequentially. Take as an example the causal relationship between success–failure and self-confidence. Failure can lead to lowered self-confidence. This, in turn, can lead to more failure. Similarly, success can lead to increased self-confidence, which in turn can lead to more successes in a continuing sequential loop. These sequences can be seen in baseball players when they go deeper and deeper into a batting slump, or when they are on a hitting binge and their self-confidence is palpable.

If this type of problem were to be approached by using static groups, the reader would be automatically on guard against insupportable statements about causality. For example, a researcher may measure the self-esteem of a group of successful salesmen and a group of unsuccessful salesmen and find that the latter have lower self-esteem. The researcher could not conclude from this that low self-esteem was caused by occupational failure, because low self-esteem may be a cause of occupational ineptitude. To establish causation, the researcher would have to manipulate one or the other of these variables in a controlled experiment.

ESTABLISHING LEVELS OF THE INDEPENDENT VARIABLE

The first decision facing a researcher is whether to make the independent variable continuous or categorical. If it consists of continuous data, the researcher must decide whether to leave it that way and treat the full range of numerical values, to make it dichotomous (i.e., low IQ–high IQ), to arrange it in graduated multiple levels (low IQ–medium IQ–high IQ), or to make the levels into discontinuous groups (70–80, 95–105, 120–130). Researchers are reluctant to tamper with a continuous variable by transforming it into dichotomies or into categorical levels, because information is lost in the process. At issue is whether the investigator really has any interest in the information to be sacrificed and whether it is worth compromising the central focus of the research to preserve that information. The researcher may have a theoretical, empirical, or pragmatic reason for formulating hypotheses in categorical terms, such as first born–later born, mentally retarded–normal IQ, or rich–poor. Any one of these dichotomies may represent the central focus of the study. The interest may not extend to intermediary levels, and, in fact, their introduction might cloud the picture.

If the hypothesis is stated in categorical terms, it is appropriate to keep the independent variable consistent. On the other hand, if the

hypothesis is stated as a relationship that varies across a graduated range, then it is not appropriate to break the independent variable into dichotomies or nominal categories. The choice is driven by the theory and the rationale of the study, not by statistical considerations. The researcher cannot be faulted for setting up the independent variable in a way that permits a direct test of the hypothesis. A reader who disagrees with the hypothesis should find fault with it, not with a design that is appropriate to test that hypothesis. All of this applies to independent variables that are inherently continuous. There is no choice with variables that are inherently categorical such as male–female or blue eyes–brown eyes.

The reader, then, should examine how the levels of the independent variable were established to see whether they are consistent with the hypotheses. Consider the hypothesis, "Older people have more difficulty learning new things than do younger people." The age groups that the researcher selects to represent older people and younger people may be crucial in this study. There are several options.

1. Extreme groups. The researcher selects a group ≥ 80 years for the older group and ≤ 20 for the younger group. This would maximize the "effect," if there is any, but would minimize the amount of information about the relationship of age and learning because it is restricted to two extremes with no attention to anyone in between. As Gottsdanker (1978) cautioned, "Use of too few levels [of the independent variable] results in poorer representation of the relation between the independent and dependent variables" (p. 247).
2. Range of categories. A continuous range of age categories could consist of the following groupings: 20–39, 40–59, 60–79, and 80–99. A discontinuous range could consist of the following: 20–25, 40–45, 60–65, and 80–85. Both continuous and discontinuous age ranges give differential information across the adult life span. Neither gets at exactly what the above hypothesis predicts, because the hypothesis calls for a comparison of "younger people" and "older people." If the researcher defines the term *younger* as 20–25 and *older* as 80–85, the hypothesis predicts nothing about anybody in-between. On the other hand, if the hypothesis predicts a progressive decline throughout the life span, the use of the intermediate levels would be necessary.
3. Median split. The strategy of dividing groups along the median, though frequently used, presents two problems. Falling above or below the median does not necessarily place a person in a category that is consistent with the theory.

