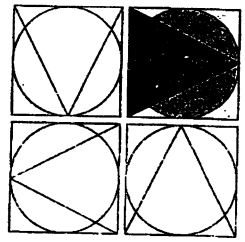


CHAPTER 18

RESEARCH DESIGN: PURPOSE AND PRINCIPLES

Research Design is the plan and structure of investigation, conceived so as to obtain answers to research questions. The plan is the overall scheme or program of the research. It includes an outline of what the investigator will do, from writing the hypotheses and their operational implications to the final analysis of data. The structure of research is harder to explain because the word *structure* is difficult to define clearly and unambiguously. Since it is a concept that becomes increasingly important as we continue our study, we here break off and attempt a definition and a brief explanation. The discourse will necessarily be somewhat abstract at this point. Later examples, however, will be more concrete. More important, we will find the concept powerful, useful, even indispensable, especially in our later study of multivariate analysis where "structure" is a key concept whose understanding is essential to understanding much contemporary research methodology.

A *structure* is the framework, organization, or configuration of elements of the structure related in specified ways. The best way to specify a structure is to write a mathematical equation that relates the parts of the structure to each other. Such a

mathematical equation, since its terms are defined and specifically related by the equation (or set of equations), is unambiguous. In short, a structure is a paradigm or model of the relations among the variables of a study. The words *structure*, *model*, and *paradigm* are troublesome because they are hard to define clearly and unambiguously. A "paradigm" is a model, an example. Diagrams, graphs, and verbal outlines are paradigms. We use "paradigm" here rather than "model" because "model" has another important meaning in science—a meaning we return to in Chapter 37 when we discuss the testing of theory using multivariate procedure and "models" of aspects of theories.

A research design expresses both the structure of the research problem and the plan of investigation used to obtain empirical evidence on the relations of the problem. We will soon encounter examples of both design and structure that will perhaps enliven this abstract discussion.

Purposes of Research Design

Research design has two basic purposes: (1) *to provide answers to research questions* and (2) *to control variance*. Design helps investigators obtain answers to the questions of research and also to control the experimental, extraneous, and error variances of the particular research problem under study. Since all research activity can be said to have the purpose of providing answers to research questions, it is possible to omit this purpose from the discussion and to say that research design has one grand purpose: to control variance. Such a delimitation of the purpose of design, however, is dangerous. Without strong stress on the research questions and on the use of design to help provide answers to these questions, the study of design can degenerate into an interesting, but sterile, technical exercise.

Research designs are invented to enable researchers to answer research questions as validly, objectively, accurately, and economically as possible. Research plans are deliberately and specifically conceived and executed to bring empirical evidence to bear on the research problem. Research problems can be, and are, stated in the form of hypotheses. At some point in the research they are stated so that they can be empirically tested. Designs are carefully worked out to yield dependable and valid answers to the research questions epitomized by the hypotheses. We can make one observation and infer that the hypothesized relation exists on the basis of this one observation, but it is obvious that we cannot accept the inference so made. On the other hand, it is also possible to make hundreds of observations and to infer that the hypothesized relation exists on the basis of these many observations. In this case we may or may not accept the inference as valid. The result depends on how the observations and the inference were made. An adequately planned and executed design helps greatly in permitting us to rely on both our observations and our inferences.

How does design accomplish this? Research design sets up the framework for study of the relations among variables. Design tells us, in a sense, what observations to make, how to make them, and how to analyze the quantitative representations of

the observations. Strictly speaking, design does not "tell" us precisely what to do, but rather "suggests" the direction of observation-making and analysis. An adequate design "suggests," for example, how many observations should be made, and which variables are active and which are attribute variables. We can then act to manipulate the active variables and to categorize and measure the attribute variables. A design tells us which type of statistical analysis to use. Finally, an adequate design outlines possible conclusions to be drawn from the statistical analysis.

An Example

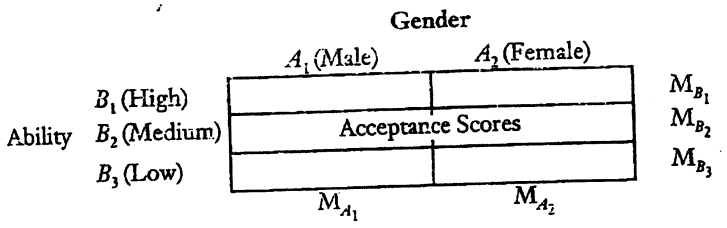
It has been said that colleges and universities discriminate against women in hiring and in admissions. Suppose we wanted to test discrimination in admissions. The idea for this example came from the unusual and ingenious experiment cited earlier: Walster, Cleary, and Clifford (1970). We set up an experiment as follows. To a random sample of 200 colleges we send applications for admission, basing the applications on several model cases selected over a range of tested ability, with all details the same except for gender. Half the applications will be those from men and half from women. Other things being equal, we expect approximately equal numbers of acceptances and rejections. Acceptance, then, is the dependent variable. It is measured on a three-point scale: full acceptance, qualified acceptance, and rejection. Call male A_1 and female A_2. The paradigm of the design is given in Figure 18.1.

The design is the simplest possible, given minimum requirements of control. The two treatments will be assigned to the colleges at random. Each college, then, will receive one application, which will be either male or female. The difference between the means, M_{A_1} and M_{A_2} will be tested for statistical significance with a t- or F-test. The substantive hypothesis is: $M_{A_1} > M_{A_2}$, or more males than females will be accepted for admission. If there is no discrimination in admissions, then M_{A_1} is statistically equal to M_{A_2}. Suppose that an F-test indicates that the means are not significantly different. Can we then be sure that there is no discrimination practiced (on the average)? While the design of Figure 18.1 is satisfactory as far as it goes, perhaps it does not go far enough.

FIGURE 18.1

Treatments	
A_1 (Male)	A_2 (Female)
Acceptance Scores	
M_{A_1}	M_{A_2}

FIGURE 18.2



A Stronger Design

Walster and her colleagues used two other independent variables, Race and Ability, in a factorial design. We drop Race—it was neither statistically significant, nor did it interact significantly with the other variables—and concentrate on gender and ability. If a college bases its selection of incoming students strictly on ability, there is no discrimination (unless, of course, ability selection is called discrimination). Add Ability to the design of Figure 18.1; use three levels. That is, in addition to the applications being designated Male and Female, they are also designated as High Ability, Medium Ability, and Low Ability. For example, three of the applicants may be: male, medium ability; female, high ability; female, low ability. Now, if there is no significant difference between genders and the interaction between Gender and Ability is not significant, this would be considerably stronger evidence for no discrimination than that yielded by the design and statistical test of Figure 18.1. We now use the expanded design to explain this statement and to discuss a number of points about research design. The expanded design is given in Figure 18.2.

The design is a 2 × 3 factorial. One independent variable, A, is gender, the same as in Figure 18.1. The second independent variable, B, is ability, which is manipulated by indicating in several ways what the ability levels of the students are. It is important not to be confused by the names of the variables. Gender and Ability are ordinarily attribute variables and thus nonexperimental. In this case, however, they are manipulated. The students' records sent to the colleges were systematically adjusted to fit the six cells of Figure 18.2. A case in the A_1B_2 cell, for instance, would be the record of a male of medium ability. It is this record that the college judges for admission.

Let's assume that we believe discrimination against women takes a more subtle form than simply across-the-board exclusion: that it is the women of lower ability who are discriminated against (compared to men). This is an interaction hypothesis. At any rate, we use this problem and the paradigm of Figure 18.2 as a basis for discussing some elements of research design.

Research problems suggest research designs. Since the hypothesis just discussed is one of interaction, a factorial design is evidently appropriate. A is Gender; B is Ability. A is partitioned into A_1 and A_2, and B into B_1, B_2, and B_3.

The paradigm of Figure 18.2 suggests a number of things. First and most obvious, a fairly large number of participants is needed. Specifically, $6n$ participants are necessary (n equals number of *Ss* in each cell). If we decide that n should be 20, then we must have 120 *Ss* for the experiment. Note the "wisdom" of the design here. If we were only testing the treatments and ignoring ability, only $2n$ *Ss* would be needed. Please note that some, such as Simon (1976, 1987); Simon and Roscoe (1984); and Daniel (1976) disagree with this approach for all types of problems. They feel that many designs contains hidden replications and that one can do with a lot fewer participants than 20 per cell. Such designs do require a lot more careful planning, but the researcher can come out with a lot more useful information and study more independent variables than just two or three.

