STUDY SUGGESTIONS

1. Some feel that society has placed too many restrictions on scientists on how to conduct their research. List the strong and weak points behind these regulations.
2. What is the purpose of debriefing? Why is it necessary?
3. A student who is a fan of daytime talk shows wants to determine if the way a woman dresses influences men’s behavior. She plans to attend two bars on a single night. In one bar she will dress provocatively and in the other she will dress in a business suit. Her dependent variable is the number of men who approach and talk to her. Do you see any ethical problems with this study design?
4. Visit the library and try to locate material pertaining to other incidences of fraud and unethical behavior by behavioral and medical scientists. How many of these can you find?
5. In the novel Arrowsmith, can you propose an alternative method that would have enabled Martin Arrowsmith to fully test his phage?
6. Locate and read at least one of the following articles:


---

Research Design: Purpose and Principles

- **PURPOSES OF RESEARCH DESIGN**
  - An Example: A Stronger Design

- **RESEARCH DESIGN AS VARIANCE CONTROL**
  - A Controversial Example

- **MAXIMIZATION OF EXPERIMENTAL VARIANCE**

- **CONTROL OF EXTRANEOUS VARIABLES**

- **MINIMIZATION OF ERROR VARIANCE**

*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, and female A. 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, $M_A$ and $M_B$, will be tested for statistical significance with a $t$- or $F$-test. The substantive hypothesis is: $M_A > M_B$, or more males than females will be accepted for admission. If there is no discrimination in admissions, then $M_A$, is statistically equal to $M_B$.

Figure 18.1 indicates that the means are not significantly different. Can we 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 18.2

<table>
<thead>
<tr>
<th>Gender</th>
<th>A₁ (Male)</th>
<th>A₂ (Female)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ability</td>
<td>B₁ (High)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B₂ (Medium)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B₃ (Low)</td>
<td></td>
</tr>
<tr>
<td>Acceptance Scores</td>
<td>M₁₁</td>
<td>M₁₂</td>
</tr>
<tr>
<td></td>
<td>M₂₁</td>
<td>M₂₂</td>
</tr>
<tr>
<td></td>
<td>M₃₁</td>
<td>M₃₂</td>
</tr>
</tbody>
</table>

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 \times 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 A₁B₁ 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 A₁ and A₂, and B into B₁, B₂, and B₃.

Chapter 18 - Research Design: Purpose and Principles

The paradigm of Figure 18.2 suggests a number of things. First and most obvious, a fairly large number of participants is needed. Specifically, 60 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 contain 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 - β, where β is the probability of committing a Type II error. If power is high (close to 1.00), this says that if the statistical test was not significant, the researcher 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 A₁ and A₂, but not to B₁, B₂, and B₃.

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 test paper, and hence teachers may unknowingly (or knowingly) “help” students with their paper, and hence teachers may unknowingly (or knowingly) “help” students with their paper, and hence teachers may unknowingly (or knowingly) “help” students with their paper.
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 \( (A) \) may be accepted more than males \( (A) \) at the higher ability level \( (B) \), whereas they may be accepted less at the lower ability level \( (B) \).

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

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 \( M_{A_1} > M_{A_2} \bigg| B_1 \) \( (M_{A_1} > M_{A_2} \bigg| B_2) \), 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 \( B_1, B_2 \), and \( B_3 \). 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 \( B_3 \), 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
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, 1973; 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—for which 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 (M) 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 dichotomized 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:

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Dependent Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Methods</strong></td>
<td><strong>Aptitude</strong></td>
</tr>
<tr>
<td>Mastery Learning, $A_1$</td>
<td>High Aptitude, $B_1$</td>
</tr>
<tr>
<td>Traditional, $A_2$</td>
<td>Low Aptitude, $B_2$</td>
</tr>
</tbody>
</table>

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
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 assigned randomly 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 “maximin” 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 maximin 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 $A_1$, $A_2$, 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 extraneous variance that is due to numerous sources and chance. Remembering this subprinciple of the maximin 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. If, however, 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 $A_1$ 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 success-fully 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 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-gender" 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 empha-sized 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:

\[ V_i = V_c + V_e. \]

where \( V_i \) is the total variance in a set of measures; \( V_c \) 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 \( V_c \) is, the smaller \( V_e \) must be, with a given amount of \( V_i \).

Consider the following equation: \( F = V_c/V_e. \) 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 needs 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 maxim in principle and F-tests and t-tests.
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 in tact, 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 in local health departments 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—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 $Y_1, Y_2$, etc., to represent parts of dependent variables, but the difference will always be clear.) The symbol $(X)$ indicates that the independent variable is not manipulated—i.e., not under the direct control of the investigator, but is measured or imagined. The dependent variable is $Y$: $Y_i$ is the dependent variable before the manipulation of $X$, and $Y_j$, the dependent variable after the manipulation of $X$. With $X$, we borrow the negation sign of set theory: $\neg X$ ("not-$X$") to indicate that the experimental variable (the independent variable $X$) is not manipulated. [Note: $(X)$ is a nonmanipulatable variable and $\neg X$ is a manipulatable 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 $\neg X$ just given is not quite accurate because in some cases $\neg X$ 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, $\neg X$. For our purposes, however, $\neg X$ 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, but they are often 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 ~X (treatment and no treatment) is generalized to X₁, X₂, X₃, ... Xₙ, different forms or degrees of treatment.

If X is enclosed 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 delinquent behavior, 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, in which X is not under the control of the investigator, or (X). Obviously, (~X) is also possible.

Faulty Designs

There are four (or more) inadequate designs of research that have 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

<table>
<thead>
<tr>
<th>(a) X</th>
<th>Y'</th>
<th>(Experimental)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(b) (X)</td>
<td>Y'</td>
<td>(Nonexperimental)</td>
</tr>
</tbody>
</table>

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 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 no 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 embodied in this design is used.

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

| (a) Y₀ X Y' | (Experimental) |
| (b) Y₀ (X) Y' | (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
Designs of Research

PART SIX

between the pretest and posttest scores. If \( r_{xy} = 1.00 \), then there is no regression effect; if \( r_{xy} = .00 \), 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 Neseroorde, 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 is 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 by the 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 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.

