Chapter 2

Comparing Group and Single-Case Designs

Throughout the behavioral and health sciences, correlational and experimental studies dominate the research design landscape. Although differing from one another both with respect to the ability to control relevant variables and in terms of the kinds of inferences supported by the method, correlational and experimental designs share one very important feature: Both tend to employ large numbers of subjects or research participants. For this reason, both kinds of research strategy can be referred to as **group designs**. In group designs, the analysis of data (group means and correlation coefficients) and conclusions, or inferences, drawn from the study occur at the level of the group, not individual participants. In this chapter, we consider the history and underlying logic of group designs as used in behavioral and health science. We also begin to consider the ways in which group designs differ from single-case designs, the latter being the major focus of this book.

GROUP DESIGN METHODOLOGY

To illustrate the research process that ordinarily characterizes group designs, we describe a research project conducted by clinical psychologists. Zabinski, Wilfley, Calfas, and Winzelberg (2004) were interested in whether an online psychoeducational intervention would be effective in treating women identified to be at risk for an eating disorder. They began by screening female college students for eating disorder risk; students who met diagnostic criteria for an eating disorder were not chosen for participation but were referred to appropriate mental health professionals. A total of 60 students were eventually identified to be at risk of developing a

disorder, and these students were randomly assigned to either a treatment group or a control group. Random assignment of participants to groups is important because it allows researchers to be fairly certain that the groups do not differ systematically on any important feature, such as intelligence, class rank, family background, and so on. The logic behind random assignment is that these other variables, which might impact the outcome or dependent variables, will be equally distributed across both groups. Thus, any difference between the groups on outcome measures should reflect only the effects of the independent variable, in this case the online intervention.

In many applied studies in health-related disciplines, clinical treatments serve as independent variables, factors intentionally manipulated or controlled by the researchers. The 30 participants in Zabinksi et al.'s (2004) treatment group took part in a psychoeducational program delivered on the Internet. These participants were provided with readings that covered important diet and other health-related topics, as well as weekly homework assignments built around this information, and they participated in a chat room discussion with other participants. The researchers anticipated that a combination of factual information and social support, the latter primarily generated through the chat room dialogue, would assist these students in reducing their risk of developing an eating disorder.

The second group of 30 participants served as a wait-list control group; that is, they did not receive the Internet intervention during the study, but they were eligible to receive the treatment after the study was terminated. To evaluate whether the Internet intervention was effective in reducing eating disorder risk, the researchers collected measures on several aspects of behavior and psychological functioning from both the experimental group and the control group at various stages during the study. Dependent variables measured during the study included self-report measures of behaviors predictive of eating disorders, self-esteem, and perceived social support. Because both groups contained many students, the researchers' first order of business was to find a way to summarize all of these numbers. This is ordinarily done through a simple mathematical calculation by adding all the scores together and dividing by the number of scores in the group. This procedure produces an arithmetic average, known as a **mean**. Then, because the researchers were interested in knowing whether the Internet treatment was effective in reducing eating disorder risk, the two groups' scores on the various dependent measures had to be compared. The answer to the study's question, as with so many studies carried out by behavioral scientists, comes down to this between-groups comparison.

Of course, we do not expect the two groups to have identical scores on any of the dependent measures (e.g., social support, self-esteem), any more than we would expect any two groups of people, chosen at random, to be the same on any particular characteristic, including intelligence, income, amount of formal education, and

so on—that is, we cannot expect the two group means to be equivalent, *even if the independent variable (Internet treatment) had no effect whatsoever!* What we want to know, then, is just how big a difference there is between the groups, and how likely this difference would be to occur merely by chance, even if study method had no effect. The formal machinery for answering this question developed within the behavioral sciences in the first few decades of this century and has long since become the standard for designing and analyzing experiments in these disciplines. The general methodological framework for conducting science in this way is often referred to as **null hypothesis significance testing,** or **NHST** research. A discussion of this research tradition can hardly ensue without introducing R. A. Fisher, whose scientific contributions to behavioral science are considered unparalleled.

FISHER AND CROP YIELDS

During the 1920s and 1930s, Sir Ronald Fisher, a scientist trained in biology and mathematics, became enthralled by the mathematical regularities that seemed to characterize much of the natural world. One of the most important of these regularities, observed by many scientists even before Fisher's time, was that certain characteristics distribute themselves in nature in a highly orderly and predictable manner. For instance, suppose we were to measure height, in inches, in a large sample of, say, just over 1,000 adult women. After measuring this variable in all of our participants, we might want to know the typical, or "average" height of participants in our sample. As we saw previously, we can acquire this basic kind of quantitative information by calculating a simple mean for the variable of height. The purpose of this measure is to summarize the information we obtained. We wish, in other words, to boil down a large number of observations to a reasonable and interpretable number. As a single number that we can use to represent an entire sample, the mean is useful for this purpose.

Such statistics are frequently reported in summarized presentations of information. Suppose you were intending to relocate to another city in a different part of the country. You would probably be interested in learning something about the community and region to which you were moving. If you were to contact the Chamber of Commerce for that city, it would probably supply you with abundant information, some of it in quantitative form, describing the community and local region. You might, for example, learn that the average temperature during the month of July is 85° and that the region receives an average yearly snowfall of 65 inches. You might also discover that the average income for residents of this city is $22,500 per year. Of course, you recognize that these are just representative numbers. They don't describe specific instances of these phenomena with precision,

only general tendencies. Nevertheless, the information helps you to familiarize yourself with some of the general features of the new region and perhaps helps you to prepare for the relocation.

Suppose, in the preceding example, we were to discover that the average woman in our sample stood 64 inches (5 feet, 4 inches) tall. This number clearly does not represent every person in the group. In fact, it is possible, because the mean is a mathematically derived number, that no single person in the group stands exactly 64 inches tall. What we would expect to find, however, is that a relatively large percentage of persons in our sample will be close to 64 inches in height, some slightly taller, some slightly shorter. In fact, most members of our sample will probably fall within a couple of inches above or below the mean of 64 inches. However, as we move further away from this mean height, we will encounter fewer members. In other words, although we expect that many women stand 65 or 66 inches tall, as we approach 70 inches (5 feet, 10 inches), we expect fewer. Once beyond 6 feet (72 inches), very few members of our sample should remain. The same is true, of course, on the other end of the spectrum. Although we would anticipate some of our group to stand less than 64 inches, the further below this mean we go, the fewer women we expect to encounter.