There are ways to determine how many participants are needed in a study. Such determination is part of the subject of "power," which refers to the ability of a test of statistical significance to detect differences in means (or other statistics) when such differences indeed exist. Chapter 8 discusses sample sizes and their relationship to research. Chapter 12, however, presents a method for estimating sample sizes to meet certain criteria. Power is a fractional value between 0 and 1.00 that is defined as $1 - \beta$, where β is the probability of committing a Type II error. The Type II error is failing to reject a false null hypothesis. If power is high (close to 1.00), this says that if the statistical test was not significant, the research can conclude that the null hypothesis is true. Power also tells you how sensitive the statistical test is in picking up real differences. If the statistical test is not sensitive enough to detect a real difference, the test is said to have low power. A highly sensitive test that can pick up true differences is said to have high power. In Chapter 16, we discussed the difference between parametric and nonparametric statistical tests. Nonparametric tests are generally less sensitive than parametric tests. As a result, nonparametric tests are said to have lower power than parametric tests. One of the most comprehensive books on the topic of power estimation is by Cohen (1988). Jaccard and Becker (1997) give an easy-to-follow introduction to power analysis.

Second, the design indicates that the "participants" (colleges, in this case) can be assigned randomly to both A and B because both are experimental variables. If Ability was a nonexperimental attribute variable, however, then the participants could be randomly assigned to A_1 and A_2, but not to B_1, B_2, and B_3.

Third, according to the design the observations made on the "participants" must be made independently. The score of one college must not affect the score of another college. Reducing a design to an outline like that shown in Figure 18.2 in effect prescribes the operations necessary for obtaining the measures that are appropriate for the statistical analysis. An F-test depends on the assumption of the independence of the measures of the dependent variable. If Ability here is an attribute variable and individuals are measured for intelligence, say, then the independence requirement is in greater jeopardy because of the possibility of one subject seeing another subject's paper, and because teachers may unknowingly (or knowingly) "help" students with answers, among other reasons. Researchers try to prevent such things—not on moral grounds but to satisfy the requirements of sound design and sound statistics.

A fourth point is quite obvious to us by now: Figure 18.2 suggests factorial analysis of variance, F-tests, measures of association and, perhaps, post hoc tests. If

the research is well designed before the data are gathered—as it certainly was by Walster et al.—most statistical problems can be solved. In addition, certain troublesome problems can be avoided before they arise, or can even be prevented from arising at all. With an inadequate design, however, problems of appropriate statistical tests may be very troublesome. One reason for the strong emphasis in this book on treating design and statistical problems concomitantly is to point out ways to avoid these problems. If design and statistical analysis are planned simultaneously, the analytical work is usually straightforward and uncluttered.

A highly useful dividend of design is this: A clear design, like that in Figure 18.2, suggests the statistical tests that can be made. A simple one-variable randomized design with two partitions, for example, two treatments, A_1 and A_2, permit, only a statistical test of the difference between the two statistics yielded by the data. These statistics might be two means, two medians, two ranges, two variances, two percentages, and so forth. Only one statistical test is ordinarily possible. With the design of Figure 18.2, however, three statistical tests are possible: (1) between A_1 and A_2; (2) among B_1, B_2, and B_3, and (3) the interaction of A and B. In most investigations, all the statistical tests are not of equal importance. The important ones, naturally, are those directly related to the research problems and hypotheses.

In the present case the interaction hypothesis [or (3) above] is the important one, since the discrimination is supposed to depend on ability level. Colleges may practice discrimination at different levels of ability. As suggested above, females (A_2) may be accepted more than males (A_1) at the higher ability level (B_1), whereas they may be accepted less at the lower ability level (B_3).

It should be evident that research design is not static. A knowledge of design can help us to plan and do better research, and can also suggest the testing of hypotheses. Probably more important, we may be led to realize that the design of a study is not adequate to the demands we are making of it. What is meant by this somewhat peculiar statement?

Assume that we formulate the interaction hypothesis as outlined above without knowing anything about factorial design. We set up a design consisting, actually, of two experiments. In one of these experiments we test A_1 against A_2 under condition B_1. In the second experiment we test A_1 against A_2 under condition B_2. The paradigm would look like that shown in Figure 18.3. (To make matters simpler, we are only

FIGURE 18.3

B_1, Condition Treatments		B_2, Condition Treatments	
A_1	A_2	A_1	A_2
M_{A_1}	M_{A_2}	M_{A_1}	M_{A_2}

using two levels of B_1, B_2, and B_3, but changing B_3 to B_2. The design is thus reduced to 2×2.)

The important point to note is that *no adequate* test of the hypothesis is possible with this design. A_1 can be tested against A_2 under both B_1 and B_2 conditions, to be sure. But it is not possible to know, clearly and unambiguously, whether there is a significant interaction between A and B. Even if $M_{A_1} > M_{A_2} \mid B_2$ (M_{A_1} is greater than M_{A_2}, under condition B_2), as hypothesized, the design cannot provide a clear possibility of confirming the hypothesized interaction, since we cannot obtain information about the differences between A_1 and A_2 at the two levels of B, B_1 and B_2. Remember that an interaction hypothesis implies, in this case, that the difference between A_1 and A_2 is different at B_1 from what it is at B_2. In other words, information of both A and B *together in one experiment* is needed to test an interaction hypothesis. If the statistical results of separate experiments showed a significant difference between A_1 and A_2 in one experiment under the B_1 condition, and no significant difference in another experiment under the B_2 condition, then there is good *presumptive* evidence that the interaction hypothesis is correct. But presumptive evidence is not good enough, especially when we know that it is possible to obtain better evidence.

In Figure 18.3, suppose the means of the cells were, from left to right: 30, 30, 40, 30. This result would seem to support the interaction hypothesis, since there is a significant difference between A_1 and A_2 at level B_2, but not at level B_1. But we could not know this to be certainly so, even though the difference between A_1 and A_2 is statistically significant. Figure 18.4 shows how this would look if a factorial design had been used. (The figures in the cells and on the margins are means.) Assuming that the main effects, A_1 and A_2; B_1 and B_2, were significant, it is still possible that the interaction is not significant. Unless the interaction hypothesis is specifically tested, the evidence for interaction is merely presumptive, because the planned statistical interaction test, that a factorial design provides, is lacking. It should be clear that a knowledge of design could have improved this experiment.

Research Design as Variance Control

The main technical function of research design is *to control variance*. A research design is, in a manner of speaking, a set of instructions to the investigator to gather and analyze data in certain ways. It is therefore a control mechanism. The statistical

FIGURE 18.4

	A_1	A_2	
B_1	30	30	30
B_2	40	30	35
	35	30	

principle behind this mechanism, as stated earlier, is: *Maximize systematic variance, control extraneous systematic variance, and minimize error variance.* In other words, we must *control* variance.

According to this principle, by constructing an efficient research design the investigator attempts to: (1) maximize the variance of the variable or variables of the substantive research hypothesis, (2) control the variance of extraneous or "unwanted" variables that may have an effect on the experimental outcomes, and (3) minimize the error or random variance, including so-called errors of measurement. Let's look at an example.

A Controversial Example

Controversy is rich in all science. It seems to be especially rich and varied in behavioral science. Two such controversies have arisen from different theories of human behavior and learning. Reinforcement theorists have amply demonstrated that positive reinforcement can enhance learning. As usual, however, things are not so simple. The presumed beneficial effect of external rewards has been questioned; research has shown that extrinsic reward can have a deleterious influence on children's motivation, intrinsic interest, and learning. A number of articles and studies were published in the 1970s showing the possible detrimental effects of using reward. In one such study Amabile (1979) showed that external evaluation has a deleterious effect on artistic creativity. Others included Deci (1971), and Lepper and Greene (1978). At the time, even the seemingly straightforward principle of reinforcement is not so straightforward. However, in recent years a number of articles have appeared defending the positive effects of reward (see Eisenberger & Cameron, 1996; Sharpley, 1988; McCullers, Fabes, & Moran, 1987; Bates, 1979).

There is a substantial body of belief and research that indicates that college students learn well under a regime of what has been called *mastery learning*. Very briefly "mastery learning" means a system of pedagogy based on personalized instruction and requiring students to learn curriculum units to a mastery criterion (see Abbott & Falstrom, 1975; Ross & McBean, 1995; Senemoglu & Fogelman, 1995; Bergin, 1995). Although there appears to be some research supporting the efficacy of mastery learning, there is at least one study—and a fine study it is—by Thompson (1980) whose results indicate that students taught through the mastery learning approach do no better than students taught with a conventional approach of lecture, discussion, and recitation. This is an exemplary study, done with careful controls, over an extended time period. The example given below was inspired by the Thompson study. The design and controls in the example, however, are much simpler than Thompson's. Note, too, that Thompson had an enormous advantage: He did his experiment in a military establishment. This means, of course, that many control problems, usually recalcitrant in educational research, were easily resolved.