<table>
<thead>
<tr>
<th>Design 19.3: Simulated Before–After</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Y</td>
</tr>
</tbody>
</table>

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)  

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>~X</td>
<td>~Y</td>
</tr>
</tbody>
</table>

(Experimental)

(b)  

<table>
<thead>
<tr>
<th></th>
<th>(X)</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>~X</td>
<td>~Y</td>
</tr>
</tbody>
</table>

(Nonexperimental)

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 ~ 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 group 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. If such random assignment of participants to groups is not possible, then every effort should be made to assign experimental treatments to 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 to 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 and observe that $q$ covaries with the manipulation of $p$. But how confident can we to the completeness and adequacy of the controls? If we use a design similar to validity of the $q$ statement, since our control of extraneous independent 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 other conditions? Perhaps the question is better put: How much can we generalize the can be asked of research data because it touches not only on technical matters (like search). In basic research, for example, generalizability is not the first consideration, related as they are. This emphasizes the internal rather than the external aspects of 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 focuses 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; Garfield, Kline, & Staner, 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. Which 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 perform to have another interest: Is discrimination practiced among colleges in general? Their study is clearly an applied research, though none 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 control 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. 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 imbalance 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 of 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? 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 urban 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 bias and the independent variable, reactive effects of experimental arrangements and multiple-treatment differences.

If 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. This 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
   - Maturity
   - Testing or Measurement
   - Instrumentation
   - Statistical Regression
   - Selection
   - Experimental Mortality or Attrition
   - Selection—Maturity 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?
STUDY SUGGESTIONS

1. How does generalizability theory differ from classical test theory?
2. Of the following, which do you think is more useful for researchers: (a) validity, or (b) reliability? Justify your choice.
3. Outline some of the problems with (a) test–retest reliability, and (b) parallel forms reliability. Give an example where you would and would not use each of these.
4. Given the following situations listed below, indicate which reliability coefficient would be the most appropriate?
   a. A typing test given to a word-processing class
   b. A psychological problem checklist used by therapists
   c. A cognitive achievement test
   d. A spelling test on four-letter words
   e. The number of “aggressive” acts by a male monkey in a zoo during the same 10-minute time period each day
   f. After a group of students completed a test, the test was divided into two parts and separate scores were computed for each student; the correlation of the two scores was .79
5. How many different components can you arrive at that would be part of the error term in the classical test theory equation: \( X = X_s + X_e \)?
6. Give an explanation as to why a “true” score or measurement can never be attained.
7. A split-half reliability is .7. What is the estimated full-length reliability?
8. If a test–retest reliability of a 50-item test is .65, what would be the estimated reliability if an additional 50 similar items were added to the test?

CHAPTER 28
VALIDITY

- TYPES OF VALIDITY
  - Content Validity and Content Validation
  - Criterion-Related Validity and Validation
  - Decision Aspects of Validity
  - Multiple Predictors and Criteria
  - Construct Validity and Construct Validation
  - Convergence and Discriminability
    - A Hypothetical Example of Construct Validation
  - The Multitrait–Multimethod Method
    - Research Examples of Criterion-Related Validation
    - Research Examples of Construct Validation
    - Other Methods of Construct Validation
- A VARIANCE DEFINITION OF VALIDITY: THE VARIANCE RELATION OF PROBABILITY AND VALIDITY
  - Statistical Relation between Reliability and Validity
- THE VALIDITY AND RELIABILITY OF PSYCHOLOGICAL AND EDUCATIONAL MEASUREMENT INSTRUMENTS

The subject of validity is complex, controversial, and peculiarly important in behavioral research. Here, perhaps more than anywhere else, the nature of reality is questioned. It is possible to study reliability without inquiring into the meaning of variables. It is not possible to study validity, however, without sooner or later inquiring into the nature and meaning of one’s variables.

When measuring certain physical properties and relatively simple attributes of persons, validity is no great problem. There is often rather direct and close congruence between the nature of the object measured and the measuring instrument. The length of an object, for example, can be measured by laying sticks marked in a
standard numbering system (feet or meters) on the object. Weight is more indirect, but not difficult: An object placed in a container displaces the container downward. The downward movement of the container is registered on a calibrated index (pounds or ounces). With some physical attributes, then, there is little doubt of what is being measured.

On the other hand, suppose an educational scientist wishes to study the relation between intelligence and school achievement or the relation between authoritarianism and teaching style. Now there are no rulers to use, no scales with which to weigh the degree of authoritarianism, no clear-cut physical or behavioral attributes to point unmistakably to teaching style. In such cases it is necessary to invent indirect means to measure psychological and educational properties. These means are often so indirect that the validity of the measurement and its products is doubtful.

Types of Validity

The most common definition of validity is epitomized by the question: Are we measuring what we think we are measuring? The emphasis in this question is on what is being measured. For example, a teacher has constructed a test to measure understanding of scientific procedures and has included in the test only factual items about scientific procedures. The test is not valid because, while it may reliably measure the pupils’ factual knowledge of scientific procedures, it does not measure their understanding of such procedures. In other words, it may measure what it measures quite well, but it does not measure what the teacher intended it to measure.

Although the most common definition of validity was given above, it must immediately be emphasized that there is no one validity. A test or scale is valid for the scientific or practical purpose of its use. Educators may be interested in the nature of high school pupils’ achievement in mathematics. They would then be interested in what a mathematics achievement or aptitude test measures. They might, for example, want to know the factors that enter into mathematics test performance and their relative contributions to this performance. On the other hand, they may be primarily interested in knowing the pupils who will probably be successful, and those who will probably be unsuccessful in high school mathematics. They may have little interest in what a mathematics aptitude test measures. They are interested mainly in successful prediction. Implied by these two uses of tests are different kinds of validity. We now examine an extremely important development in test theory: the analysis of study of different kinds of validity. Although there are different types, the researcher should design the validation study with only one type in mind. Some researchers consider all validity coefficients only to discover that each gives a different value.

The most important classification of types of validity is that prepared by a joint committee of the American Psychological Association, the American Educational Research Association, and the National Council on Measurements used in Education. Three types of validity are discussed: content, criterion-related, and construct. Each of these will be examined briefly, though we put the greatest emphasis on construct validity, since it is probably the most important form of validity from the scientific research point of view.