Being more than 50 years old is not necessarily "older" for the purposes of this particular hypothesis. A better example of the problem would be a sample split in the middle of the IQ range into a high IQ group (> 100) and a low IQ group (< 100). The expectation that the classification yields a highly intelligent group is vitiated by the presence of so many whose scores are only slightly higher than the median. The same holds for the classification of those who fall slightly below the median. An individual with an IQ of 101 is labeled *high IQ*, whereas one with an IQ of 99 is labeled *low IQ*. The 2-point difference is smaller than the standard error of the measure, and the two individuals are essentially indistinguishable on this variable. The placement of them in contrasting groups that are expected to perform differently on some other variable would work against the hypothesis. A more rational way to divide the groups would be to separate them by some standard deviational unit such as + 0.5 σ or + 1 σ. Researchers who wish to maximize the contrast by using extreme groups can use an interval of 2 σ or 3 σ.

Continuous Full-Range Distribution

With a continuous variable like age, as opposed to a categorical variable like sex, there may be an advantage in studying the question at all points rather than breaking the continuous variable down into arbitrary categories. The hypothesis informs the decision about the best way to proceed. Returning to the age and learning study, the researcher could treat age as a continuous variable and would obtain a progression of learning scores across the whole range of ages. This, however, is not how the hypothesis is worded. If in fact it is not until an advanced age is reached that there is a relationship between age and learning new things, linear correlation across the life span would not be very revealing. Suppose, for example, that in a study of visual acuity, presbyopic changes begin to show around the age of 40, after which there is steady decline. Correlation across life span is not linear. The graph of visual acuity is flat for the years before 40 and then begins to drop off. Curvilinear correlation or the use of categories like 30–35, 36–39, 40–43, 44–47, and 48–51 might show the picture more accurately.

Theory-Driven Levels

Instead of selecting levels arbitrarily, researchers are advised to use the theory that underlies the study as the guide for selection. Predictions

would come from the literature on age and learning. The age zone for the onset of learning deficits would be identified and then bracketed with other age-level categories for comparison. The hypothesis would be stated in a way that was consistent with the theory.

Strength of Independent Variable (Magnitude of Effect)

The strength of the independent variable is important. In the above example of extreme groups, the potential for effect was maximized. In a study of the effects of stress, use of a powerful stressor increases the magnitude of effect, but decreases generality because the findings of the study are limited to extreme conditions. A researcher who obtained predicted results with weaker stress conditions could probably safely assume that in most cases results would apply for greater stress as well. The results of such a study would therefore have greater generality.

Examination of the levels of the independent variable give the reader an idea about whether they match the hypothesis and whether the choice was meaningful and appropriate. The reader might ask whether the results would have been the same if other levels had been selected.

Time Sequencing

PROSPECTIVE

In prospective studies the researcher predicts consequences or effects from known antecedents or causes. This can be done with or without manipulating the independent variable. To illustrate the former condition, consider a medical researcher who places a random half of a group of people who have moderately high cholesterol on a cholesterol-lowering medication. The other half are given a placebo. Over time, the coronary illness rates of the two groups are compared. In a prospective study that does not involve manipulating the independent variable, the future coronary illness rates of a group of people who have moderately high cholesterol is compared with the rate for a group whose cholesterol is well within the normal range. In examining this study, which uses preexisting groups, the informed reader focuses attention on how the researcher attempts to control all of the other variables that could contribute to coronary illness. The reader knows that the researcher, who was not in control of the independent variable, has to be exceedingly cautious about claiming causality.

RETROSPECTIVE

In retrospective (ex post facto) studies, the researcher postdicts (i.e., tells backward) antecedents or causes from known consequences or effects. Obviously, the independent variable cannot be manipulated because what is under investigation has already happened. In the example that is being used, the best that the researcher can do is to go back in time and look at the cholesterol content of the premorbid diets of people who have coronary disease as compared with those who are free of this ailment.