Controversy enters the picture because mastery learning adherents seem so strongly convinced of its virtues, while its doubters are almost equally skeptical. Will research decide the matter? Hardly. But let's see how one might approach a relatively modest study capable of yielding at least a partial *empirical* answer.

An educational investigator decides to test the hypothesis that achievement in science is enhanced more by a mastery learning method (*ML*) than by a traditional method (*T*). We ignore the details of the methods and concentrate on the design of the research. Call the mastery learning method A_1 and the traditional method A_2. As investigators we know that other possible independent variables influence achievement: intelligence, gender, social class background, previous experience with science, motivation, and so on. We would have reason to believe that the two methods work differently with different kinds of students. They may work differently, for example, with students of differing scholastic aptitudes. The traditional approach is effective, perhaps, with students of high aptitude, whereas mastery learning is more effective with students of low aptitude. Call aptitude *B:* high aptitude is B_1 and low aptitude B_2. In this example, the variable Aptitude was dichotomous into high and low groups. This is not the best way to handle the Aptitude variable. When a continuous measure is dichotomized or trichotomized, variance is lost. In a later chapter we will see that leaving a continuous measure and using multiple regression is a better method.

What kind of design should be set up? To answer this question it is important to label the variables and to know clearly what questions are being asked. The variables are:

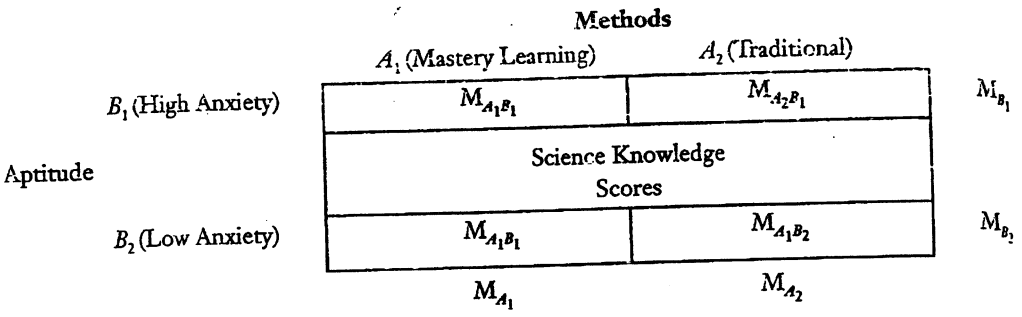
Independent Variables		Dependent Variable
Methods	*Aptitude*	*Science Achievement*
Mastery Learning, A_1	High Aptitude, B_1	Test scores in science
Traditional, A_2	Low Aptitude, B_2	

We may as investigators also have included other variables in the design, especially variables potentially influential on achievement: general intelligence, social class, gender, high school average, for example. We also would use random assignment to take care of intelligence and other possible influential independent variables. The dependent variable measure is provided by a standardized science knowledge test.

The problem seems to call for a factorial design. There are two reasons for this choice: (1) There are two independent variables. (2) We have quite clearly an interaction hypothesis in mind, though we may not have stated it in so many words. We do have the belief that the methods will work differently with different kinds of students. We set up the design structure shown in Figure 18.5.

Note that all the marginal and cell means have been appropriately labeled. Note, too, that there is one *active variable*, Methods; and one *attribute variable*, Aptitude. You might remember from Chapter 3 that an *active variable* is an experimental or manipulated variable. An *attribute variable* is a measured variable or a variable that is a characteristic of people or groups; for example, intelligence, social class, and occupation (people); and cohesiveness, productivity, and restrictive–permissive atmosphere (organizations, groups, and the like). All we can do is to categorize the

FIGURE 18.5



participants as high aptitude and low aptitude and assign them accordingly to B_1 and B_2. We can, however, assign the students randomly to A_1 and A_2, the Methods groups. This is done in two stages: (1) the B_1 (high aptitude) students are randomly assigned to A_1 and A_2 and (2) the B_2 (low aptitude) students are assigned randomly to A_1 and A_2. By so randomizing the participants we can assume that before the experiment begins, the students in A_1 are approximately equal to the students in A_2 in all possible characteristics.

Our present concern is with the different roles of variance in research design and the variance principle. Before going further, we name the variance principle for easy reference the "maxmincon" principle. The origin of this name is obvious: maximize the systematic variance under study; control extraneous systematic variance; and minimize error variance—with two of the syllables reversed for euphony.

Before tackling the application of the maxmincon principle in the present example, an important point should be discussed. Whenever we talk about variance, we must be sure to know which variance we are talking about. We speak of the variance of the methods, of intelligence, of gender, of type of home, and so on. This sounds as though we were talking about the independent variable variance. This is true and not true. We always mean the *variance of the dependent variable, and the variance of the dependent variable measures*, after the experiment has been done. This is not true in so-called correlational studies where, when we say "the variance of the independent variable," we mean just that. When correlating two variables, we study the variances of the independent and dependent variables "directly." Our way of saying "independent variable variance" stems from the fact that, by manipulation and control of independent variables, we influence, presumably, the variance of the dependent variable. Somewhat inaccurately put, we "make" the measures of the dependent variable behave or vary as a presumed result of our manipulation and control of the independent variables. In an experiment, it is the dependent variable measures that are analyzed. Then, from the analysis we infer that the variances present in the total

variance of the dependent variable measures are due to the manipulation and control of the independent variables, and to error. Now, back to our principle.

Maximization of Experimental Variance

The experimenter's most obvious, but not necessarily most important, concern is to maximize what we will call the ***experimental variance***. This term is introduced to facilitate subsequent discussions and, in general, simply refers to the variance of the dependent variable, influenced by the independent variable or variables of the substantive hypothesis. In this particular case, the experimental variance is the variance in the dependent variable, presumably due to methods, A_1 and A_2, and aptitude, B_1 and B_2. Although experimental variance can be taken to mean only the variance due to a manipulated or *active* variable, like methods, we shall also consider *attribute* variables, like intelligence, gender and, in this case, aptitude, experimental variables. One of the main tasks of an experimenter is to maximize this variance. The methods must be "pulled" apart as much as possible to make A_1 and A_2 (and A_3, A_4, and so on, if they are in the design) as unlike as possible.

If the independent variable does not vary substantially, there is little chance of separating its effect from the total variance of the dependent variable. It is necessary to give the variance of a relation a chance to show itself, to separate itself, so to speak, from the total variance, which is a composite of variances due to numerous sources and chance. Remembering this subprinciple of the maximincon principle, we can write a research precept: *Design, plan, and conduct research so that the experimental conditions are as different as possible.* There are, of course, exceptions to this subprinciple, but they are probably rare. An investigator might want to study the effects of small gradations of, say, motivational incentives on the learning of some subject matter. Here one would not make the experimental conditions as different as possible. Still, they would have to be made to vary somewhat or there would be no discernible resulting variance in the dependent variable.

In the present research example, this subprinciple means that the investigator must take pains to make A_1 and A_2, the mastery learning and traditional methods, as different as possible. Next, B_1 and B_2 must also be made as different as possible on the aptitude dimension. This latter problem is essentially one of measurement, as we will see in a later chapter. In an experiment, the investigator is like a puppeteer making the independent variable puppets do what he or she wants. The strings of the A_1 and A_2 puppets are held in the right hand and the strings of the B_1 and B_2 puppets in the left hand. (We assume there is no influence of one hand on the other, that is, the hands must be independent.) The A_1 and A_2 puppets are made to dance apart just as the B_1 and B_2 puppets are made to dance apart. The investigator then watches the audience (the dependent variable) to see and measure the effect of the manipulations. If one is successful in making A_1 and A_2 dance apart, and if there is a relation between A and the dependent variable, the audience reaction—if separating A_1 and A_2 is funny, for instance—should be laughter. The investigator may even observe that he or she only gets laughter when A_1 and A_2 dance apart and, at the same time, B_1 or B_2 dance apart (interaction again).