Content Validity and Content Validation

A university psychology professor has given a course to seniors in which she has emphasized the understanding of principles of human development. She prepares an objective-type test. Wanting to know something of its validity, she critically examines each of the test’s items for its relevance to understanding principles of human development. She also asks two colleagues to evaluate the content of the test. Naturally, she tells the colleagues what it is she is trying to measure. She is investigating the content validity of the test.

Content validity is the representativeness or sampling adequacy of the content—the substance, the matter, the topic—of a measuring instrument. Content validation is guided by the question: Is the substance or content of this measure representative of the content or the universe of content of the property being measured? Any psychological or educational property has a theoretical universe of content consisting of all the things that can possibly be said or observed about the property. The members of this universe, U, can be called “items.” The property might be “arithmetic achievement,” to take a relatively easy example. U has an infinite number of members: all possible items using numbers, arithmetic operations, and concepts. A test high in content validity would theoretically be a representative sample of U. If it was possible to draw items from U at random in sufficient numbers, then any such sample of items would presumably form a test high in content validity. If U consists of subsets A, B, and C, which are arithmetic operations, arithmetic concepts, and number manipulations, respectively, then any sufficiently large sample of U would represent A, B, and C approximately equally. The test’s content validity would be satisfactory.

Ordinarily, and unfortunately, it is not possible to draw random samples of items from a universe of content. Such universes exist only theoretically. True, it is possible and desirable to assemble large collections of items, especially in the achievement area, and to draw random samples from the collections for testing purposes. But the content validity of such collections, no matter how large and how “good” the items, is always in question.

If it is not possible to satisfy the definition of content validity, how can a reasonable degree of content validity be achieved? Content validation consists essentially in judgment. Alone or with others, one judges the representativeness of the items. One may ask: Does this item measure Property M? To express it more fully, one might ask: Is this item representative of the universe of content of M? If U has subsets, such as those indicated above, then one has to ask additional questions; for example: Is this item a member of the subset M1 or the subset M2?

Some universes of content are more obvious and much easier to judge than others; the content of many achievement tests, for instance, would seem to be obvious. The content validity of these tests, it is said, can be assumed. While this statement seems reasonable, and while the content of most achievement tests is “self-validated” in the sense that the individual writing the test, to a degree, defines the
property being measured (for example, a teacher writing a classroom test of spelling or arithmetic), it is dangerous to assume the adequacy of content validity without systematic efforts to check the assumption. For example, an educational investigator, testing hypotheses about the relations between social studies achievement and other variables, may assume the content validity of a social studies test. The theory from which the hypotheses were derived, however, may require understanding and application of social studies ideas, whereas the test used may be almost purely factual in content. The test lacks content validity for the purpose. In fact, the investigator is not really testing the stated hypotheses.

Content validation, then, is basically judgmental. The items of a test must be studied, each item being weighed for its presumed representativeness of the universe. This means that each item must be judged for its presumed relevance to the property being measured, which is no easy task. Usually other “competent” judges should judge the content of the items. The universe of content must, if possible, be clearly defined; that is, the judges must be furnished with specific directions for making judgments, as well as with specification of what they are judging. Then, some method for pooling independent judgments can be used. An excellent guide to the content validity of achievement tests is Bloom (1956). This is a comprehensive attempt to outline and discuss educational goals in relation to measurement. Bloom’s work has been termed “Bloom’s Taxonomy.”

There is another type of validity that is very similar to content validity. It is called face validity. Face validity is not validity in the technical sense. It refers to what the test appears to measure. Trained or untrained individuals would look at the test and decide whether or not the test measures what it was supposed to measure. There is no quantification of the judgment or any index of agreement that is computed between judges. Content validity is quantifiable through the use of agreement indices of judges’ evaluations. One such index is Cohen’s Kappa (Cohen, 1960).

Criterion-Related Validity and Validation

As the unfortunately clumsy name indicates, criterion-related validity is studied by comparing test or scale scores with one or more external variables, or criteria, known or believed to measure the attribute under study. One type of criterion-related validity is called predictive validity. The other type is called concurrent validity, which differs from predictive validity in the time dimension. Predictive validity involves the use of future performance of the criterion, whereas concurrent validity measures the criterion at about the same time. In this sense, the test serves to assess the present status of individuals.

Concurrent validity is often used to validate a new test. At least two concurrent measures are taken on each examinee. One of these would be the new test and the other would be some existing test or measure. Concurrent validity would be found by correlating the two sets of scores. In the intelligence testing area, the new tests and even revisions of older tests, generally use the Stanford–Binet or the Wechsler test as the concurrent criterion.

Decision Aspects of Validity

Criterion-related validity, as indicated earlier, is ordinarily associated with practical problems and outcomes. Interest is not so much in what is behind test performance as it is in helping to solve practical problems and to make decisions. Tests are used by the hundreds for the predictive purposes of screening and selecting potentially successful candidates in education, business, and other occupations. Does a test, or a set of tests, materially aid in deciding on the assignment of individuals to jobs, classes, schools, and the like? Any decision is a choice among treatments, assignments, or programs. Cronbach (1971) points out that to make a decision, one predicts the person’s success under each treatment and then use, some rule to translate the prediction into an assignment. A test high in criterion-related validity is one that helps investigators make successful decisions in assigning people to treatments, conceiving treatments broadly. An admissions committee or administrator decides to admit or not admit an applicant to college on the basis of a test of academic aptitude. Obviously, such use of tests is highly important, and the tests’
predictive validity is also highly important. The reader is referred to Cronbach's essay for a good exposition of the decision aspects of tests and validity.

A major contribution in this area is by Taylor and Russell (1939). These researchers demonstrated that tests with low validity can still be effectively used for decision purposes. They developed the Taylor–Russell Table, which utilizes three pieces of information: validity coefficient, selection ratio, and base rate. The selection ratio pertains to the number of people (applicants) who will be selected out of the total number of people. If there were only 10 positions and 100 people applying, the selection ratio would be 0.10 or 10%. The base rate is that proportion of people in the population with a certain characteristic. This figure is generally reported in the press. The base rate for women, for example, is 52 or 52% of the population in the United States. Without using a test, if we gathered randomly 100 persons in a room, 52 of them would be women. Any of these three components can be varied and in so doing has an effect on the accuracy of selection. That is, it can help make a better decision. Anastasi and Urbina (1997) give a good account of how this method works. The interested reader would need to consult the original Taylor and Russell article to see the complete range of tables. Essentially better prediction can be made using a low validity test if the selection ratio is small. Since 1939, there have been a few modification, and additions to this method. These include Abrahams, Alf, and Wolfe (1971); Pritchard and Kazar (1979); and Thomas, Owen, and Gunst (1977).