What we eventually end up with in this kind of exercise is a variable—in this case, adult female height—that distributes itself in a fairly symmetrical manner about a mean of 64 inches. This symmetrical distribution, characteristic of many features observed in nature, is referred to as a **normal curve,** or **normal distribution.** As you can see in Figure 2.1, the majority of participants in our sample fall close to the mean of 64 inches, and as we move further away from this mean on either side, the number and proportion of participants become fewer. This is an important feature of normal distributions because it allows us to determine how likely, or probable, any particular measure is in the distribution. For example, 65 inches is just one inch beyond our mean, and because it represents a small departure from the average height, we see this as a fairly likely outcome. Thus, being 65 inches tall is a fairly probable event, and we would expect perhaps several observations of women this tall. On the other hand, a height of 75 inches is markedly different from our mean. Consequently, the probability that any one member of our sample is 75 inches tall is very low, because this is an unlikely, though not impossible, occurrence.

One of the advantages of normally distributed variables is that we can calculate the likelihood, or probability, of obtaining any particular measurement from our sample. Obviously, the further away from the group mean a particular score is, the less probable its occurrence. Essentially the same logic is used in group research designs when comparing participants in experimental and control groups on some outcome or dependent measure. Although the details of doing so get somewhat more complex than this, we still end up utilizing the properties of the normal curve

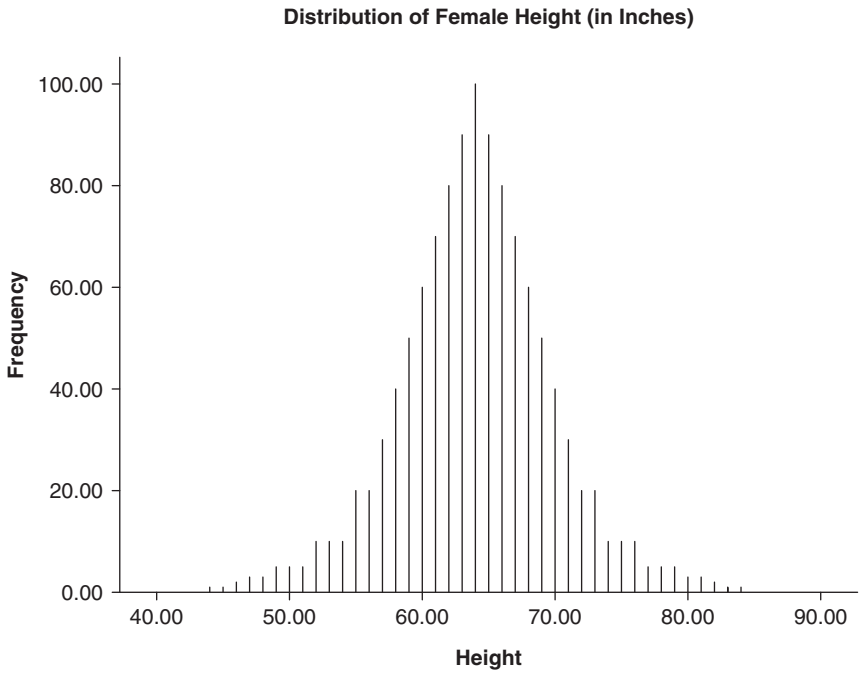

Figure 2.1 Normal curve of adult female height

to draw conclusions about the effects of independent variables. The larger the difference between group means, the less probable this outcome and the more convinced we are that the difference is due to our independent variable, as opposed to mere chance or random error. Thus, the normal curve has proved to be an immensely useful tool in the conduct of research in the behavioral sciences.

Among the first applications of the normal curve to scientific research was Fisher's application to agricultural science. Fisher and his colleagues devised several statistical tests, all founded on the basic properties of the normal curve, that allowed agriculturists to study the effects of numerous variables, such as soil characteristics, fertilization, watering schedule, and so on, on crop productivity. In such research, two crops containing the same plant (e.g., corn) would be grown under identical conditions with the exception of the independent variable of interest; that is, one crop might be treated with a new fertilizer while the other crop received no special fertilization at all. At some predetermined time, the crops would be compared with respect to yield. Of interest in such a comparison is the overall or total yield of a particular crop, not comparisons of individual plants. It makes sense that an agriculturist would benefit from information about how to increase crop yield through changes in fertilization, watering practices, and so on.

THE NULL HYPOTHESIS

Other researchers came to see the value of the statistical methods used in agricultural research and considered how they might be applied as well to the behavioral sciences. By the late 1930s, the formal research design and decision-making machinery introduced by Fisher had become a mainstay of psychology and related fields. Although today's researchers have at their disposal an array of statistical tools (particularly in the form of high-powered computer programs) that would have been the envy of Fisher and his contemporaries, the formal logic driving data analysis in the behavioral sciences has not changed since Fisher's revolutionary contributions.

Whether comparing crop yields—or, in our case, group means on a scholastic test—data analysis in the Fisherian tradition reduces to evaluation of the researcher's hypothesis about the subject matter. A **hypothesis** is a testable statement about the relation between variables. The purpose of conducting a formal research project in the Fisherian tradition is to assess whether one's formal hypothesis is tenable. In the case of the Internet eating disorder treatment study, the researchers' hypothesis suggests that measures of social support, self-esteem, and eating disorder risk will be related to the independent variable (treatment or no treatment) in some systematic way. We can, of course, propose as specific a hypothesis as we feel confident in making. For instance, perhaps the researchers believe that students who receive the Internet treatment will demonstrate higher self-esteem after treatment than those in the control group. We would therefore expect that the mean self-esteem score for the Internet treatment group would be higher than the mean self-esteem score for the control group.

In the Fisherian tradition, the results of the Internet treatment study will be evaluated not against any specific experimental hypothesis but against the **null hypothesis**, which suggests that there will be no difference between the groups on any of the dependent measures, such as self-esteem or eating disorder risk. In other words, the null hypothesis states that treatment will have no effect on the various outcome measures. Clearly, the researchers believe otherwise, or they would have no reason to conduct the study in the first place. Researchers, then, are in reality attempting to reject the null hypothesis, or show it to be false. In doing so, they provide support for the original research hypothesis, that Internet treatment will influence outcome or dependent measures.

In actuality, the null hypothesis doesn't really refer to what the researchers would expect to find among their specific sample of students. Instead, it refers to the entire population from which the sample has been selected, all college students. Researchers operate as if the study were actually being carried out on the complete population of college students, which would probably number in the hundreds of thousands. This is important, because the researchers obviously are

not conducting the study because they believe their results are meaningful only to their students. Instead, they think that online interventions may prove effective in reducing eating disorder risks in large numbers of people and that this knowledge may be very broadly applied. The null hypothesis states that, at the population level, there would be no difference between groups on the dependent variable as a function of treatment versus no treatment. In demonstrating that their sample means are in fact different, the researchers are in effect casting doubt on the validity of this null hypothesis.