Control of Extraneous Variables

The control of extraneous variables means that the influences of those independent variables extraneous to the purposes of the study are minimized, nullified, or isolated. There are three ways to control extraneous variables. The first is the easiest, if it is possible: to eliminate the variable as a variable. If we are worried about intelligence as a possible contributing factor in studies of achievement, its effect on the dependent variable can be virtually eliminated by using participants of only one intelligence level, say intelligence scores within the range of 90 to 110. If we are studying achievement, and racial membership is a possible contributing factor to the variance of achievement, it can be eliminated by using only members of one race. The principle is: *To eliminate the effect of a possible influential independent variable on a dependent variable, choose participants so that they are as homogeneous as possible on that independent variable.*

This method of controlling unwanted or extraneous variance is very effective. If we select only one gender for an experiment, then we can be sure that gender cannot be a contributing independent variable. But then we lose generalization power; for instance, we can say nothing about the relation under study with girls if we use only boys in the experiment. If the range of intelligence is restricted, then we can discuss only this restricted range. Is it possible that the relation, if discovered, is nonexistent or quite different with children of high intelligence or children of low intelligence? We simply do not know; we can only surmise or guess.

The second way to control extraneous variance is through randomization. This is the best way, in the sense that you can have your cake and eat some of it, too. Theoretically, randomization is the only method for controlling all possible extraneous variables. Another way to phrase it is: if proper randomization has been accomplished, then the experimental groups can be considered statistically equal in all possible ways. This does not mean, of course, that the groups are equal in all the possible variables. We already know that by chance the groups can be unequal, but the probability of their being equal is greater, with proper randomization, than the probability of their not being equal. For this reason, control of the extraneous variance by randomization is a powerful method of control. All other methods leave open many possibilities of inequality. If we match for intelligence, we may successfully achieve statistical equality in intelligence (at least in those aspects of intelligence measured), but we may suffer from inequality in other significantly influential independent variables like aptitude, motivation, and social class. A precept that springs from this equalizing power of randomization, then, is: *Whenever it is possible to do so, assign subjects to experimental groups and conditions randomly, and assign conditions and other factors to experimental groups randomly.*

The third method of controlling an extraneous variable is to build it right into the design as an independent variable. For example, assume that gender was to be controlled in the experiment discussed earlier and it was considered inexpedient or unwise to eliminate it. One could add a third independent variable, gender, to the design. Unless one were interested in the actual difference between the genders on the dependent variable or wanted to study the interaction between one or two of the other variables and gender, however, it is unlikely that this form of control would be used. One might want information of the kind just mentioned and also want to

control gender, too. In such a case, adding it to the design as a variable might be desirable. The point is that building a variable into an experimental design "controls" the variable, since it then becomes possible to extract from the total variance of the dependent variable the variance due to the variable. (In the above case, this would be the "between-genders" variance.)

These considerations lead to another principle: *An extraneous variable can be controlled by building it into the research design as an attribute variable, thus achieving control and yielding additional research information about the effect of the variable on the dependent variable and about its possible interaction with other independent variables.*

The fourth way to control extraneous variance is to match participants. The control principle behind matching is the same as that for any other kind of control, the control of variance. Matching is similar—in fact, it might be called a corollary—to the principle of controlling the variance of an extraneous variable by building it into the design. The basic principle is to split a variable into two or more parts in a factorial design, say into high and low intelligence, and then randomize within each level as described above. Matching is a special case of this principle. Instead of splitting the participants into two, three, or four parts, however, they are split into $N/2$ parts, N being the number of participants used; thus the control of variance is built into the design.

In using the matching method several problems may be encountered. To begin with, the variable on which the participants are matched must be substantially related to the dependent variable or the matching is a waste of time. Even worse, it can be misleading. In addition, matching has severe limitations. If we try to match, say, on more than two variables, or even more than one, we lose participants. It is difficult to find matched participants on more than two variables. For instance, if one decides to match intelligence, gender, and social class, one may be fairly successful in matching the first two variables but not in finding pairs that are fairly equal on all three variables. Add a fourth variable and the problem becomes difficult, often impossible to solve.

Let us not throw out the baby with the bath water, however. When there is a substantial correlation between the matching variable or variables and the dependent variable (>.50 or .60), then matching reduces the error term and thus increases the precision of an experiment, a desirable outcome. If the same participants are used with different experimental treatments—called repeated measures or randomized block design—we have powerful control of variance. How can one match better on all possible variables than by matching a subject with oneself? Unfortunately, other negative considerations usually rule out this possibility. It should be forcefully emphasized that matching of any kind is no substitute for randomization. If participants are matched, *they should then be assigned to experimental groups at random.* Through a random procedure, like tossing a coin or using odd and even random numbers, the members of the matched pairs are assigned to experimental and control groups. If the same participants undergo all treatments, then the order of the treatments should be assigned randomly. This adds randomization control to the matching, or repeated measures control.

A principle suggested by this discussion is: *When a matching variable is substantially correlated with the dependent variable, matching as a form of variance control can be*

profitable and desirable. Before using matching, however, carefully weigh its advantages and disadvantages in the particular research situation. Complete randomization or the analysis of covariance may be better methods of variance control.

Still another form of control, statistical control, was discussed at length in previous chapters, but one or two further remarks are in order here. Statistical methods are, so to speak, forms of control in the sense that they isolate and quantify variances. But statistical control is inseparable from other forms of design control. If matching is used, for example, an appropriate statistical test must be used, or the matching effect, and thus the control, will be lost.

Minimization of Error Variance

Error variance is the variability of measures due to random fluctuations whose basic characteristic is that they are self-compensating, varying now this way, now that way, now positive, now negative, now up, now down. Random errors tend to balance each other so that their mean is zero.

There are a number of determinants of error variance, for instance, factors associated with individual differences among participants. Ordinarily we call this variance due to individual differences "systematic variance." But when such variance cannot be, or is not identified and controlled, we have to lump it with the error variance. Because many determinants interact and tend to cancel each other out (or at least we assume that they do), the error variance has this random characteristic.

Another source of error variance is that associated with what are called errors of measurement: variation of responses from trial to trial, guessing, momentary inattention, slight temporary fatigue, lapses of memory, transient emotional states of participants, and so on.

Minimizing error variance has two principal aspects: (1) the reduction of errors of measurement through controlled conditions, and (2) an increase in the reliability of measures. The more uncontrolled the conditions of an experiment, the more the many determinants of error variance can operate. This is one of the reasons for carefully setting up controlled experimental conditions. In studies under field conditions, of course, such control is difficult; still, constant efforts must be made to lessen the effects of the many determinants of error variance. This can be done, in part, by specific and clear instructions to participants and by excluding from the experimental situation factors that are extraneous to the research purpose.

To increase the reliability of measures is to reduce the error variance. Pending fuller discussion later in the book, reliability can be taken to be the accuracy of a set of scores. To the extent that scores do not fluctuate randomly, they are reliable. Imagine a completely unreliable measurement instrument. This instrument does not allow us to predict the future performance of individuals. It gives a set of rank ordering values for a sample of participants at one time and a completely different set of rank ordering at another time. With such an instrument, it would not be possible to identify and extract systematic variances, since the scores yielded by the instrument would be like the numbers in a table of random numbers. This is the extreme case. Now, imagine differing amounts of reliability and unreliability in the measures of the

dependent variable. The more reliable the measures, the better we can identify and extract systematic variances and the smaller the error variance in relation to the total variance.

Another reason for reducing error variance as much as possible is to give systematic variance a chance to show itself. We cannot do this if the error variance, and thus the error term, is too large. If a relation exists, we seek to discover it. One way to discover the relation is to find significant differences between means. But if the error variance is relatively large due to uncontrolled errors of measurement, the systematic variance—earlier called "between" variance—will not have a chance to appear. Thus, the relation, although it exists, will probably not be detected.

The problem of error variance can be put into a neat mathematical nutshell. Remember the equation:

$$V_t = V_b + V_e$$

where V_t is the total variance in a set of measures; V_b is the between-groups variance, the variance presumably due to the influence of the experimental variables; and V_e is the error variance (in analysis of variance, the within-groups variance and the residual variance). Obviously, the larger V_e is, the smaller V_b must be, with a given amount of V_t.

Consider the following equation: $F = V_b/V_e$. For the numerator of the fraction on the right to be accurately evaluated for significant departure from chance expectation, the denominator should be an accurate measure of random error.

A familiar example may make this clear. Recall that in the discussions of factorial analysis of variance and the analysis of variance of correlated groups, we talked about variance due to individual differences being present in experimental measures. We said that, while adequate randomization can effectively equalize experimental groups, there will be variance in the scores due to individual differences, for instance, differences due to intelligence, aptitude, and so on. Now, in some situations, these individual differences can be quite large. If they are, then the error variance and, consequently, the denominator of the F equation above, will be "too large" relative to the numerator; that is, the individual differences will have been randomly scattered among, say, two, three, or four experimental groups. Still they are sources of variance and, as such, will inflate the within-groups or residual variance, the denominator of the above equation.