**Multiple Predictors and Criteria**

Both multiple predictors and multiple criteria can be and are used. Later, when we study multiple regression, we will focus on multiple predictors and how to handle them statistically. Multiple criteria can be handled separately or together, though it is not easy to do the latter. In practical research, a decision must usually be made. If there is more than one criterion, how can we best combine them for decision-making? The relative importance of the criteria, of course, must be considered. Do we want an administrator high in problem-solving ability, high in public-relations ability, or both? Which is more important in the particular job? It is highly likely that the use of both multiple predictors and multiple criteria will become common, as multivariate methods become better understood and the computer is used routinely in prediction research.

**Construct Validity and Construct Validation**

Construct validity is one of the most significant scientific advances of modern measurement theory and practice. It is a significant advance because it links psychometric notions and practices to theoretical notions. The classic work in this area is Cronbach and Meehl (1955). Measurement experts, when they inquire into the construct validity of tests, usually want to know which psychological or other property or properties can "explain" the variance of tests. They wish to know the "meaning" of tests. If a test is an intelligence test, they want to know which factors lie behind test performance. They ask: Which factors or constructs account for variance in test performance? Does this test measure verbal ability and abstract reasoning ability? Does it also "measure" social class membership? They ask, for example, what proportion of the total test variance is accounted for by each of the constructs—verbal ability, abstract reasoning ability, and social class membership. In short, they seek to explain individual differences in test scores. Their interest is usually more in the properties being measured, than in the tests used to accomplish the measurement.

Researchers generally start with the constructs or variables entering into relations. Suppose that a researcher has discovered a positive correlation between two measures, one a measure of educational traditionalism and the other a measure of the perception of the characteristics associated with a "good" teacher. Individuals high on the traditionalism measure see the "good" teacher as efficient, moral, thorough, industrious, conscientious, and reliable. Individuals low on the traditionalism measure may see the "good" teacher in a different way. The researcher now wants to know why this relation exists, what is behind it. To accomplish this, the meaning of the constructs entering the relation, "perception of the 'good' teacher" and "traditionalism" must be studied. How to study these meanings is a construct validity problem. This example was extracted from Kerlinger and Pedhazur (1968).

One can see that construct validation and empirical scientific inquiry are closely allied. It is not simply a question of validating a test. One must try to validate the theory behind the test. Cronbach (1990) says that there are three parts to construct validation: suggesting which constructs possibly account for test performance, deriving hypotheses from the theory involving the construct, and testing the hypotheses empirically. This formulation is but a piece of the general scientific approach discussed in earlier chapters.

The significant point about construct validity that sets it apart from other types of validity, is its preoccupation with theory, theoretical constructs, and scientific empirical inquiry, involving the testing of hypothesized relations. Construct validation in measurement contrasts sharply with approaches that define the validity of a measure, primarily by its success in predicting a criterion. For example, a purely empirical test might say that a test is valid if it efficiently distinguishes individuals high and low in a trait. Why the test succeeds in separating the subsets of a group is of no great concern. It is enough that it does.

**Convergence and Discriminability**

Note that the testing of alternative hypotheses is particularly important in construct validation, because both convergence and discriminability are required. Convergence means that evidence from different sources gathered in different ways all indicate the same or similar meaning of the construct. Different methods of measurement should converge on the construct. The evidence yielded by administering the measuring instrument to different groups in different places should yield similar meanings or, if not, should account for differences. A measure of the self-concept of children, for instance, should be capable of similar interpretation in different parts of the country. If
it is not capable of such interpretation in some locality, the theory should be able to explain why—indeed, it should predict such a difference.

Discriminability means that one can empirically differentiate the construct from other constructs that may be similar, and that one can point out what is unrelated to the construct. We point out, in other words, that other variables are correlated with the construct and how they are so correlated. But we also indicate which variables should be uncorrelated with the construct. We point out, for example, that a scale to measure Conservatism should and does correlate substantially with measures of Authoritarianism and Rigidity—the theory predicts this—but not with measure of Social Desirability (see Kerlinger, 1970). Let us illustrate these ideas.

A Hypothetical Example of Construct Validation
Let us assume that an investigator is interested in the determinants of creativity and the relation of creativity to school achievement. The investigator notes that the most sociable persons, who exhibit affection for others, also seem to be less creative than those who are less sociable and affectionate. The goal is to test the implied relation in a controlled fashion. One of the first tasks is to obtain or construct a measure of the sociable—affectionate characteristic. The investigator, surmising that this combination of traits may be a reflection of a deeper concern of love for others, calls it amorism. An assumption is made about individual differences in amorism: that is, some people have a great deal of it, others a moderate amount, and still others very little.

The first step is to construct an instrument to measure amorism. The literature gives little help, since scientific psychologists have rarely investigated the fundamental nature of love. Sociability, however, has been measured. The investigator must construct a new instrument, basing its content on intuitive and reasoned notions of what amorism is. The reliability of the test, tried out with large groups, runs between .75 and .85.

The question now is whether or not the test is valid. The investigator correlates the instrument, calling it the $A$-scale, with independent measures of sociability. The correlations are moderately substantial, but additional evidence is needed to claim that the test has construct validity. Certain relations are deduced that should and should not exist between amorism and other variables. If amorism is a general tendency to love others, then it should correlate with characteristics like cooperativeness and friendliness. Persons high in amorism will approach problems in an ego-oriented manner as contrasted to persons low in amorism, who will approach problems in a task-oriented manner.

Acting on this reasoning, the investigator administers the $A$-scale and a scale to measure subjectivity to a number of tenth-grade students. To measure cooperativeness, an observation of the classroom behavior of the same group of students is made. The correlations between the three measures are positive and significant. Note that we would not expect high correlation between the measures. If the correlations were too high, we would then suspect the validity of the $A$-scale. It would be measuring, perhaps, subjectivity or cooperativeness, but not amorism.