STATISTICAL INFERENCE

Testing the null hypothesis is a formal procedure that allows us to state, with some degree of confidence, that our independent variable had an effect on the dependent variable. If this turns out to be the case, then we will use the quantitative measures obtained from our study sample to draw conclusions about the effects of the independent variable at the population level. The act of using sample data to infer certain characteristics of a population is referred to as **statistical inference**. So how are such decisions made? Remember that in an experiment such as the one on eating disorder risk, we would not necessarily expect the group means to be identical if the independent variable (Internet treatment) had no effect. What statistical decision making comes down to is determining how likely it is that any particular difference between means would have occurred by chance, that is, not as a result of the independent variable but simply as a result of measurement error and individual differences. Researchers use probability theory and knowledge of the normal curve to make such decisions. Statisticians have created numerical tables containing distributions of test statistics reflecting differences in group means and their specific probabilities of occurrence. If consultation of these tables reveals a difference between group means that would occur by chance less than 5% of the time, then most researchers consider this to be sufficient as a scientific criterion for drawing conclusions. Such a result allows the researcher to be confident that the difference in means obtained in the study was not due to random error or chance, that is, the difference is considered to be *statistically significant*. Finally, if the study was conducted in a methodologically sound manner, eliminating alternative explanations, then the most probable interpretation is that the difference was in fact brought about by the independent variable (Internet treatment).

Needless to say, the actual statistical mechanics of testing the null hypothesis entail more detail than this account describes. Such detail, however, is not the primary focus of this book. Just about any basic book on statistics found in your college library or local bookstore will have adequate coverage of this material. The

purpose of describing these features of group designs is to provide a benchmark against which single-subject designs can be compared.

INTERIM SUMMARY

Traditional group designs involve data collection from many subjects, often in the hundreds, and the use of inferential statistics to draw conclusions about group differences. This statistical strategy involves formal testing of the null hypothesis, which states that no differences exist between the groups. The purpose of statistical inference is to determine whether the null hypothesis can be rejected, thus indicating that differences between the groups were statistically significant. If the study was conducted according to strong methodological standards, then rejection of the null hypothesis leads to the conclusion that differences between groups were due to the independent variable.

SHORTCOMINGS OF GROUP DESIGNS

Group designs and statistical decision making have influenced the research process in the behavioral and health sciences for many decades, ever since Fisher introduced the logic of NHST. As a result, generations of social and behavioral scientists have cut their methodological teeth on the logic and tactics of null hypothesis testing. One unfortunate consequence of this fact, however, is that contemporary researchers tend to be somewhat myopic about research design, viewing NHST as the only way to conduct scientific research. Despite being a time-honored method of research in the behavioral sciences, NHST has its detractors, and research methodologists have directed increasing attention to the shortcomings of null hypothesis testing as well as the advantages of alternative data collection and analysis strategies (Cohen, 1990, 1994; Loftus, 1993, 1996). In fact, psychologists' discontent with NHST has reached such a fever pitch that a symposium was held at the annual convention of the American Psychological Association (Wilkinson & The Task Force on Statistical Inference, 1999) in 1996 to consider whether significance tests should be banned from their journals! No resolutions emerged from this provocative meeting, but a clear signal was sent that perhaps better methods might be available to behavioral researchers. Let's consider for a moment what all the fuss has been about.

The Meaninglessness of the Null Hypothesis

Many critics charge that the null hypothesis as a standard of comparison is of little value to researchers because it is almost invariably false anyway (Loftus,

1993, 1996; Meehl, 1978; Michael, 1974). Remember, the null hypothesis states that the population means, from which the researcher has obtained the observed group means, are equal. It is, in fact, nearly impossible that population means would be precisely equal, even if an independent variable had no effect on a dependent variable. Thus, we know that the null hypothesis is almost necessarily false, and this means that proving it to be so is not a very impressive scientific achievement. Indeed, psychologist Geoffrey Loftus (1996) claimed that rejecting the null hypothesis in most psychological research is like "rejecting the proposition that the moon is made of green cheese. The appropriate response would be 'Well, yes, okay . . . but so what?'" (p. 163). Moreover, a number of controllable factors influence how readily the null hypothesis can be proved false, the most salient of which is simply increasing sample sizes, which explains why traditional group researchers have historically paid so much attention to sample size.

You may recall, too, that rejecting the null hypothesis simply means that the difference we obtained between our group means is unlikely to be due to chance. Rejecting the null hypothesis *does not* tell us how large our effect is or whether this effect would make a difference in the real world. For this reason, critics of null hypothesis testing argue that we often end up asking the wrong question about our subject matter, especially when you consider that we pretty much know beforehand what the answer will be. Thus, Loftus (1996) suggested that the use of null hypothesis testing and statistical inference is "akin to trying to build a violin using a stone mallet and a chain saw. The tool-to-task fit is not very good, and, as a result, we wind up building a lot of poor quality violins" (p. 161). Some research methodologists believe that research in the behavioral sciences has placed too much emphasis on the goal of rejecting the null hypothesis at the expense of asking the more relevant question of "How big is the difference between group means, and what implications might this have for the real world?" Highly celebrated clinician and researcher Paul Meehl (1978) was perhaps most critical of psychology's dependence on significance tests and the null hypothesis testing tradition:

> I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology. (p. 817)

The Limited Information Value of Aggregated Data

Perhaps the most conspicuous characteristic of group designs is the fact that the primary data in such studies are dependent variables summarized across many subjects. As we have seen, the most common such measure is the group mean.

There is no question that a group mean, or *arithmetic average*, is easy to calculate and makes the business of evaluating many data points more manageable, but researchers often forget that sample statistics, such as group means, should not be used to predict or draw inferences about individual behavior. The group means calculated in such studies are interpreted as estimates of the population group means only. They are not, nor should they be taken as, indicative of the behavior of any one person, even a member of the group on which the mean was calculated. As we have already seen, means can be quite unrepresentative of the individuals who comprise the group, and this is especially true when groups exhibit considerable variability.