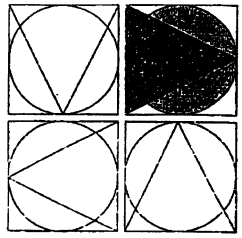
CHAPTER SUMMARY

1. Research designs are plans and structures used to answer research questions.
2. Research designs have two basic purposes: (i) provide answers to research questions, and (ii) control variance.
3. Research designs work in conjunction with research hypotheses to yield a dependable and valid answer.
4. Research designs can also tell us what statistical test to use to analyze the data collected from that design.

5. When speaking of controlling variance, we can mean one or more of three things:
   - maximize systematic variance
   - control extraneous variance
   - minimize error variance
6. To maximize systematic variance, one should have an independent variable where the levels are very distinct from one another.
7. To control extraneous variance the researcher need to eliminate the effects of a potential independent variable on the dependent variable. This can be done by:
   - holding the independent variable constant; for example, if one knows gender has a possible effect, gender can be held constant by doing the study with only one gender (i.e., females).
   - randomization; meaning to choose participants randomly and then assigning each group of participants to treatment conditions randomly (levels of the independent variable).
   - build the extraneous variable into the design by making it an independent variable.
   - matching participants—this method of control might be difficult in certain situations; a researcher will never be quite sure that a successful match was made on all of the important variables.
8. Minimizing error variance involves measurement of the dependent variable. By reducing the measurement error one will have reduced error variance. The increase in the reliability of the measurement would also lead to a reduction of error variance.

STUDY SUGGESTIONS

1. We have noted that research design has the purpose of obtaining answers to research questions and controlling variance. Explain in detail what this statement means. How does a research design control variance? Why should a factorial design control more variance than a one-way design? How does a design that uses matched participants or repeated measures of the same participants control variance? What is the relation between the research questions and hypotheses and a research design? Invent a research problem to illustrate your answers to these questions (or use an example from the text).
2. Sir Ronald Fisher (1951), the inventor of analysis of variance, said in one of his books, it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis. Whether you agree or disagree with Fisher's statement, what do you think he meant by it? In framing your answer, remember the maxmincon principle and *F*-tests and *t*-tests.



CHAPTER 19

INADEQUATE DESIGNS AND DESIGN CRITERIA

All disciplined creations of humans have form. Architecture, poetry, music, painting, mathematics, scientific research—all have form. People put great stress on the content of their creations, often not realizing that without streng structure, no matter how rich and how significant the content, the creations may be weak and sterile.

So it is with scientific research. The scientist needs viable and plastic form with which to express scientific aims. Without content—without good theory, good hypotheses, good problems—the design of research is empty. But without form, without structure adequately conceived and created for the research purpose, little of value can be accomplished. Indeed, it is no exaggeration to say that many of the failures of behavioral research have been failures of disciplined and imaginative form.

The principal focus of this chapter is on inadequate research designs. Such designs have been so common that they must be discussed. More important, the student should be able to recognize them and understand why they are inadequate.

This negative approach has a virtue: The study of deficiencies forces one to ask why something is deficient, which in turn centers attention on the criteria used to judge both adequacies and inadequacies. So the study of inadequate designs leads us to the study of the criteria of research design. We take the opportunity, too, to describe the symbolic system to be used, and to identify an important distinction between experimental and nonexperimental research.

Experimental and Nonexperimental Approaches

Discussion of design must be prefaced by an important distinction: that between experimental and nonexperimental approaches to research. Indeed, this distinction is so important that a separate chapter (Chapter 23) will be devoted to it later. An *experiment* is a scientific investigation in which an investigator manipulates and controls one or more independent variables and observes the dependent variable or variables for variation concomitant to the manipulation of the independent variables. An *experimental design*, then, is one in which the investigator *manipulates* at least one independent variable. In an earlier chapter we briefly discussed Hurlock's classic study (1925). Hurlock manipulated incentives to produce different amounts of retention. In the Walster, Cleary, and Clifford (1970) study (discussed in Chapter 18), sex, race, and ability levels were manipulated to study their effects on college acceptance: the application forms submitted to colleges differed in descriptions of applicants as male–female; white–black; and high, medium, or low ability levels.

In nonexperimental research one cannot manipulate variables or assign participants or treatments at random because the nature of the variables is such as to preclude manipulation. Participants come to us with their differing characteristics intact, so to speak. They come to us with their sex, intelligence, occupational status, creativity, or aptitude "already there." Wilson (1996) used a nonexperimental design to study the readability, ethnic content, and cultural sensitivity of patient education material used by nurses at local health department and community health centers. Here, the material preexisted. There was no random assignment or selection. Edmondson (1996) also used a nonexperimental design to compare the number of medication errors by nurses, physicians, and pharmacists in eight hospital units at two urban teaching hospitals. Edmondson did not choose these units or hospitals at random, neither were the medical professionals chosen at random. In many areas of research, likewise, random assignment is unfortunately not possible, as we will see later. Although experimental and nonexperimental research differ in these crucial respects, they share structural and design features that will be pointed out in this and subsequent chapters. In addition, their basic purpose is the same: to study relations among phenomena. Their scientific logic is also the same: to bring empirical evidence to bear on conditional statements of the form If p, then q. In some fields of behavioral and social sciences the nonexperimental framework is unavoidable. Keith (1988) states that a lot of studies conducted by school psychologists are of the nonexperimental nature. School psychology researchers as well as many in educational psychology must work within a practical framework. Many times, schools, classrooms,

or even students are given to the researcher "as-is." Stone-Romero, Weaver, and Glenar (1995) have summarized nearly 20 years of articles from the *Journal of Applied Psychology*, concerning the use of experimental and nonexperimental research designs.

The ideal of science is the controlled experiment. Except, perhaps, in taxonomic research—research with the purpose of discovering, classifying, and measuring natural phenomena and the factors behind such phenomena—where the controlled experiment is the desired model of science. It may be difficult for many students to accept this rather categorical statement since its logic is not readily apparent. Earlier it was said that the main goal of science was to discover relations among phenomena. Why then assign a priority to the controlled experiment? Do not other methods of discovering relations exist? Yes, of course they do. The main reason for the preeminence of the controlled experiment, however, is that researchers can have more confidence that the relations they study are the relations they think they are. The reason is not hard to see: They study the relations under the most carefully controlled conditions of inquiry known. The unique and overwhelmingly important virtue of experimental inquiry, then, is control. In a perfectly controlled experimental study, the experimenter can be confident that the manipulation of the independent variable affected the dependent variable and nothing else. In short, a perfectly conducted experimental study is more trustworthy than a perfectly conducted nonexperimental study. Why this is so should become more apparent as we advance in our study of research design.

Symbolism and Definitions

Before discussing inadequate designs, explanation of the symbolism to be used in these chapters is necessary. X is used to define an *experimentally manipulated* independent variable (or variables). X_1, X_2, X_3, and so on represent independent variables 1, 2, 3, and so on, though we usually use X alone, even when it can mean more than one independent variable. (We also use X_1, X_2, etc., to represent partitions of an independent variable, but the difference will always be clear.) The symbol (X) indicates that the independent variable is not *manipulated*—is not under the direct control of the investigator, but is *measured* or *imagined*. The dependent variable is Y: Y_b is the dependent variable *before* the manipulation of X, and Y_a the dependent variable *after* the manipulation of X. With $\sim X$, we borrow the negation sign of set theory: $\sim X$ ("not-X") to indicate that the experimental variable (the independent variable X) is *not* manipulated. [Note: (X) is a nonmanipulable variable and $\sim X$ is a manipulable variable that is not manipulated.] The symbol (R) will be used for the random assignment of participants to experimental groups and the random assignment of experimental treatments to experimental groups.

The explanation of $\sim X$ just given is not quite accurate because in some cases $\sim X$ can represent a different aspect of the treatment X, rather than merely the absence of treatment. In an older language, the experimental group was the group that was given the so-called experimental treatment, X; while the control group did not receive it, $\sim X$. For our purposes, however, $\sim X$ will do well enough, especially if we

understand the generalized meaning of *control* discussed below. An *experimental group*, then, is a group of participants receiving some aspect or treatment of X. In testing the frustration–aggression hypothesis, the experimental group is the group whose participants are systematically frustrated. In contrast, the control group is one that is given "no" treatment.