Knowing the pitfalls of psychological measurement, the investigator is not satisfied. These positive correlations may be due to a factor common to all three tests, but irrelevant to amorism; for example, the tendency to give "right" answers. (This would probably be ruled out, however, because the observation measure of cooperativeness correlates positively with amorism and subjectivity.) So, taking a new group of participants, the investigator administers the amorism and subjectivity scales, has the participants’ behavior rated for cooperativeness, and, in addition, administers a creativity test that has been found in other research to be reliable.

The investigator states the relation between amorism and creativity in hypothesis form: The relation between the $A$-scale and the creativity measure will be negative and significant. The correlations between amorism and cooperativeness and between amorism and subjectivity will be positive and significant. "Check" hypotheses are also formulated: The correlation between cooperativeness and creativity will not be significant, it will be near zero; but the correlation between subjectivity and creativity will be positive and significant. This last relation is predicted on the basis of previous research findings. The six correlation coefficients are given in the correlation matrix of Table 28.1. The four measures are labeled as follows: $A$, amorism; $B$, cooperativeness; $C$, subjectivity; and $D$, creativity.

The evidence for the construct validity of the $A$-scale is good. All the $r$s are as predicted; especially important are the $r$s between $D$ (creativity) and the other variables. Note that there are three different kinds of prediction: positive, negative, and zero. All three kinds are as predicted. This illustrates what might be called differential prediction or differential validity—or discriminability. It is not enough to predict, for instance, that the measure presumably reflecting the target property is positively correlated with one theoretically relevant variable. One should, through deduction from the theory, predict more than one such positive relation. In addition, one should predict zero relations between the principal variable and variables "irrelevant" to the theory. In the example above, although cooperativeness was expected to correlate with amorism, there was no theoretical reason to expect it to correlate at all with creativity.

An example of a different kind is the investigator who deliberately introduces a measure that would, if it correlates with the variable whose validity is under study,

### Table 28.1 Intercorrelations of Four Hypothetical Measures ($N = 90^*$)

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.50</td>
<td>.60</td>
<td>-.30</td>
</tr>
<tr>
<td>B</td>
<td>.40</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>.50</td>
<td></td>
</tr>
</tbody>
</table>

$^*$ A = Amorism; B = Cooperativeness; C = Subjectivity; D = Creativity. Correlation coefficients .25 or greater are significant at the .01 level.
invalidate other positive relations. One bugaboo of personality and attitude scales is the social desirability phenomenon mentioned earlier. The correlation between the target variable and a theoretically related variable may be due to both instruments measuring social desirability, rather than the variables they were designed to measure. One can partly check against this tendency by including a measure of social desirability along with the other measures.

Despite all the evidence leading the investigator to believe that the A-scale has construct validity, there may still be doubt. So a study is developed in which pupils who are high and low in amorism solve problems. The prediction is that pupils low in amorism will solve problems more successfully than those high in amorism. If the data support the prediction, this is further evidence of the construct validity of the amorism measure. It is, of course, a significant finding in and of itself. Such a procedure, however, is probably more appropriate with achievement and attitude measures. One can manipulate communications, for example, in order to change attitudes. If attitude scores change according to theoretical prediction, this would be evidence of the construct validity of the attitude measure, since the scores would probably not change according to prediction if the measure were not measuring the construct.

The Multitrait–Multimethod Matrix Method

A significant and influential contribution to testing validity is Campbell and Fiske’s (1959) use of the ideas of convergence and discriminability and correlation matrices to bring evidence to bear on validity. To explain the method, we use some data from a study of social attitudes by Kerlinger (1967, 1984). It has been found that there are two basic dimensions of social attitudes, which correspond to philosophical, sociological, and political descriptions of liberalism and conservatism. Two different kinds of scales were administered to graduate students of education and groups outside the universities in New York, Texas, and North Carolina. One instrument, Social Attitudes Scale, had the usual attitude statements, 13 liberal and 13 conservative items.

The second instrument, Refers-1, or REF-1, used attitude refers (single word and short phrases: private property, religion, and civil rights, for example) as items, 25 liberal referents and 25 conservative referents. The samples, the scales, and some of the results are described in Kerlinger (1972). The data reported in Table 28.2 were obtained from a Texas sample, N = 227 graduate students.

We have, then, two completely different kinds of attitude instruments, one with referent items and the other with statement items, or Method 1 and Method 2. The two basic dimensions being measured were liberalism (L) and conservatism (C). In the I. and C. subscales of the two scales measure liberalism and conservatism. Part of the evidence is given in Table 28.2, which presents the correlations among the four subscales of the two instruments, as well as the subscale reliability coefficients, calculated from the responses to the two scales.

In a multitrait–multimethod analysis, more than one attitude and more than one method are used in the validation process. The results of correlating variables within and between methods can be presented in a so-called multitrait–multimethod matrix. The matrix (matrices) given in Table 28.2 is the simplest possible form of such an analysis: two variables and two methods. Ordinarily one would want to use more variables.

The most important part of the matrix is the diagonal of the cross-method correlation matrix (Table 28.2), this is the Method 1–Method 2 matrix in the lower-left section of the table. The diagonal values should be substantial, since they reflect the magnitudes of the correlations between the same variables measured differently. These values, italicized in the table (.53 and .54), are fairly substantial.

In this example, the theory calls for near-zero or low negative correlations between L and C (see Kerlinger, 1967 for a more complete development of this). The correlation between L1 and C1 is −.07 and between L1 and C2 is +.09, both in accord with the theory. The cross-correlation between L and C, that is, the correlation between L of Method 1 and C of Method 2, or between L1 and C1 is −.37, higher than the theory predicts (an upper limit of .30 was adopted). With the exception of the cross-correlation of −.37 between L1 and C1, then, the construct validity of the social attitudes scale is supported. One will, of course, want more evidence than the results obtained with one sample. And one will also want an explanation of the substantial cross-method negative correlation between L1 and C1. The example, however, illustrates the basic idea of the multitrait-multimethod approach to validity.