Unfortunately, the temptation to use group means as representative of individual behavior, or as some kind of comparison or standard measure, is irresistible. For example, a typical textbook in developmental psychology may state that the "average" infant is able to walk unsupported at approximately 14 months. This fact may be viewed, especially by anxious first-time parents, as a rigid criterion or temporal milestone that must be achieved by all normal infants. Deviation from this mean, which is actually quite common, may be an additional source of stress or worry to the parents. At such times, it may be helpful to remind ourselves that a mean is simply a mathematical shortcut or economical way of reducing the variability that characterizes all natural phenomena. The biologist Stephen Jay Gould (1996) suggested that scientists are often too quick to render their observations in manageable, summarized form, while in the process neglecting to appreciate the inherent complexity and variability of the natural world. As a case in point, Gould pointed to the domain of evolutionary theory, in which genetic variability, produced both by sexual recombination and mutation, plays an essential role. Genetic variability is the raw material on which natural selection operates, and in its absence the very process of evolution would be unthinkable. Nevertheless, recognizing the importance of variability required a historic shift in perception, according to Gould:

> What can be more discombobulating than a full inversion, or "grand flip," in our concept of reality: in Plato's world, variation is accidental, while essences record a higher reality; in Darwin's reversal, we value variation as a defining (and concrete earthly) reality, while averages (our closest operational approach to "essences") become mental abstractions. (p. 41)

Gould reminded us that mathematical averages are convenient abstractions that allow us to make sense of large amounts of data but that often distort or misrepresent the individual case. The world is full of rich variation that often goes unappreciated in our haste to reduce multiple observations to singular, discrete measures.

Limitations of Single or Few Measures of the Dependent Variable

Clearly, a defining feature of group designs is the use of large samples of subjects, sometimes numbering in the hundreds. This ordinarily means that it would be impractical to take more than one or two measures of the dependent variable from each subject, because doing so would take unreasonable amounts of time and effort. As a consequence, group researchers ordinarily measure the dependent variable, after manipulating the independent variable, one time in each experimental and control subject. Even this practice can prove Herculean, depending on the nature of the dependent variable and the size of the sample. Recall, though, that the dependent variable is most likely to be an outcome measure of behavior or psychological functioning. Although perhaps unintentionally, group researchers, by measuring the dependent variable once, are treating the variable as a discrete, singular phenomenon (much like Fisher's measure of crop yield).

However, the subject matters of the behavioral and health sciences are anything but discrete. Behavior and psychological and physical functioning are highly dynamic and time-dependent phenomena, continuously waxing and waning from one moment to the next. In other words, we must ask how well a continuous subject matter can be circumscribed by so few observations (usually only one or two). Imagine, if you will, walking into a busy preschool for the purpose of observing the interaction between the children and the teacher and to get a feel for the overall ambience of the environment. Suppose you were allotted only a brief period in which to observe, say, the first 15 minutes of the day. To your considerable dismay, what you witness is utter chaos, with children running about the room, discarding coats and hats here and there, singing, and playing tag, all the while oblivious of the teacher's instructions to sit down and be quiet. Having made your observation, you bid the teacher good-bye (and, perhaps, "Good luck"), and leave the room. The important question now becomes *How do you summarize your observation of the classroom? What conclusions will you draw about the teacher–child interaction and the climate in the classroom?*

Perhaps more important, let's consider the teacher's perspective on this observation. Is there any reason for the teacher to be concerned about the conclusions that you might draw from your observation? Of course! First, she might be rather disappointed that you showed up at the beginning of the day, shortly after both she and the children had arrived. This is ordinarily a very busy, often frantic time, when children are making the transition from the home environment to the school. Lunch boxes need to be put away, coats hung up, greetings made, and other sundry matters attended to before settling down into a daily routine. In other words, the first few minutes of the school day may be the absolute worst time to be observing, *if one is interested in obtaining a representative picture of the behavior of*

interest. The teacher no doubt would want you to hang around a bit longer to observe a larger sample of behavior and obtain a more representative picture of her interactions with the students and the overall classroom climate.

What this example suggests is that brief, discrete observations of behavior are likely to offer a distorted or, at best, incomplete picture of the subject matter. This is the case simply because behavior changes, sometimes quite dramatically, over time, and it is difficult to appreciate this natural ebb and flow if your measurement strategy does not allow for prolonged or repeated observations. The use of large numbers of subjects in group designs often precludes making multiple or long-term measures of the dependent variable. Were we not dealing with a continuous phenomenon that exhibits serial dependence and cyclic trends, this would not be problematic, but behavior, unlike crop yield, is a dynamic, continuous phenomenon whose natural dimensions seem to call out for continuous or repeated measurement. The structure of group designs ordinarily renders continuous measurement unmanageable.

Studying Numbers Rather Than People

Whatever else could be said about them, statistical significance tests are, in essence, devices used by researchers to help in decision making. We want to know whether an independent or treatment variable had an effect on some dependent or outcome variable. Much of the current controversy swirling around hypothesis testing and statistical inference has to do with whether these methods contribute positively to the business of scientific decision making. Jack Michael (1974) offered a very compelling argument that statistical techniques actually impair good decision making, in part because more intellectual effort is expended on evaluating the properties of the significance tests themselves than paying attention to the actual subject matter. Michael claimed that inordinate amounts of time are spent in professional and graduate education, training researchers in the intricacies of hypothesis testing, probability theory, and statistical significance, when this time could be spent putting the student into better contact with the actual phenomenon of interest. In addition, today's high-powered computer programs may offer a seductive alternative to the painstaking work of conceptualizing, operationalizing, observing, and measuring behavior. It is quite easy for today's scientists to get caught up in the bells and whistles of powerful and convenient statistical analysis while forgetting why the analysis is being conducted in the first place. Thus, some researchers may become more infatuated with the behavior of numbers than the behavior of organisms. The situation is analogous to the "gadget fanatic" who simply delights in any new technology, regardless of whether the technology proves helpful in completing a task or solving a problem. Much of this could be forgiven if statistical inference proved to be an indispensable tool for drawing conclusions in the behavioral

sciences. This is clearly not the case, however, as attested to by the ongoing dialogue among research methodologists.

The limitations of group designs and the NHST strategy are by now quite apparent to many researchers in psychology and the behavioral sciences. Nevertheless, group methodology remains predominant in the behavioral sciences because most researchers were trained in its use and continue to teach the methods to their students. Researchers, like everybody else, get used to doing things the way they have always done them, and breaking habits and exploring new ways of doing things is always a challenge. Avoiding this challenge would be understandable if, in fact, there existed no alternative research strategy to group designs and NHST. This is not the case, however, as we will soon see.

INTERIM SUMMARY

Despite dominating the behavioral science research landscape for more than six decades, group design and its attendant data analysis strategies possess serious limitations. In particular, the improbability of the null hypothesis makes its rejection a rather weak benchmark of scientific progress. Also, the practice of summing behavioral measures across subjects and calculating group averages may produce artificial descriptions that misrepresent the degree of variability characterizing behavior. Finally, the practices of statistical inference have taken on a life of their own in the behavioral sciences, and some scientists believe that unjustifiable amounts of time and effort are spent training researchers on these questionable decision-making criteria, when instead more time should be devoted to learning about the subject matter relevant to one's profession.