In modern multivariate research, it is necessary to expand these notions. They are not changed basically; they are only expanded. It is quite possible to have more than one experimental group, as we have seen. Different degrees of manipulation of the independent variable are not only possible, they are often also desirable or even imperative. Further, it is possible to have more than one control group, a statement that at first seems like nonsense. How can one have different degrees of "no" experimental treatment? This occurs because the notion of *control* is generalized. When there are more than two groups, and when any two of them are treated differently, one or more groups serve as "controls" on the others. Recall that control is always control of variance. With two or more groups treated differently, variance is engendered by the experimental manipulation. So the traditional notion of X and $\sim X$ (treatment and no treatment) is generalized to $X_1, X_2, X_3, \ldots, X_k$, different forms or degrees of treatment.

If X is encased inside parentheses (X), this means that the investigator "imagines" the manipulation of X, or assumes that X occurred and that it is the X of the hypothesis. It may also mean that X is measured and not manipulated. Actually, we are saying the same thing here in different ways. The context of the discussion should make the distinction clear. Suppose a sociologist is studying delinquency and the frustration–aggression hypothesis. The sociologist observes delinquency, Y, and imagines that the delinquent participants were frustrated in their earlier years, or (X). All nonexperimental designs will have (X). Generally, then, (X) represents an independent variable *not under the experimental control of the investigator.*

One more point—each design in this chapter will ordinarily have an *a* and a *b* form. The *a* form will be the experimental form, or that in which X is manipulated. The *b* form will be the nonexperimental form, that in which X is not under the control of the investigator, or (X). Obviously, $(\sim X)$ is also possible.

Faulty Designs

There are four (or more) inadequate designs of research that have often been used—and are occasionally still used—in behavioral research. The inadequacies of the designs lead to poor control of independent variables. We number each such design, give it a name, sketch its structure, and then discuss it.

Design 19.1: One Group

(*a*) X	Y	(Experimental)
(*b*) (X)	Y	(Nonexperimental)

Design 19.1(a) has been called the "One-Shot Case Study," an apropos name given by Campbell and Stanley (1963). The (*a*) form is experimental, the (*b*) form nonexperimental. An example of the (*a*) form: a school faculty institutes a new curriculum and wishes to evaluate its effects. After one year, *Y*, student achievement, is measured. It is concluded, say, that achievement has improved under the new program. With such a design the conclusion is weak. Design 19.1(b) is the nonexperimental form of the one-group design. *Y*, the outcome, is studied, and *X* is assumed or imagined. An example would be to study delinquency by searching the past of a group of juvenile delinquents for factors that may have led to their antisocial behavior. The method is problematic because the factors (variables) may be confounded. When the effect of two or more factors (variables) cannot be separated, the results are difficult to interpret. Any number of possible explanations might be plausible.

Scientifically, Design 19.1 is worthless. There is virtually no control of other possible influences on outcome. As Campbell (1957) pointed out, the minimum of useful scientific information requires at least one formal comparison. The curriculum example requires, *at the least*, comparison of the group that experienced the new curriculum with a group that did not experience it. The presumed effect of the new curriculum, say such-and-such achievement, might well have been about the same under any kind of curriculum. The point is not that the new curriculum did or did not have an effect. It was that without any formal, controlled comparison of the performance of the members of the "experimental" group with the performance of the members of some other group not experiencing the new curriculum, little can be said about its effect.

An important distinction should be made. It is not that the method is entirely worthless, but that it is *scientifically* worthless. In everyday life, of course, we depend on such scientifically questionable evidence; we have to. We act, we say, on the basis of our experience. We hope that we use our experience rationally. The everyday-thinking paradigm implied by Design 19.1 is not being criticized. Only when such a paradigm is used and said or believed to be scientific do difficulties arise. Even in high intellectual pursuits, the thinking implied by this design is used. Freud's careful observations and brilliant and creative analysis of neurotic behavior seem to fall into this category. The quarrel is not with Freud, then, but rather with assertions that his conclusions are "scientifically established."

Design 19.2: One Group, Before–After (Pretest, Posttest)

(*a*) Y_b	X	Y_a	(Experimental)
(*b*) Y_b	(X)	Y_a	(Nonexperimental)

Design 19.2 is only a small improvement on Design 19.1. The essential characteristic of this mode of research is that a group is compared to itself. Theoretically, there is no better choice, since all possible independent variables associated with the

participants' characteristics are controlled. The procedure dictated by such a design is as follows. A group is measured on the dependent variable, Y, before experimental manipulation. This is usually called a *pretest*. Assume that the attitudes toward women of a group of participants are measured. An experimental manipulation designed to change these attitudes is used. An experimenter might expose the group to expert opinion on women's rights, for example. After the interposition of this X, the attitudes of the participants are again measured. The difference scores, or $Y_a - Y_b$, are examined for change in attitudes.

At face value, this would seem a good way to accomplish the experimental purpose. After all, if the difference scores are statistically significant, does this not indicate a change in attitudes? The situation is not so prosaic. There are a number of other factors that may have contributed to the change in scores. Hence, the factors are confounded. Campbell (1957) gives an excellent detailed discussion of these factors, only a brief outline of which can be given here.

Measurement, History, Maturation

First is the possible effect of the measurement procedure: measuring participants changes them. Can it be that the post-X measures were influenced not by the manipulation of X but by increased sensitization due to the pretest? Campbell (1957) calls such measures *reactive* measures, because they themselves cause the subject to react. Controversial attitudes, for example, seem to be especially susceptible to such sensitization. Achievement measures, though probably less reactive, are still affected. Measures involving memory are susceptible. If you take a test now, you are more likely to remember later things that were included in the test. In short, observed changes may be due to reactive effects.

Two other important sources of extraneous variance are *history* and *maturation*. Between the Y_b and Y_a testings, many things can occur other than X. The longer the period of time, the greater the chance of extraneous variables affecting the participants, and thus the Y_a measures. This is what Campbell (1957) calls *history*. These variables or events are *specific* to the particular experimental situation. *Maturation*, on the other hand, covers events that are *general*—not specific to any particular situation. They reflect change or growth in the organism studied. Mental age increases with time, an increase that can easily affect achievement, memory, and attitudes. People can learn in any given time interval, and the learning may affect dependent variable measures. This is one of the exasperating difficulties of research that extends over considerable time periods. The longer the time interval, the greater the possibility that extraneous, unwanted sources of systematic variance will influence dependent variable measures.

The Regression Effect

A statistical phenomenon that has misled researchers is the so-called *regression effect*. Test scores change as a statistical fact of life: on retest, on the average, they regress toward the mean. The regression effect operates because of the imperfect correlation

between the pretest and posttest scores. If $r_{ab} = 1.00$, then there is no regression effect; if $r_{ab} = .00$, the effect is at a maximum in the sense that the best prediction of any posttest score from pretest score is the mean. With the correlations found in practice, the net effect is that lower scores on the pretest tend to be higher, and higher scores lower on the posttest—when, in fact, no real change has taken place in the dependent variable. Thus, if low-scoring participants are used in a study, their scores on the posttest will probably be higher than on the pretest due to the regression effect. This can deceive the researcher into believing that the experimental intervention has been effective when it really has not. Similarly, one may erroneously conclude that an experimental variable has had a depressing effect on high pretest scorers. Not necessarily so. The higher and lower scores of the two groups may be due to the regression effect. How does this work? There are many chance factors at work in any set of scores. Two excellent references on the discussion of the regression effect are Anastasi (1958) and Thorndike (1963). For a more statistically sophisticated presentation, see Nesselroade, Stigler, and Baltes (1980). On the pretest some high scores are higher than "they should be" due to chance, and similarly with some low scores. On the posttest it is unlikely that the high scores will be maintained, because the factors that made them high were chance factors—which are uncorrelated on the pretest and posttest. Thus the high scorer will tend to drop on the posttest. A similar argument applies to the low scorer—but in reverse.

Research designs have to be constructed with the regression effect in mind. There is no way in Design 19.2 to control it. If there was a control group, then one could "control" the regression effect, since both experimental and control groups have pretest and posttest. If the experimental manipulation has had a "real" effect, then it should be apparent over and above the regression effect. That is, the scores of both groups, other things being equal, are affected the same by regression and other influences. So if the groups differ in the posttest, it should be due to the experimental manipulation.

Design 19.2 is inadequate, not so much because extraneous variables and the regression effect can operate (the extraneous variables operate whenever there is a time interval between pretest and posttest), but *because we do not know whether they have operated, whether they have affected the dependent variable measures.* The design affords no opportunity to control or to test such possible influences.