Campbell and Fiske (1959) used specific terminology to describe each correlation in the table. The monotrait–monomethod are the reliabilities. These are found in the main diagonal of the matrix. In Table 28.2, these are the values .85, .88, .81, and .82, enclosed in parentheses. The heteromethod–monotrait are the reliabilities which we discussed above. They are .53 and .54 in Table 28.2. There are two other types of correlation: the monotrait–heteromethod (the values −.07 and +.09), and the heteromethod–heteromethod (these were −.37 and −.15). Campbell and Fiske state that in order to have complete evidence of construct validity, the correlations must follow a set pattern. Failure to meet the requirements weakens validity concerns. There

<table>
<thead>
<tr>
<th>Method 1 (1)</th>
<th>Method 2 (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>C1</td>
</tr>
<tr>
<td>(.85)</td>
<td>−.07</td>
</tr>
<tr>
<td>(.88)</td>
<td>.54</td>
</tr>
</tbody>
</table>

Method 1: Referents; Method 2: Statements; L = Liberalism; C = Conservatism. The diagonal parenthesized entries are internal consistency reliabilities; the italicized entries (.53 and .54) are cross-method L-C correlations (validities).
A Measure of Anti-Semitism

In an unusual attempt to validate their measure of anti-Semitism, Glock and Stark (1966) used responses to two incomplete sentences about Jews: “It’s a shame that Jews . . .” and “I can’t understand why Jews . . .” Coders considered what each subject had written and characterized the responses as negative, neutral, or positive images of Jews. Each subject, then, was characterized individually as having one of the three different perceptions of Jews. When the responses to the Index of Anti-Semitic Beliefs, the measures being validated, were divided into None, Medium, Medium High, and High Anti-Semitism, the percentages of negative responses to the two open-ended questions were, respectively: 28, 41, 61, 75. This is good evidence of validity because the individuals categorized None to High Anti-Semitism by the measure to be validated, the Index of Anti-Semitic Beliefs, responded to an entirely different measure of anti-Semitism, the two open-ended questions, in a manner congruent with their categorization by the index.

A Measure of Personality

In a later chapter we will be discussing an important analytic tool called factor analysis. However, it is necessary to mention this method in light of construct validation. In recent years, factor analysis seems to be the method of choice for many involved with construct validity. Factor analysis is essentially a method of finding those variables that have something in common. If some items of a personality test are designed to measure extraversion, then in a factor analysis, those items should have high loadings on one factor and low on the others.

In the mid-1950s Professor Andrew L. Comrey at University of California, Los Angeles, undertook a task to examine all of the existing well-known, published personality tests. His initial goal was to try to determine who had the correct (valid) measure of personality. To do this, Dr. Comrey used factor analysis. Contrary to his initial expectations, a new personality test of its own unique character emerged. Comrey’s personality test, now called the Comrey Personality Scales (CPS) was among the first to be developed using factor analysis. In 1970, after approximately 15 years of research and test construction, the Comrey Personality Scale was published (see Comrey & Lee, 1992 for a summary and procedure). Comrey’s construct for personality consists of eight major dimensions:

- Trust versus Defensiveness
- Orderliness versus Lack of Compulsion
- Social Conformity versus Rebelliousness
- Activity versus Lack of Energy
- Emotional Stability versus Neuroticism
- Extraversion versus Introversion
- Masculinity versus Femininity (renamed as Mental Toughness versus Sensitivity)
- Empathy versus Egocentrism
Since 1970, Comrey has published a number of articles supporting the validity of the Comrey Personality Scales. This was done by first administering the CPS, or a translated form of the CPS, to different groups of people. After obtaining the data, each set of data was factor analyzed. In each case, the same eight factors emerged. Although this does not say exclusively there are eight factors of personality, the data support it. In recent research by Brief, Comrey, and Collins (1994), the CPS was translated into Russian and administered to 287 male and 170 female Russian participants. The data supported six of the eight subscales. The only subscales that did not receive enough support were Empathy versus Egocentrism and Activity versus Lack of Energy.

In a short article, Comrey, Wong, and Backer (1978) present a simple procedure for validating the Social Conformity versus Rebelliousness scale. In one study, Comrey et al. recruited two groups of participants: Asians and non-Asians. The traditional view of Asians is that they are more socially conforming than non-Asians. There is some evidence to support this claim, such as strong parental influence, strong traditional values, and so on. [Scattone and Saertmo (1997) is one research study that demonstrated this.] Hence in this study, by Comrey and others, the established notion concerning the difference between Asians and non-Asians on social conformity was used as the criterion or “outside measure.” The participants all took both the Comrey Personality Scales; however, only the Social Conformity versus Rebelliousness was of interest for this study. Using a t-test, these researchers showed a statistically significant difference between Asians and non-Asians on the Social Conformity versus Rebelliousness scale. This study could be used as an example illustrating discriminant validity.

The second study in this article demonstrated convergent validity. One expects that Social Conformity is related to political affiliation and philosophy. It is generally thought that Conservatives are more socially conforming than Liberals, who are considered more rebellious. In this study, persons completed the Comrey Personality Scales and answered questions about political affiliation. Comrey et al. found a statistically significant correlation between political affiliation and scores on the Social Conformity versus Rebelliousness scale. This provided additional information as to the validity of that scale. Even though this article is short, it is well presented. The student learn a great deal from reading this article.

The Measurement of Democracy

What do we mean by democracy? The word is used constantly, but what do we mean when we use it? Even more difficult, how is it measured? Bollen (1980) defined and measured “democracy,” used it as a construct, and demonstrated the construct validity of his Index of Political Democracy. He examined previous uses and definitions carefully, explained the theory behind the construct, and extracted from earlier measures important facets of political democracy to construct his measure. It has two large aspects—political liberty, and popular sovereignty—which can be called latent variables. Each aspect has three facets: press freedom, freedom of group opposition, and governmental sanctions (absence of) for political liberties; and freedom of elections, executive selection, and legislative selection for popular sovereignty. It is these six “indica-

tors” that are used to measure the political democracy of countries. Each indicator is defined operationally and a t-point scale used to apply to any country. Popular sovereignty, for instance, is measured by assessing to what extent the elite of a country are accountable to the people: wide franchise, equal weighting of votes, and fair electoral processes. The six indicators are combined into a single index or score (see Bollen, 1979, for a detailed description of the index and its scoring). Note that “Indicator,” or “Social Indicator,” is an important term in contemporary social research. Unfortunately, there is little agreement on just what indicators are. They have been variously defined as indices of social conditions, statistics, and even variables. In Bollen’s paper, they are variables. For a discussion of definitions, see Jaeger (1978).