Because the investigator has no control over the amount, duration, or timing of the levels of the antecedent; no control over the selection or assignment of participants to the antecedent conditions; and no control over other events, situations, and circumstances that could have a bearing on the dependent variable, the reader should be cautious when weighing the credibility of causal statements and causal inferences from retrospective studies.

LONGITUDINAL VERSUS CROSS-SECTIONAL

Another aspect of time sequencing applies to studies in which the passage of time is a factor. In a longitudinal study of human development over time, for instance, the independent variable, age, is a marker of levels of the passage of time. The same individuals are reassessed at specified age intervals. Structural and experiential events are intervening variables that operate during the passage of time to bring about observed changes. The study is prospective, but instead of manipulating the independent variable, the researcher just waits for it to happen.

If the strategy is to make the study cross-sectional, the researcher selects representative samples of children who have already attained different age levels but are equivalent in other respects. The researcher assumes that they have all passed through early levels and have had similar kinds of relevant life experiences before reaching their present age. The cross-sectional strategy is much quicker and more feasible to carry out, but there is more chance for error. Individuals at different age levels are not the same individuals as those at other levels and may not be as equivalent in important respects as the investigator would like. Assumptions about their intervening growth and experience may not be entirely justified. For example, a researcher who is doing a cross-sectional study of the physical development of children from ages 1 to 4 should not choose a cohort of 1- and 2-year-olds who differ from the older children in socioeconomic status (e.g., drawing

younger children from economically disadvantaged neighborhoods and with a history of substandard nutrition if these conditions are not shared by the older group of children).

Actuality

GENUINE SITUATIONS

In some studies, a real-life experience, with or without independent variable manipulation, is introduced or is found for the study. When possible, such studies can be powerful, but ethical considerations can prevent this strategy if any harm can come from it. For example, simulation would be required for a study on the effect of alcohol or marijuana on the in-flight performance of commercial airline pilots. In other instances, the genuine situation is abandoned because the time or cost is prohibitive or because real facilities are not available to the researcher.

ANALOG–SIMULATED SITUATIONS

Experimental arrangements can be set up to be analogous to real-life situations. They are "as-if" experiments, with simulation of time, place, persons, or situations. The study with the airline pilots could be done in a flight simulator, which realistically duplicates the genuine experience. That there are no consequences for crashing a simulated flight and that it might be sobering to be piloting a real airplane full of real passengers while "under the influence" do not override the necessity of an analog approach. In some psychotherapy research, the therapist, the patient, and the therapy are all simulated. At best, such studies can have implications for the real therapeutic enterprise. Some analog studies are of considerable interest and great ingenuity, but readers must scrutinize the kinds of claims and generalizations that are made.

Setting

Research studies can take place in the field or in the laboratory. Strictly speaking, the setting merely describes the venue of the study. In prac-

tice, some correlates of location may be important, because people may behave differently in different settings.

FIELD STUDIES

Field studies are meant to take place in their natural habitat. In field studies, chimpanzees are studied in the jungle as opposed to the zoo, educational studies are conducted in the classroom, and psychotherapy clients are studied in the therapist's office. These are frequently mislabeled *in vivo* studies, a term that refers to occurrences within the living organism as opposed to being isolated from the living organism and artificially maintained in a test tube (*in vitro* = under glass). As psychological experiments generally deal with living organisms, it is more appropriate to describe the setting as *in situ* (i.e., in its original place).

LABORATORY STUDIES

In laboratory experiments, participants are removed from their natural habitat and are brought into a special room that is used for research purposes. The setting does not necessarily make any difference, but it might under some circumstances. The setting, the actuality, and the use of manipulation of the independent variable are all orthogonal.