Design 19.3: Simulated Before–After

	X	Y_a
Y_b		

The peculiar title of Design 19.3 stems in part from its very nature. Like Design 19.2 it is a before–after design. Instead of using the before and after (or pretest–posttest) measures of one group, we use as pretest measures the measures of another group, which are chosen to be as similar as possible to the experimental

group, and thus a control group of a sort. (The line between the two levels above indicates separate groups.) This design satisfies the condition of having a control group, and is thus a gesture toward the comparison that is necessary to scientific investigation. Unfortunately, the controls are weak, a result of our inability to know that the two groups were equivalent before X, the experimental manipulation.

Design 19.4: Two Groups, No Control

(a)	X	Y	(Experimental)
	$\sim X$	$\sim Y$	
(b)	(X)	Y	(Nonexperimental)
	$(\sim X)$	$\sim Y$	

Design 19.4 is common. In (*a*) the experimental group is administered treatment X. The "control" group, taken to be, or assumed to be, similar to the experimental group, is not given X. The Y measures are compared to ascertain the effect of X. Groups or participants are taken "as they are," or they may be matched. The nonexperimental version of the same design is labeled (*b*). An effect, Y, is observed to occur in one group (top line) but not in another group, or to occur in the other group to a lesser extent (indicated by the $\sim Y$ in the bottom line). The first group is found to have experienced X, the second group not to have experienced X.

This design has a basic weakness: The two groups are *assumed* to be equal in independent variables other than X. It is sometimes possible to check the equality of the groups roughly by comparing them on different pertinent variables, for example, age, sex, income, intelligence, ability, and so on. This should be done if it is at all possible, but, as Stouffer (1950, p. 522) says, "there is all too often a wide-open gate through which other uncontrolled variables can march." Because randomization is not used—that is, the participants are not assigned to the groups at random—it is not possible to assume that the groups are equal. Both versions of the design suffer seriously from lack of control of independent variables due to lack of randomization.

Criteria of Research Design

After examining some of the main weaknesses of inadequate research designs, we are in a good position to discuss what can be called *criteria* of research design. Along with the criteria, we will enunciate certain principles that should guide researchers. Finally, the criteria and principles will be related to Campbell's (1957) notions of internal and external validity, which, in a sense, express the criteria another way.

Answer Research Questions?

The main criterion or desideratum of a research design can be expressed in a question: *Does the design answer the research questions?* or *Does the design adequately test the*

hypotheses? Perhaps the most serious weakness of designs often proposed by the neophyte is that they are not capable of answering the research questions adequately. A common example of this lack of congruence between the research questions and hypothesis, on the one hand, and the research design, on the other, is matching participants for reasons irrelevant to the research and then using an experimental group—control type of design. For instance, students often assume that because they match pupils on intelligence and sex that their experimental groups are equal. They have heard that one should match participants for "control" and that one should have an experimental group and a control group. Frequently, however, the matching variables may be irrelevant to the research purposes. That is, if there is no relation between, say, sex and the dependent variable, then matching on sex is irrelevant.

Another example of this weakness is the case where three or four experimental groups are needed. For example, three experimental groups and one control group, or four groups with different amounts or aspects of X, the experimental treatment is required. However, the investigator uses only two because he or she has heard that an experimental group and a control group are necessary and desirable.

The example discussed in Chapter 18 of testing an interaction hypothesis by performing, in effect, two separate experiments is another example. The hypothesis to be tested was that discrimination in college admissions is a function of both sex and ability level, that it is women of low ability who are excluded (in contrast to men of low ability). This is an interaction hypothesis and probably calls for a factorial-type design. To set up two experiments, one for college applicants of high ability and another for applicants of low ability, is poor practice because such a design, as shown earlier, cannot decisively test the stated hypothesis. Similarly, to match participants on ability and then set up a two-group design would miss the research question entirely. These considerations lead to a general and seemingly obvious precept:

Design research to answer research questions.

Control of Extraneous Independent Variables

The second criterion is *control*, which refers to control of independent variables: the independent variables of the research study and extraneous independent variables. Extraneous independent variables are, of course, variables that may influence the dependent variable but that are not part of the study. Such variables are confounded with the independent variable under study. In the admissions study of Chapter 18, for example, geographical location (of the colleges) may be a potentially influential extraneous variable that can cloud the results of the study. If colleges in the east, for example, exclude more women than colleges in the west, then geographical location is an extraneous source of variance in the admissions measures—which should somehow be controlled. The criterion also refers to control of the variables of the study. Since this problem has already been discussed and will continue to be discussed, no more need be said here. But the question must be asked: *Does this design adequately control independent variables?*

The best single way to answer this question satisfactorily is expressed in the following principle:

> *Randomize whenever possible: select participants at random; assign participants to groups at random; assign experimental treatments to groups at random.*

While it may not be possible to select participants at random, it may be possible to assign them to groups at random; thus "equalizing" the groups in the statistical sense discussed in earlier chapters. If such random assignment of participants to groups is not possible, then every effort should be made to assign experimental treatments to experimental groups at random. And, if experimental treatments are administered at different times with different experimenters, times and experimenters should be assigned at random.

The principle that makes randomization pertinent is complex and difficult to implement:

> *Control the independent variables so that extraneous and unwanted sources of systematic variance have minimal opportunity to operate.*

As we have seen earlier (Chapter 8), randomization theoretically satisfies this principle. When we test the empirical validity of an If p, then q proposition, we manipulate p and observe that q covaries with the manipulation of p. But how confident can we be that our If p, then q statement is really "true"? Our confidence is directly related to the completeness and adequacy of the controls. If we use a design similar to designs 19.1 through 19.4, we cannot have too much confidence in the empirical validity of the If p, then q statement, since our control of extraneous independent variables is weak or nonexistent. Because such control is not always possible in much psychological, sociological, and educational research, should we then give up research entirely? By no means. But we must be aware of the weaknesses of intrinsically poor design.

Generalizability

The third criterion, *generalizability*, is independent of other criteria because it is different in kind. This is an important point that will shortly become clear. It means simply: *Can we generalize the results of a study to other participants, other groups, and other conditions?* Perhaps the question is better put: *How much* can we generalize the results of the study? This is probably the most complex and difficult question that can be asked of research data because it touches not only on technical matters (like sampling and research design), but also on larger problems of basic and applied research. In basic research, for example, generalizability is not the first consideration, because the central interest is the relations among variables and why the variables are related as they are. This emphasizes the internal rather than the external aspects of the study. These studies are often designed to examine theoretical issues such as motivation or learning. The goal of basic research is to add information and knowledge

to a field of study, but usually without a specific practical purpose. Its results are generalizable, but not in the same realm as results found in applied research studies. In applied research, on the other hand, the central interest forces more concern for generalizability, because one certainly wishes to apply the results to other persons and to other situations. Applied research studies usually have their foundations in basic research studies. Using information found in a basic research study, applied research studies apply those findings to determine if it can solve a practical problem. Take the work of B. F. Skinner for example. His early research is generally considered as basic research. It was from his research that schedules of reinforcement were established. However, later, Skinner and others (Skinner, 1968; Garfinkle, Kline, & Stancer, 1973) applied the schedules of reinforcement to military problems, educational problems, and behavioral problems. Those who do research on the modification of behavior are applying many of the theories and ideas tested and established by B. F. Skinner. If the reader will ponder the following two examples of basic and applied research, he or she can get closer to this distinction.

In Chapter 14 we examined a study by Johnson (1994) on rape type, information admissibility and perception of rape victims. This is clearly basic research: the central interest was in the relations among rape type, information admissibility, and perception. While no one would be foolish enough to say that Johnson was not concerned with rape type, information admissibility, and perception in general, the emphasis was on the relations among the variables of the study. Contrast this study with the effort of Walster et al. (1970) to determine whether colleges discriminate against women. Naturally, Walster and her colleagues were particular about the internal aspects of their study. But they perforce had to have another interest: Is discrimination practiced among colleges in general? Their study is clearly applied research, though one cannot say that basic research interest was absent. The considerations of the next section may help to clarify generalizability.

Internal and External Validity

Two general criteria of research design have been discussed at length by Campbell (1957) and by Campbell and Stanley (1963). These notions constitute one of the most significant, important, and enlightening contributions to research methodology in the past three or four decades.

Internal validity asks the question: Did *X*, the experimental manipulation, really make a significant difference? The three criteria of Chapter 18 are actually aspects of internal validity. Indeed, anything affecting the *controls* of a design becomes a problem of internal validity. If a design is such that one can have little or no confidence in the relations, as shown by significant differences between experimental groups, this is a problem of internal validity.