Through factor analysis and other procedures, Bollen brought empirical evidence to bear on the reliability and construct validity of the index. He showed, for example, that the six indicators are manifestations of an underlying latent variable, which is “political democracy.” He also showed that the index is highly correlated with other measures of democracy. Finally, index values were calculated for a large number of countries. These values seem to agree with the extent of democracy (a scale of 0–100) in the countries, for example, U.S., 92.4; Canada, 99.5; Cuba, 5.2; United Arab Republic, 38.7; Sweden, 99.9; Soviet Union, 18.2; Israel, 96.8. Bollen has evidently successfully measured a highly complex and difficult construct.

Other Methods of Construct Validation

In addition to the multivariate–multimethod approach and the methods used in the above studies, there are other methods of construct validation. Any test is familiar with the technique of correlating items with total scores. In using the technique, the total score is assumed to be valid. To the extent that an item measures the same thing as the total score does, the item is valid (see Chapter 27, or Friedenberg, 1995, for discussion on item analysis).

In order to study the construct validity of any measure, it is always helpful to correlate the measure with other measures. The researcher might illustrate the method and the ideas behind it. But, would it not be more valuable to correlate a measure with a large number of other measures? Is there any better way to learn about a construct than to know its correlates? Factor analysis is a refined method of doing this. It tells us, in effect, what measures measure the same thing and to what extent they measure what they measure.

Factor analysis is a powerful and indispensable method of construct validation. Bollen (1980) used it in his validation of the Index of Political Democracy and Comrey used it to develop an entire personality test. Although it has been briefly characterized earlier and will be discussed in detail in a later chapter, its great importance in validating measures warrants characterization here. It is a method for reducing a large number of measures to a smaller number, called factors, by discovering which ones “go together” (i.e., which measures measure the same thing) and the relations between the clusters of measures that go together. For example, we may give a group of individuals 20 tests, each presumed to measure something different. We may find,
A Variance Definition of Validity: The Variance Relation of Reliability and Validity

The variance treatment of validity presented here is an extension of the treatment of reliability presented in Chapter 27. Both treatments follow Guilford’s presentation of validity.

In the last chapter, reliability was defined as

$$r_n = \frac{V_x}{V_t} \tag{28.1}$$

the proportion of “true” variance to total variance. It is theoretically and empirically useful to define validity similarly:

$$V_{al} = \frac{V_{al}}{V_t} \tag{28.2}$$

where $V_{al}$ is the validity, $V_x$ the common factor variance, and $V_t$ the total variance of a measure. Validity is thus seen as the proportion of the total variance of a measure that is common factor variance.

Unfortunately, we are not yet in a position to present the full meaning of this definition. An understanding of so-called factor theory is required, but factor theory will not be discussed until later in the book. Despite this difficulty, we must attempt an explanation of validity in variance terms if we are to have a well-rounded view of the subject. Besides, expressing validity and reliability mathematically will unify and clarify both subjects. Indeed, reliability and validity will be seen to be parts of one unified whole.

Common factor variance is the variance of a measure that is shared with other measures. In other words, common factor variance is the variance that two or more tests have in common.

In contrast to the common factor variance of a measure is its specific variance, $V_s$, the systematic variance of a measure that is not shared by any other measure. If a test measures skills that other tests measure, we have common factor variance; if it also measures a skill that no other test measures, we have specific variance. Figure 28.1 expresses these ideas and also adds the notion of error variance. The $A$ and $B$ circles represent the variances of Tests $A$ and $B$. The intersection of $A$ and $B$, $A \cap B$, is the relation of the two sets. Similarly, $V(A \cap B)$ is common factor variance. The specific variances and the error variances of both tests are also indicated.

From this viewpoint, then, and following the variance reasoning outlined in the last chapter, any measure's total variance has several components: common factor variance, specific variance, and error variance. This is expressed by the equation:

$$V_t = V_{al} + V_s + V_e \tag{28.3}$$

To be able to talk of proportions of the total variance, we divide the terms of Equation 28.3 by the total variance:

$$\frac{V_t}{V_t} = \frac{V_{al}}{V_t} + \frac{V_s}{V_t} + \frac{V_e}{V_t} \tag{28.4}$$

How do Equations 28.1 and 28.2 fit into this picture? The first term on the right of the equal sign, $V_{al}/V_t$, is the right-hand member of (28.2). Therefore validity can be viewed as that part of the total variance of a measure that is not specific variance and not error variance. This is easily seen algebraically:

$$\frac{V_{al}}{V_t} = \frac{V_t}{V_t} - \frac{V_s}{V_t} - \frac{V_e}{V_t} \tag{28.5}$$

By a definition of the previous chapter, reliability can be defined as

$$r_n = 1 - \frac{V_e}{V_t} \tag{28.6}$$

This can be written:

$$r_n = \frac{V_t}{V_t} - \frac{V_e}{V_t} \tag{28.7}$$

The right-hand side of the equations, however, is part of the right-hand side of (28.5). If we rewrite (28.5) slightly, we obtain
\[
\frac{V_\alpha}{V_t} = \frac{V_i}{V_t} - \frac{V_e}{V_t} = \frac{V_s}{V_t} \quad (28.8)
\]

This must mean, then, that validity and reliability are close variance relations. Reliability is equal to the first two right-hand members of (28.8). So, bringing in (28.1):

\[
r_n = \frac{V_i}{V_t} - \frac{V_e}{V_t} = \frac{V_s}{V_t} \quad (28.9)
\]

If we substitute in (28.8), we get

\[
\frac{V_\alpha}{V_t} = \frac{V_s}{V_t} - \frac{V_e}{V_t} \quad (28.10)
\]

Thus we see that the proportion of the total variance of a measure is equal to the proportion of the total variance that is "true" variance minus the proportion that is specific variance. Or, the validity of a measure is that portion of the total variance of the measure that shares variance with other measures. Theoretically, valid variance includes no variance due to error, neither does it include variance that is specific to this measure and this measure only.