PHILOSOPHY OF THE SINGLE-CASE ALTERNATIVE

Among the more fruitful outcomes of the recent dialogue concerning NHST research is an enhanced willingness by researchers to consider alternative research strategies. One such alternative, and the major subject of this book, is the *single-case research design*, also referred to variably as *single-subject*, *small n*, and $N = 1$ designs. We use the moniker *single-case design* to describe this class of research strategies because, depending on the context, a *case* may refer to a particular client, a social unit, such as a family or a support group, or even to an institutional or organizational body, such as a group of employees. In most health care environments, the "case" would in fact be a client receiving some kind of medical or health-related service.

Although the single-case method represents one alternative to the NHST strategy, it is anything but a newcomer. In fact, single-case research was in use in the behavioral sciences long before the development of group designs and statistical inference, and it continues to make important contributions to the study of behavior, particularly in applied settings (Blampied, 1999, 2000, 2001; Blampied, Barabasz, & Barabasz, 1996; Morgan & Morgan, 2001). There is much more, however, to this research strategy than simply choosing to measure variables at the level of the individual instead of the group. In fact, single-case research design differs from traditional group designs along several dimensions, and these differences are reflected in all aspects of the research design, from observation and measurement, to manipulation of independent variables, to data analysis and interpretation, to the drawing of conclusions. Although we will have a chance later to examine these differences in detail, let's take a brief look at the more fundamental assumptions that inform single-case research.

Behavior as an Individual Organism Phenomenon

The pursuit of any scientific enterprise begins with some basic assumptions about the phenomenon of interest. How scientists view their subject matter has important consequences for methodological practices, including measurement, research design, and the drawing of inferences or conclusions. Among the most fundamental assumptions guiding single-case research is the recognition that behavior is a natural phenomenon that takes place at the level of the individual organism. Although we can speak of abstractions such as *groups*, *communities*, *societies*, and the like, much of what interests us about behavior is readily observed at the individual level. Individual organisms, after all, must solve certain problems in order to survive, including the acquisition of food and water, safety from predators or the elements, and selection of a mate for purposes of reproduction. These are universal requirements of all biological creatures, and while you and I as humans may achieve them through very different means than do other animals, we all do so only through behaving or acting on the world around us.

The study of human behavior is necessarily an enormous undertaking because it includes every activity imaginable, from simplistic actions, such as reflexively blinking when encountering a bright light, to complex activity, such as programming a computer. The entirety of human experience, including all of our thoughts, emotions, and actions, is open to investigation. However, it is essential to understand that our unit of analysis, *at least within the behavioral sciences*, is the individual. Even when we study group behavior, such as decision making in a jury, most of our observations will be made initially at the level of the individual—and it is at precisely this level that most human service personnel intervene in human

behavior. The clinical psychologist treating a client's snake phobia, the home care nurse teaching a patient to operate his or her own intravenous equipment, and the physical therapist assisting an accident victim in relearning physical movements are all contributing to the adaptive behavior of their individual clients.

Keep in mind, too, that behavior is a fairly inclusive concept. Although for methodological reasons many behavioral scientists have restricted their observations to overt, readily observable behavior, we well recognize the importance of the less conspicuous processes of thinking, imagining, and feeling. Over the years, researchers have developed some ingenious ways of enhancing the accessibility to scientific scrutiny of these covert processes, referred to by Skinner (1953) as *private events*. Through self-report instruments, explicit training in self observation and self-monitoring, and even through sophisticated scanning technology, previously hidden dimensions of behavior have increasingly been opened up to analysis. An important point about all of these developments, however, is that they reflect advances in our ability to observe, measure, and record activity at the level of the individual organism.

Clearly, the most apparent feature of single-case research is that observation and measurement always take place at the level of the individual subject. Of course, this is also true of group designs. Except under highly unusual circumstances—perhaps if one is measuring, in decibels, the total noise level of a crowd—most observation and measurement in the behavioral and health sciences are initially conducted at the level of the individual subject. It is common, however, in group designs, to combine the individual data from all participants in the research, usually by calculating means or other quantitative summary measures. These aggregate measures are then treated statistically during the data analysis phase of the study. Finally, in such research it is understood that the conclusions are to be applied only at the group level, not to individuals. For example, in the experiment on eating disorder risk, the results demonstrated that the group of participants who received the Internet treatment improved on several dependent measures relative to members of the control group. This improvement, however, manifested itself only in differences between the group means on relevant measures, such as self-esteem and social support. These group differences do not allow the researchers to make any specific statements about individuals within the groups. In other words, just because the treatment group showed increased self-esteem after treatment does not mean that every person in that group benefited from treatment on this measure or that members of the control group did not evidence increases in self-esteem.

Unlike group designs, single-case research designs do not involve aggregation of data across multiple participants for the purpose of creating group statistics. Data collection, analysis, and presentation are all conducted on individual data

only, and summary measures are not calculated across subjects. In this way, single-case designs avoid the problem of referring to the "average" subject, because this label is understood to reflect only the abstract consequence of a somewhat arbitrary mathematical calculation.

It is important to keep in mind that the phrase *single-case research* does not mean that only *one* subject participates in the study. The designation *single case* simply means that all data collection and analysis are conducted on the data from individual subjects, not at the group level. Actually, several subjects may serve in a single-case research study and, in fact, most single-case research studies include more than one subject. The participation of more than one subject is crucial if researchers are to reach conclusions that can be broadly applied. Only when the effect of an independent variable can be shown over and over again, through experimental replication, will such an effect take on the status of a general principle or law. Science relies quite heavily on the replication of important findings. We will have more to say later on the special place of replication in single-case research.

Measurement Must Be Sensitive to Behavioral Continuity

In addition to occurring at the individual level, behavior also manifests itself continuously. Learning to ride a bicycle, writing a term paper, discussing discipline with a teenager, studying for an exam, and practicing therapeutic exercises are all quite difficult to conceive of as discrete events, and for very good reason: They are not discrete, one-time occurrences but dynamic events that unfold and change, sometimes dramatically, over time. It makes sense, then, to expect observational and measurement strategies within the behavioral sciences to take the temporal and serial dependence of behavior into account. Unfortunately, this has not been the case historically. A discrete group mean, measured at only one point in time, regardless of how many participants' data contributed to the calculation, does not adequately reflect behavior's continuous nature. Instead, it is a little like looking at a single snapshot of a gymnast, frozen in time in the middle of a routine. The picture does little to portray the complexity and drama of the rapid changes of pace and sequential transitions from one difficult move to the next that characterize the full routine. For this reason, single-case researchers believe the discrete measurement practices of group designs are ill founded within the behavioral and health sciences, not because such measurement is inherently poor but because it does not sufficiently map onto a continuous, dynamic phenomenon such as behavior.