Earlier in this chapter we presented four possible threats to internal validity. Some textbook authors have referred to these as "alternative explanations" (see Dane, 1990) or "rival hypotheses" (see Graziano & Raulin, 1993). These were listed as measurement, history, maturation, and statistical regression. Campbell and Stanley (1963) also list four other threats. They are instrumentation, selection,

attrition, and the interaction between one of more of those previously listed (total of eight).

Instrumentation is a problem if the device used to measure the dependent variable changes over time. This is particularly true in studies using a human observer. Human observers or judges can be affected by previous events or fatigue. Observers may become more efficient over time, and thus the later measurements are more accurate than earlier ones. On the other hand, with fatigue, the human observer would become less accurate in the later trials than the earlier ones. When this happens, the values of the dependent variable will change and that change will not be due solely to the manipulation of the independent variable.

With selection, Campbell and Stanley (1963) are talking about the type of participants the experimenter selects for the study. This is especially likely if the researcher is not careful in studies that do not use random selection or assignment. The researcher could have selected participants in each group that are very different on some characteristic, and as such could account for a difference in the dependent variable. It is important for the researcher to have the groups equal prior to the administration of treatment. If the groups are the same before treatment, then logic follows that if they are different following treatment then it was the treatment (independent variable) that caused the difference and not something else. However, if the groups are different to begin with and different after treatment it is very difficult to make a statement that the difference was due to treatment. Later when discussing quasi-experimental designs, we will see how we can strengthen the situation.

Attrition or experimental mortality deals with the drop out of participants. If too many participants in one treatment condition leave the study, the unbalance is a possible reason for the change in the dependent variable. Attrition also includes the departure of participants with certain characteristics.

Any of the previous seven threats to internal validity could also interact with one another. Selection could interact with maturation. This threat is especially possible when using participants who are volunteers. If the researcher is comparing two groups—one group consists are volunteers (self-selected), the other group consists of nonvolunteers—the performance between these two on the dependent variable may be due to the fact that volunteers are more motivated. Student researchers sometimes use the volunteer subject pool and members of their own family or social circle as participants. There may be a problem of internal validity if volunteers are placed in one treatment group and their friends are put into another.

A difficult criterion to satisfy—*external validity*—defines *representativeness* or *generalizability*. When an experiment has been completed and a relation found, to what populations could it be generalized? Can we say that *A* is related to *B* for all schoolchildren? All eighth-grade children? All eighth-grade children in this school system? or All eighth-grade children of only this school? Or must the findings be limited to the eighth-grade children with whom we worked? These very important scientific questions should always be *asked and answered.*

Not only must sample generalizability be questioned, it is also necessary to ask questions about the ecological and variable representativeness of studies. If the social setting in which the experiment was conducted is changed, will the relation of *A* and

B still hold? Will *A* be related to *B* if the study is replicated in a lower-class school? In a western school? In a southern school? These are questions of *ecological representativeness.*

Variable representativeness is more subtle. A question not often asked, but that should be asked, is: Are the variables of this research representative? When an investigator works with psychological and sociological variables, one assumes that the variables are "constant." If the investigator finds a difference in achievement between boys and girls, one can assume that sex as a variable is "constant."

In the case of variables like achievement, aggression, aptitude, and anxiety, can the investigator assume that the "aggression" of the suburban participants is the same "aggression" to be found in city slums? Is the variable the same in a European suburb? The representativeness of "anxiety" is more difficult to ascertain. When we talk of "anxiety," what kind of anxiety do we mean? Are all kinds of anxiety the same? If anxiety is manipulated in one situation by verbal instructions and in another situation by electric shock, are the two induced anxieties the same? If anxiety is manipulated by, say, experimental instruction, is this the same anxiety as that measured by an anxiety scale? Variable representativeness, then, is another aspect of the larger problem of external validity, and thus of generalizability.

Unless special precautions are taken and special efforts made, the results of research are frequently not representative, and hence not generalizable. Campbell and Stanley (1963) say that internal validity is the sine qua non of research design, but that the ideal design should be strong in both internal validity and external validity, even though they are frequently contradictory. This point is well taken. In these chapters, the main emphasis is on internal validity, with a vigilant eye on external validity.

Campbell and Stanley (1963) present four threats to external validity. They are reactive or interaction effects of testing, the interaction effects of selection biases and the independent variable, reactive effects of experimental arrangements and multiple-treatment interference.

In the reactive or interaction effect of testing, the reference is to the use of a pretest prior to administering treatment. Pretesting may decrease or increase the sensitivity of the participant to the independent variable. This would make the results for the pretested population unrepresentative of the treatment effect for the nonpretested population. The likelihood of an interaction between treatment and pretesting seems first to have been pointed out by Solomon (1949).

The interaction effects of selection bias and the independent variable indicates that selection of participants can very well affect generalization of the results. A researcher using only participants from the subject pool at a particular university, which usually consists of freshmen and sophomores, will find it difficult to generalize the findings of the study to other students in the university or at other universities.

The mere participation in a research study can be a problem in terms of external validity. The presence of observers, instrumentation, or laboratory environment could have an effect on the participant that would not occur if the participant was in a natural setting. The fact that one is participating in an experimental study may alter

one's normal behavior. Whether the experimenter is male or female, African American or white American could also have an effect.

If participants are exposed to more than one treatment condition, performance on later trials is affected by performance on earlier trials. Hence, the results can only be generalized to people who have had multiple exposures given in the same order.

The negative approach of this chapter was taken in the belief that an exposure to poor but commonly used and *accepted* procedures, together with a discussion of their major weaknesses, would provide a good starting point for the study of research design. Other inadequate designs are possible, but all such designs are inadequate on structural principles alone. This point should be emphasized because in Chapter 20 we will find that a perfectly good design structure can be poorly used. Thus it is necessary to learn and understand the two sources of research weakness: intrinsically poor designs and intrinsically good designs poorly used.

CHAPTER SUMMARY

1. Studying faulty designs helps researchers design better studies by knowing what pitfalls to avoid.
2. Nonexperimental designs are those with nonmanipulated independent variables, absence of random assignment or selection.
3. Faulty designs include the "one-shot case study," the one group before–after design, simulated before–after design, and the two group no-control design.
4. Faulty designs are discussed in terms of internal validity.
5. Internal validity is concerned with how strongly the experimenter can state the effect of the independent variable on the dependent variable. The more confidence the experimenter has about the manipulated independent variable, the stronger the internal validity.
6. Nonexperimental studies are weaker in internal validity than experimental studies.
7. There are eight basic classes of extraneous variables which, if not controlled, may be confounded with the independent variable. These eight basic classes are called threats to internal validity.
8. Campbell's threats to internal validity can be outlined as follows:

   - History
   - Maturation
   - Testing or Measurement
   - Instrumentation
   - Statistical Regression
   - Selection
   - Experimental Mortality or Attrition
   - Selection–Maturation Interaction

9. External validity is concerned with how strong a statement the experimenter can make about the generalizability of the results of the study.
10. Campbell and Stanley give four possible sources of threats to external validity:

    - Reactive or interaction effect of testing
    - Interaction effects of selection biases and the independent variable
    - Reactive effects of experimental arrangements
    - Multiple-treatment interference

STUDY SUGGESTIONS

1. Suppose a liberal arts college decides to begin a new curriculum for all undergraduates. It asks the faculty to form a research group to study the program's effectiveness for two years. The research group, wanting to have a group with which to compare the new curriculum group, requests that the present program be continued for two years and that students be allowed to volunteer for the present or the new program. The research group believes that it will then have an experimental group and a control group.
   Discuss the research group's proposal critically. How much faith would you have in the findings at the end of two years? Give reasons for your positive or negative reactions to the proposal.
2. Imagine that you are a graduate school professor and have been asked to judge the worth of a proposed doctoral thesis. The doctoral student is a school superintendent who is instituting a new type of administration into her school system. She plans to study the effects of the new administration for a three-year period and then write her thesis. She says that she will not study any other school situation during the period so as not to bias the results. Discuss the proposal. When doing so, ask yourself: Is the proposal suitable for doctoral work?
3. In your opinion should all research be held rather strictly to the criterion of generalizability? Explain why or why not. Which field is likely to have more basic research: psychology or education? Why? What implications do your conclusions have for generalizability?
4. What does replication of research have to do with generalizability? Explain. If it were possible, should all research be replicated? Explain why or why not. What does replication have to, do with external and internal validity?