It is possible to have a therapy field study using real therapists in a simulated (also called *analog*) therapy situation with simulated films of patients or actors playing the part of patients. The film is stopped at strategic points and the therapist is asked to "respond" to the patient's remarks (see Strupp, 1955, for a prototype). This could be done equally well in the therapists' offices, or they could come to a laboratory to participate. One can also manipulate a condition such as therapist interventions, in a study of the effect of these interventions on clients' impressions of the value of the session. This could be done in the regular therapy context and setting with real therapist–client dyads. The real dyads may be brought into the laboratory, simulating the place and circumstances. An analog study could be set up with participants who are not real clients being interviewed by people who are not their therapists.

From the point of view of critical evaluation, one should recognize that it is as possible to do analog studies in the field as it is to make the field into a laboratory. One should also understand that it is possible to do careless and imprecise studies in the laboratory or in the field and precise cause–effect studies in either setting as well.

External Validity and Generalizability

The concept of *field* may refer to more than the literal setting (i.e., where the study took place). Questions can always be raised about the generalizability of the study: whether the research represent how things actually are and whether it reflects how things happen in the real world. Dialogues about the comparative merits of field and laboratory have been going on intermittently throughout this century. In discussing reaction time measurements outside of the laboratory, Woodworth (1938) stated:

> The automobile driver cannot hope to equal the short R.T. of the laboratory, because his preparation is not so good, he does not get a Ready signal two seconds before the emergency. He has to shift his own internal transmission when the stimulus arrives. A second or two must be allowed him for shifting his set and adjusting himself to the new situation which has arisen. If he is really startled several seconds may be needed. (p. 339)

Egon Brunswick (1955) introduced the term *ecological validity* in advocating the use of study participants and settings that are representative of the real world. Laboratory studies were subsequently challenged by many critics as lacking in ecological validity. Berkowitz and Donnerstein (1982) deplored the "widespread equation of experimental value with ecological validity" (p. 2). They asserted that laboratory experiments are designed to test causal hypotheses, not to determine how probable it is that they happen in particular situations. Berkowitz and Donnerstein insisted that laboratory experiments "give us a truer image of human complexity than do uncontrolled, naturalistic investigations" (p. 247).

Mook (1982) stated that when one "isolate[s] a single factor from that complexity [the natural environment] and var[ies] it independently of the rest of nature . . . the complexity of nature is taken directly in hand and discarded [in the laboratory]" (p. 126). This is not necessarily a liability; as Mook (1983) asserted, some experiments are designed for the purpose of testing generalizations instead of making them. Conducting studies in "unnatural" settings such as laboratories can be a virtue, according to Mook, if you are trying to find out whether something *can* happen. This is a valid point: If experiments are viewed as attempts to falsify hypotheses by putting them to the sternest test in the laboratory, credence is awarded to a hypothesis that holds up under conditions that are far more rigorous than in real life.

In a detailed treatment of the issues, Kerlinger (1986) listed the virtues of laboratory experiments as follows: (a) complete control is

possible, (b) independent variables can be manipulated, (c) random assignment can be done, (d) precision of measurement is possible, and (e) internal validity is high. Weaknesses include (a) lack of strength of independent variables, (b) artificiality, and (c) weak external validity. The virtues of field experiments are that they (a) are suitable to social and educational problems, (b) are subject to independent variable manipulation, and (c) are subject to random assignment. The principal weakness is that controls may not be as tight as one might like.

Ray (1993) added another dimension to the discussion: the difference between the scientist as observer (as is typical in field research that involves the use of naturalistic observations) and the scientist as participant (as is true in laboratory experiments). He suggested that there is less control over environmental factors but more ecological validity and fewer demand characteristics in field studies. Stern and Kalof (1996) stressed that naturalistic observations in field settings require complete and accurate recording of events as they occur with as little interference from the observer as possible.

Experience has shown that all field studies are not alike. For example, Seligman's (1996) differentiation between "efficacy studies" and "effectiveness studies" of psychotherapy reflect the difference between two levels of field study. One has tight controls that narrow the scope of the study, whereas the other portrays things the way they actually are.