The logic of single-case research mandates that researchers observe and measure behavior as continuously as possible within the practical constraints of any particular study. Continuous measurement is advantageous for several reasons, perhaps the most important being the fact that many variables, including those that

might be explicitly manipulated by the researcher, affect behavior differently with continued exposure. Many therapeutic regimens, in particular, produce their effects unevenly over the course of treatment. As a result, the nature and intensity of the treatment itself may have to be adjusted in response to the client's level of improvement. Such moment-to-moment adjustments are not possible in designs that employ single outcome measures.

Continuous measurement also serves the purpose of ensuring the representativeness of our data. Many natural phenomena, including behavior, exhibit cycles or fluctuations over time (Beasley, Allison, & Gorman, 1997). For example, suppose a corporate executive named Bob, on the advice of his physician, decides to begin an exercise regimen. He decides to walk a trail near his workplace during each day's lunch break and to walk in his neighborhood on weekends. Unfortunately, standard meetings that occur every Tuesday and Thursday prevent walking on those days. Therefore, a graph of the number of miles Bob walked each day over a 2-week period might look something like Figure 2.2. Notice that the distance walked ranges from a low of 0 miles to a high of 4 miles. This means that if we were to take any one day's measure as an accurate representation of the typical amount of walking Bob has done, we would have to conclude that he either walks a good deal (2–4 miles) or not at all. Neither of these single measures seems to do a very good job of capturing or representing the behavior of interest. On the other hand, measuring and recording the behavior every day allows us to follow the natural pattern this behavior exhibits and to appreciate the typical fluctuations that characterize any behavior when viewed through a temporal window. This measurement scheme helps us to see the process by which behavior change comes about, not just the end product. Also, this fairly refined observational strategy often proves useful in identifying naturally occurring moderating variables, some of which may serve as independent variables in a treatment regimen.

Participants Serve as Their Own Controls

Remember that in group designs the effects of a manipulated independent variable are assessed by comparing group means on the dependent variable of interest. This between-groups comparison is evaluated statistically to determine whether such a difference could have occurred by chance, or if something other than chance was operating. If all procedural aspects of the study, including the elimination of extraneous variables, were properly conducted, then that something else, within a certain degree of probability, was the independent variable.

In single-case research, group means are not even calculated, so between-group mean comparisons are not possible. What are available, however, are several

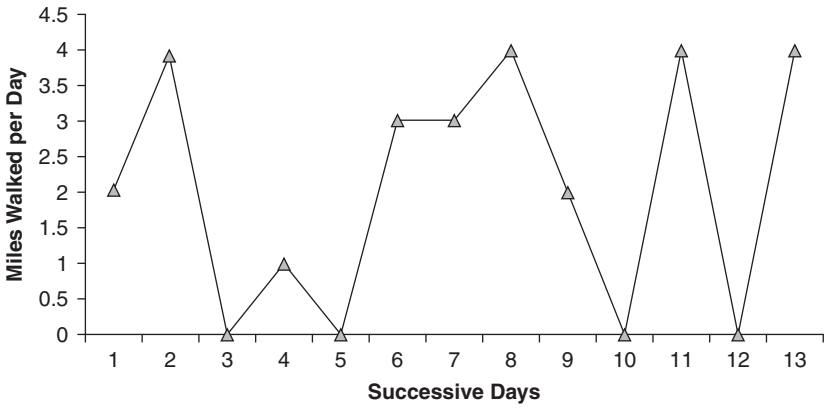

Figure 2.2 Bob's walking distance

measures of the dependent variable for each subject, obtained both prior to and during or after intervention. Thus, the meaningful comparison in this type of design is between the same person's behavior at different times. To some extent, single-case designs resemble, in logic, the before-and-after portraits of patients presented in testimonials or advertisements for weight loss, hair replacement, and other clinical treatment programs. Because different patients begin such a program at different levels of the dependent variable (e.g., different weights), simply calculating a group mean can be somewhat misleading. A weight loss of 18 pounds may be rather inconsequential to someone who begins a program at 350 pounds but considered quite respectable for someone who begins the program at 169 pounds. Collecting the data for each individual and having each person serve as the benchmark against which to measure his or her own amount of change offer a more refined assessment than that provided by group averages. There will, of course, be differences between subjects on any measure of change, but such differences are at least easily observed in single-case designs. They are obscured by group averages in group designs, and only a purposeful analysis of individual data will reveal these differences in such a study.

Drawing Inferences in Single-Case Research

In group designs we wish to draw conclusions about the larger population from which our sample was taken. We do so using statistical tests that take advantage of well-known mathematical properties of normal curves. It is important to understand,

however, that the inferences we draw hold true only at the group level; that is, if the study on eating disorder risk demonstrates improvement for participants who received the Internet treatment, then we can infer that such an outcome would hold true at the population level. In other words, we could conclude that the mean differences in improvement between the treatment and control groups in Zabinski et al.'s (2004) study are reflective of a similar mean difference in the population. This difference attends only to the group mean or average, though, not to individuals. As a consequence, we cannot really recommend the Internet treatment to any person at risk of developing an eating disorder, because the data do not allow inferences about individual behavior. We are on safe ground only so long as we discuss average group differences.

Because in single-case studies individuals are compared only with themselves under different experimental conditions, the most meaningful inference pertains to that subject. Instead of asking whether the difference between two calculated group means is larger than would be expected by chance, we want to know whether an individual's behavior following an experimental treatment is noticeably different from its pretreatment level. Although statistical techniques can be of assistance in making such decisions (Jones, Vaught, & Reid, 1975; Kazdin, 1976), single-case researchers seldom call on statistical inference to interpret data. Instead, the plotting of data on real-time charts, and pre- and posttreatment comparisons of variability and upward and downward trends in data, serve as the primary vehicle for data interpretation in this research tradition (Parsonson & Baer, 1986; Sidman, 1960). This graphic display of data is more reminiscent of the data analysis methods of the natural sciences than of the behavioral sciences. Indeed, B. F. Skinner (1956), among the most influential proponents of single-case research, often contended that the study of behavior was in fact a specialized branch of biological science.

Of course, at some level, single-case researchers want to draw the same conclusions desired by group design researchers. We are always interested in whether a particular finding will extend beyond our study sample, whether that sample included a handful or 1,000 subjects. The question of whether an empirical finding exhibits **generality**, or applies to subjects or settings beyond a particular research project, is a difficult one to answer. In general, the size of a study's sample will have little bearing on establishing the generality of a particular finding. Only through *replication*, the systematic repetition of a study, can the generality of a phenomenon be seriously confirmed or disconfirmed (Cohen, 1990; Sidman, 1960). In other words, both group and single-subject researchers must rely on the logic of replication to establish the widespread reliability of a behavioral phenomenon.

INTERIM SUMMARY

Single-subject research is informed by certain basic assumptions about the subject matter of the behavioral sciences, including the claim that behavior is best understood as a phenomenon that occurs at the level of individual organisms. Consequently, the single-subject method places a premium on observation and measurement of behavior in individual subjects. An additional assumption is that behavior is a continuous phenomenon, changing dynamically with time, thus justifying a continuous measurement strategy. Measuring behavior at the individual level also allows the subject to serve as his or her own comparison because behavioral measures can be taken prior to, during, and after manipulation of a variable or treatment. Finally, conclusions drawn from single-subject research seldom rely on statistical criteria, but instead on real-time data displayed graphically for each subject.

A BRIEF HISTORY OF SINGLE-CASE RESEARCH

We have already seen that single-case research actually predated the large-group research designs that would eventually become the norm in the behavioral sciences. This is particularly true of psychology, which was, prior to the turn of the century, closely aligned with experimental physiology. In its infancy, psychological science was a staunchly laboratory-oriented discipline because its practitioners were adamant about distinguishing the young science from the "armchair thought" experiments of philosophy. Only by adopting the trappings of "real science," including laboratory preparations, instrumentation, and objective empiricism, could psychology be taken seriously as a scientific discipline. In addition, many of psychology's early pioneers were trained in the scientific methods of biology and physiology.

Early Experimental Psychology

Among the earliest formal psychological experiments were studies of the various human senses and their capabilities. Two German scientists, Ernst Weber (1795–1878) and Gustav Fechner (1801–1887), established many of psychology's first quantitative principles by mapping out the sensory thresholds for vision, audition, and touch. Because such studies require the presentation of various amounts of stimulation to the subject over lengthy experimental sessions, large-group studies are often impractical. Indeed, the study of sensory thresholds, known as **psychophysics**, is still dominated by single-case designs. Among the fascinating topics being addressed by contemporary psychophysicists are the sensory capacities of

many nonhuman animals, as well as the practical benefits accruing from such knowledge. A case in point is the government and/or police trained canine whose tremendous olfactory abilities allows it to detect minuscule amounts of illicit drugs like cocaine, heroin, and marijuana; materials used in explosives; and even the ink on illegal bank notes. Using single-case methodology and a sophisticated combination of psychophysical techniques and operant conditioning, scientists have trained dogs to identify illegal substances and, perhaps just as important, disregard irrelevant stimuli, in such busy environments as schools, airports, and office buildings. These animals are so well trained and so skilled that they often thwart ingenious criminals, such as the smugglers who placed heroin inside concrete statues (Wren, 1999). In fact, these animals have proven to be both more sensitive and more reliable in detecting substances than are computerized electronic detectors (Wren, 1999).

The first laboratory-based study of human memory can also be traced to Germany and the work of Hermann Ebbinghaus (1850–1909). Ebbinghaus (1885/1913) created the nonsense syllable, three-letter nonword combinations (*dal, fep, bil, hos,* etc.) that possessed no pre-experimental meaning. This was an important methodological innovation because Ebbinghaus was interested in how novel material is processed in memory, and conventional words would have been contaminated by previous associations and usage.

Ebbinghaus then rehearsed a series of such nonsense syllables until he could recite the list without error. Thus, he could quantify how frequently a list of given length had to be rehearsed in order to commit the list to memory. Ebbinghaus also studied how learning one list affected the learning of subsequent lists, leading to the discovery of what is referred to today as *interference effects*. He was also able to track the amount of forgetting that occurs as a function of time, length of list, and other variables. Ebbinghaus's research, carried out in the 1880s, represented the first systematic effort to identify the lawful properties of human memory, and he served by himself as the sole subject in this research. Indeed, for this reason his research program may represent the most pure instance of single-case research in psychology's history. Impressively, many of Ebbinghaus's basic findings hold up remarkably well more than a century later, serving as ample testimony not only to his scientific ingenuity and rigor but also to the strength of the single-case method.

Behavior Analysis and Cognitive Psychology

Although eclipsed in popularity by large-group research and statistical inference during the 1920s and 1930s, the single-case method retained a strong cadre of proponents and practitioners within psychology. Particularly ardent support for the method would come from the field of behavior analysis and its articulate and

influential spokesman, B. F. Skinner (1904–1990). As a psychology graduate student at Harvard in the late 1920s, Skinner wrestled with the problem of research design in psychology as he studied fundamental learning processes in nonhumans. A wonderful, often-humorous account of Skinner's (1956) development as a scientist includes a description of the various laboratory instruments this eminent scholar created to study learning in individual organisms, including the famous instrument which bears his name, the Skinner box. Skinner eventually became frustrated not only by the impracticality of monitoring the behavior of many different subjects during learning experiments but also by psychology's growing tendency to group data from many subjects together for the purpose of calculating average scores. Skinner was of the opinion that this aggregation of data actually obscured the nature of the learning process by producing smooth, gradual learning curves that failed to represent the complexity of learning. Arguing that psychology was a natural science, Skinner chose instead to observe and record the behavior of a single subject over the course of long experimental sessions, both prior to and after the manipulation of experimental variables.

Over the course of several decades, Skinner and his students accumulated large amounts of data representing the orderly effects of independent variables on various dimensions of behavior. Despite the orderliness of the data and the rigorous experimental conditions under which they were collected, Skinner and his colleagues had difficulty getting their work published in major psychological journals, largely because of the reluctance of editors to publish reports of studies based on data from single subjects. By the 1930s and 1940s, null hypothesis testing and statistical inference were standard aspects of most psychological research, and the idea of conducting an experiment with fewer than 30 or 40 subjects was viewed with considerable disdain. In response to this rather politicized climate of scientific publishing, Skinner and his colleagues established the *Journal of the Experimental Analysis of Behavior* (*JEAB*) in 1958. *JEAB* eventually became one of experimental psychology's most prestigious journals. To this day, *JEAB* remains the primary outlet for research on operant behavior (e.g., reinforcement, punishment, extinction, stimulus control) and publishes only articles in which the data from individual organisms are presented without aggregation.

Behavior analysts have not been alone in their endorsement of single-case research within psychology. In 1972, Newell and Simon published a landmark book in cognitive psychology, *Human Problem Solving*, which outlined both a theory and detailed methodological strategies for studying problem solving. An interesting feature of their method was the intense analysis of individual subjects attempting to solve an experimenter-imposed problem. Newell and Simon used a process known as **protocol analysis**, in which a participant verbalized his or her moment-to-moment decisions and strategies throughout a problem-solving

session. Thus, a written transcript of the participant's verbalizations served as the primary data in this research and, for obvious reasons, Newell and Simon refrained from combining the transcripts of multiple participants for purposes of analysis.