> **Efficacy studies**, which are rigorously controlled studies of patients, usually with a single, well-defined disorder, randomly assigned to a fixed number of sessions, as prescribed by a treatment manual; and
>
> **Effectiveness studies**, which evaluate the benefits of treatment of multiple-disordered patients working without a guiding treatment manual, and with flexible duration of therapy—that is, therapy as it is really done. (p. 120)

Seligman suggested that the setting and approach of efficacy studies are most suitable for predicting what short-term treatments will work when they are applied in practice; effectiveness studies are most suitable for providing evidence of the effectiveness of long-term therapy. Both are field studies, but the first lends itself to tighter controls than would be feasible with the second. On seeing that the first approach has been used, the reader is attuned to the issue of whether the same results would hold outside of this particular experimental situation (external validity). When encountering the second approach, the reader is more concerned with whether the results are truly attributable to the experimental treatment (internal validity). Fortunately,

the choice available to the researcher is not necessarily dichotomous. Studies can be designed that close the major potential loopholes while still preserving the real-life components.

GENERALIZABILITY

For Campbell and Stanley (1963), *external validity* and *generalizability* are synonymous. Mook (1983) distinquished between them: "To what populations, settings, and so on, do we want the effects to be generalized? Do we want to generalize at all. . . . The question of external validity is not the same as the question of generalizability" (p. 379).

The statement of the problem gives the reader an idea about the purpose of the research and the intent of the investigator. It tells us whether the intent is to describe how things are in the real world; to determine whether a phenomenon or a relationship can exist under any circumstance; or to predict, to explain, or to test a theory. Knowledge of the purpose of the research enables the reader to judge whether the investigator *intends* to generalize and to decide from reading the study whether the researcher is *entitled* to generalize, and to what extent generalizations are justified. On the other side of the coin, a researcher cannot be faulted for not doing something that the reader wishes had been done. If there is no intent to generalize, the author does the right thing by not generalizing. Criticism is justified only when unfounded claims of generality are made.

Consider a study of the effect of diet on the mating behavior of Orca whales done under controlled conditions at Sea World. The researcher does not care whether the findings do not generalize to other populations or settings if the purpose of the study is to gather information that aids their captive breeding program. Generalizability of the data to Orcas in the wild is not the intent. In this example, the researcher is not hoping to generalize to other sea mammals or to nonmammalian marine creatures and does not expect that the findings extend to fish-eating pelicans or to vegetarian giraffes, far less to human beings. External validity is bounded by intent and by claims. Because there is no universal study that generalizes to everything, everywhere, and every time, the observation that the study has questionable external validity is itself questionable unless the author claims generality without furnishing the grounds for making such an assertion. If it is described as "external validity refers collectively to all of the dimensions of generality" (Kazdin, 1992, p. 25), all studies will be overburdened. Kazdin, however, went on to state:

> The task of the reviewer or the consumer of research (e.g., other professionals, lay persons) is to provide a plausible account of

> why the generality of the findings may be limited. Only further investigation can attest to whether the potential threats to external validity actually limit generality and truly make a theoretical or practical difference. (p. 34)

The importance of generalizability, and whether it makes a difference, depends on the nature of the research question and on the intent of the study. Generalizability is of no import if the study addresses an existence question. One would have to produce only a single chimpanzee who could act appropriately in response to printed symbols to show that it is possible.

Applied research that does not aspire to generalize cannot be faulted for not doing so. Studies can be designed to answer a question about a single setting, such as the case outcomes in a particular Mental Health Clinic, with no interest in the generality of the findings and no claims that the findings apply to anywhere else. On the contrary, most research that is designed to establish a principle is expected to be generalizable. Nobody would have much interest in such a study if those findings did not extend beyond the walls of that study.

An alternative to thinking of external validity as collective generality is to conceptualize it in study-specific terms (as one does with internal validity). One can view it as the demonstrated validity of the generalizations that the researcher intended the research to make at the outset and the validity of the generalized inferences that the researcher offers at the end. With this view, the universal challenge about faulty external validity that can be made about any study can be replaced by a focused appraisal of intentions achieved, and an assessment of generalized inferences can be drawn. In summary, an automatic criticism about the generalizability of the study is not especially fruitful. It does not matter whether the study lacks some unreachable kind of collective generalizability. Instead, the spotlight is on what, to whom, under what conditions, and how far you can extend the results beyond this single study. Some of the main aspects of generalizability are as follows:

1. Persons. Do the results apply to people who were not research participants but who share the same subclass memberships as the participants (i.e., diagnosis, age, sex, socioeconomic status, IQ, education, etc.)? Do they apply to people who belong to other subclasses as well?
2. Researchers. Would the same results be obtained with a different researcher, data collector, judge, or rater?
3. Places, Environments, Settings. Would the same results be obtained if the study were conducted in a different environment, place, or setting? If the study were done in a labora-

tory, would the results generalize to the natural environment (the issue of ecological validity)? Would it generalize to other geographical locations?

4. Time. Is there temporal generalizability? Would the findings that were obtained at one time apply as well at other times, or are the results time bound? This includes year, month, day, and hour and other historical eras (Goodwin, 1995).
5. Levels of Treatment. Do the findings pertaining to the levels of treatment used in this study apply to that treatment in general, or are the conclusions restricted to the kind, amount, intensity, and frequency of the treatment given in this research? For example, can general conclusions about the lasting benefits of psychotherapy for the treatment of severe anxiety of long duration be made on the basis of a study in which three sessions of behavior therapy were used?
6. Procedures, Conditions, and Measurement. Would one get the same answer to the question under different conditions and with different procedures and apparatus or with different methods of measurement? In other words, are the results generalizable beyond the specific ones used in this study?

It should be evident by now that at least some of these questions could be raised about almost any research. Some might be dismissed a priori in a given study. A researcher on memory loss in advanced Alzheimer's disease might see no reason why the study that was done in a New York nursing home should not hold as well in a New Orleans retirement home if the effects of the disease transcend geography or setting. Each of the other aspects of generalizability would have to be addressed in turn by the reader if not by the author. The reader should be sensitive to claims made by the researcher that go beyond what appears to be reasonable. Sometimes generalized conclusions are grossly overblown. A generalized claim that exercise elevates mood and self-esteem, based on the study of a few women in an aerobics class, would lead a discriminating reader to ask questions about the kinds and amount of exercise, and the characteristics of the women who are susceptible, before accepting the conclusion as a generalized fact.

REPRODUCIBILITY

The key to generalizability is whether the study can be reproduced or was a one-shot phenomenon that came about by an accidental confluence of participants and conditions that were uniquely accountable for the obtained results. A replication is an exact-as-possible repeat of

the same procedures (direct replication), but usually with different participants. A successful replication would demonstrate that the results were not a chance or sample-specific happening. Replication using identical procedures but a different experimenter and different participants would extend generality and would show that the results were not unique for the original researcher. Changing the setting as well as the participants and the experimenter (systematic replication) would add further generality. Successful replications on a broader sample of participants and under additional or altered conditions would further extend the generality.

For example, an original study demonstrates the successful use of a drug on a sample of people with tension headaches. It is then replicated successfully by another researcher on a different sample. Still another researcher obtains similar results on a sample of people with vascular headaches. Each replication makes claims of generality more credible.

Summary

This chapter focused on the independent variable and the establishment of independent variable levels. The advantages and disadvantages of using extreme groups, a range of continuous or discontinuous categories, median splits, or a full range distribution have been considered. The implications of adopting strategic choices about time sequencing (prospective or retrospective), actuality (real or simulated), and setting (laboratory or field) were elaborated. Generalizing the findings of a given study to other persons, researchers, settings, times, levels of treatment, or other procedures, conditions, or measurements must be done with caution.