Protocol analysis represents an important example of how researchers must develop observational and measurement strategies that respect the natural dimensions of their subject matter. Newell and Simon were not merely interested in whether their participants solved the problem; instead, they wished to examine the thoughts, decision criteria, and detailed strategies the participants used throughout the experiment. In other words, they were interested in studying the process, not product, of problem solving. This required a data collection procedure that allowed for ongoing measurement, because problem solving is inherently a continuous process, in which responses and decisions made at one juncture affect later responses and options. Rather than producing a discrete and uninformative group mean, *protocol analysis* offers a refined and detailed account of the dynamic interplay between the subject and the problem space. Protocol analysis would eventually become a standard procedure for the study of human problem solving (Ericsson & Simon, 1984).

Case Studies and Clinical Psychology

A focus on the individual case has played an integral role in the health sciences, particularly psychiatry and clinical psychology. Serving an especially instructive role in the professional literatures of these disciplines is the **case study**, a detailed description of diagnosis, treatment, and outcome for a particular client. Information in a case study is often very thorough, including family background, history of illness, possible precipitating factors, and other information that helps to provide context for understanding the specific case. Also, case studies are often presented for cases that appear quite novel or unique in terms of symptoms, or that prove challenging or unresponsive to treatment. In general, case studies are usually published for their educational value and are often considered invaluable by individuals training in medicine and other health sciences.

It is important to note, however, that a case study is not, first and foremost, a research endeavor; that is, such written accounts ordinarily do not entail formal research design, manipulation of variables, and data analysis and interpretation consistent with that encountered in the scientific research literature. Nor is the purpose of a case study to discover some invariant principle of physical or psychological pathology or the causal relationship between a given treatment regimen and clinical outcome. Case studies are important to the extent that they help inform the practicing clinician of certain features of a disorder or its treatment that *may* be of relevance or help shed light on similar cases in the future. The primary

limitation of case studies, at least from a scientific standpoint, is not that they represent a single case but that strong conclusions regarding cause and effect are not possible because of a lack of rigor in research design. In this sense, case studies are to be distinguished from the single-case research designs that are the focus of this book. As Aeschleman (1991) pointed out, single-case designs are experimental research designs that allow for a significant amount of control over relevant variables and consequently support scientifically sound inferences. Also, most single-case studies do not use only one subject. The relevant independent variables are manipulated, and data collected and analyzed, at the level of individual subjects, but in most cases several subjects serve in such studies.

INTERIM SUMMARY

The single-case research strategy has played a historically important role in the behavioral and health sciences. Indeed, much of the work conducted by pioneering experimental psychologists, prior to the advent of Fisherian designs and statistical inference, used single-case designs. More recently, both behavioral and cognitive psychologists, despite divergent worldviews, have benefited from studies of individual participants. A focus on the single case has also characterized the work of clinicians in both psychology and medicine, particularly in the form of case studies. In Chapter 3, we continue our discussion of single-case designs by focusing on observation. Systematic observation, frequently overlooked, is one of the most fundamental and critical components of the scientific enterprise.

KEY TERMS GLOSSARY

Group designs Studies in which data are aggregated across many subjects and inferences are drawn on the basis of between-group comparisons.

Mean An arithmetic average calculated by adding scores together and then dividing this sum by the number of scores added.

Normal curve A probability distribution of scores on a variable in which scores are symmetrically distributed about a sample mean.

Hypothesis A testable statement about a relationship between two variables.

Null hypothesis The hypothesis that no relationship exists between the independent and the dependent variables in a study. In a particular study, for example, sample data, usually in the form of experimental and control group means, are being

evaluated against the null hypothesis that there is no difference between population group means.

Null hypothesis significance testing (NHST) A procedure whereby the null hypothesis is either rejected or accepted according to whether the value of a sample statistic yielded by an experiment falls within a certain predetermined "rejection region" of possible values.

Statistical inference The act of drawing conclusions about a population based on observed sample statistics.

Generality The extent to which the findings of a particular study can be extended to, or are representative of, the larger population from which subjects were sampled.

Psychophysics The study of sensory thresholds and stimulus discrimination in humans and nonhuman animals.

Protocol analysis A data collection strategy in which research participants describe, in written or verbal form, their strategies in real time during problem-solving tasks.

Case study A common nonexperimental practice in medicine in which detailed information is collected from a patient before and during the course of treatment.

SUPPLEMENTS

Review Questions

1. In group research, what is the purpose of comparing the experimental group with the control group? Also, how are group means calculated?
2. What does the phrase "Intelligence is normally distributed" mean? Why is the normal curve of use to researchers?
3. Suppose a researcher is studying a new anti-anxiety medication in a large group of people with social phobia. The researcher will be comparing a control group that receives a placebo with the experimental group that receives the new medication. What would the null hypothesis be in this study?
4. Why is the calculation of group means and the aggregating of data across subjects incompatible with the study of behavioral continuity? How do single-case researchers assess behavioral continuity?
5. What does it mean to say that subjects in single-case studies "serve as their own controls"? Describe how this concept would be exhibited in a single-case drug study similar to that described in Question 3.

6. A cognitive psychologist is using protocol analysis to study how children solve logic problems. Why would the psychologist not aggregate or sum together the data from several children for purposes of analysis?

7. How are case studies in medicine and psychiatry different from the single-case research designs described in this text?

SUGGESTED READINGS/HELPFUL WEB SITES

Leary, M. R. (2004). *Introduction to behavioral research methods* (4th ed.). Boston: Pearson.

This text is a very readable introduction to research methods in the behavioral sciences. The general logic of experimental design and null hypothesis testing is described.

www.tushar-mehta.com/excel/charts/normal_distribution/

This Web page offers descriptions and explanations of the normal distribution and its contribution to research methodology. The page uses Microsoft Excel exercises to demonstrate the normal curve and its relationship to such statistics as the mean and standard deviation.

Skinner, B. F. (1956). A case history in scientific method. *American Psychologist, 11,* 221–233.

In this engaging and at times humorous article, Skinner describes his development as a scientist, some of the apparatus which he created to study learning, and his rationale for developing single-case methods for studying behavior change.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

This groundbreaking work by two pioneers in cognitive psychology describes the use of protocol analysis to collect and analyze individual subject data. The book is important in arguing for a method of data analysis appropriate to behavior that changes over time (problem solving) and in recognizing the limitations of traditional group designs and null hypothesis testing.