This can all be summed up in two ways. First, we sum it up in an equation or two. Let us assume that we have a method of determining the common factor variance (or variances) of a test. (Later we shall see that factor analysis is such a method.) For simplicity, suppose that there are two sources of common factor variance in a test—and no others. Call these factors A and B. They might be verbal ability and arithmetic ability, or they might be liberal attitudes and conservative attitudes. If we add the variance of A to the variance of B, we obtain the common factor variance of the test, which is expressed by the equations,

\[
\frac{V_\alpha}{V_t} = \frac{V_A}{V_t} = \frac{V_B}{V_t} \quad (28.11)
\]

Then, using (28.2) and substituting in (28.12), we obtain

\[
\frac{V_{\alpha d}}{V_t} = \frac{V_A}{V_t} + \frac{V_B}{V_t} \quad (28.13)
\]

The total variance of a test, as we said before, includes the common factor variance, the variance specific to the test and to no other test (at least as far as present information goes), and error variance. Equations 28.3 and 28.4 express this. By substituting in (28.4) the equality of (28.12), we obtain

\[
\frac{V_i}{V_t} = \frac{V_A}{V_t} + \frac{V_B}{V_t} + \frac{V_e}{V_t} \quad (28.14)
\]

The first two terms on the right-hand side of (28.14) are associated with the validity of the measure, and the first three terms on the right are associated with the reliability of the measure. These relations have been indicated. Common factor variance, or the validity component of the measure, is labeled \(k^2\) (commonality), a symbol customarily used to indicate the common factor variance of a test. Reliability, as usual, is labeled \(r_n\).

To discuss all the implications of this formulation of validity and reliability would take us too far afield at this time. All that is needed now is to try to clarify the formulation with a diagram and a brief discussion.

Figure 28.2 is in an attempt to express Equation 28.14 diagrammatically. The figure represents the contributions of the different variances to the total variance (taken to be equal to 100%). Four variances, three systematic variances and one error variance, comprise the total variance in this theoretical model. Naturally, practical outcomes never look this neat. It is remarkable, however, how well the model works. The variance thinking, too, is valuable in conceptualizing and discussing measurement outcomes.

The contribution of each source of variance is indicated. Of the total variance, 80% is reliability variance. Of the reliability variance, Factor A contributes 30% and Factor B contributes 25%, and 25% is specific to this test. The remaining 20% of the total variance is error variance. The test may be interpreted as quite reliable, since a
sizable proportion of the total variance is reliable or "true" variance. The interpretation of validity is more difficult. If there were only one factor, say A, and it contributed 55% of the total variance, then we could say that a considerable proportion of the total variance was valid variance. We would know that a good bit of the reliable measurement would be the measurement of the property known as A. This would be a construct validity statement. Practically speaking, individuals measured with the test would be rank-ordered on A with adequate reliability.

With the above hypothetical example, however, the situation is more complex. The test measures two factors, A and B. There could be three sets of rank orders, one resulting from A, one from B, and one from specific. While repeat reliability might be high, if we thought we were measuring only A, to the extent we thought so, the test would not be valid. We might, however, have a score for each individual one on A and one on B. In this case the test would be valid. Note that even if we thought the test was measuring only A, predictions to a criterion might well be successful, especially if the criterion had a lot of both A and B in it. The test could have predictive validity even though its construct validity was questionable.

Indeed, modern developments in measurement indicate that such multiple scores have become more and more a part of accepted procedure.

**Statistical Relation between Reliability and Validity**

Although they appear in different chapters, the topics of reliability and validity are not separate—both deal with the level of excellence of a measuring instrument. We have seen in past discussions that we can have a reliable measure that is not valid. However, a measuring instrument without reliability would automatically designate it to the "poor" stack. Also, we had briefly mentioned that if we have a valid measure then we also have a reliable one. In Chapter 27 we discussed what happens to the reliability coefficient when we increase the length of the test. What happens to validity with an increase in length? Is it equally affected by the increase in length as reliability? The answer is "no." Gullikson (1950) classical work present formulas to show the relationship. If enough items are added to the test to double the reliability coefficient, the validity coefficient only increases by 41%. The prophetic formulas for validity usually involve the reliability coefficient in some shape and form. For example, there is a formula to predict the validity coefficient based on the reliability coefficient. Using this formula it may be possible to obtain a validity coefficient higher than the reliability. However, in practice it is very difficult to obtain a validity coefficient that is larger than the reliability. The thinking here is that one would expect that a test correlated with itself should be higher than the same test correlated with an outside measure or criterion.

If it was possible to eliminate measurement errors of the test and the criterion, we would essentially have a correlation between the true scores of both measures. We have seen that measurement errors tend to lower the coefficient values. We can, in a hypothetical realm, find what the validity coefficient might be if measurement error could be eliminated in (i) both criterion and test, (ii) criterion only, and (iii) test only. Such corrections are referred to as corrections for attenuation. If we let $r_{xy}$ be the correlation between criterion $x$ and test, $y$, the formula to correct for attenuation in both is

$$Corrected\ r_{xy} = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}}$$

The formula to determine what the validity might be if we had a perfect criterion is

$$r_{xy} = \frac{r_{xy}}{\sqrt{r_{xx}}}$$

The formula to determine the validity coefficient if we had a perfect test is

$$r_{xy} = \frac{r_{xy}}{\sqrt{r_{yy}}}$$

These formulas should not be used to make decisions about individuals, but are useful in determining if making a test or criterion more reliable is worth the effort. These formulas show what would happen to validity as changes are made in reliability.

**The Validity and Reliability of Psychological and Educational Measurement Instruments**

Poor measurement can invalidate any scientific investigation. Most of the criticisms of psychological and educational measurement, by professionals and laypeople alike, center on validity. This is as it should be. Achieving reliability is to a large extent a technical matter. Validity, however, is much more than technique. It borrows from the essence of science itself. It also borrows into philosophy. Construct validity, particularly, since it is concerned with the nature of "reality" and the nature of the properties being measured, is heavily philosophical.

Despite the difficulties of achieving reliable and valid psychological, sociological, and educational measurements, great progress has been made in this century. There is growing understanding that all measuring instruments must be critically and empirically examined for their reliability and validity. The day of tolerance of inadequate measurement has ended. The demands imposed by professionals, the theoretical and statistical tools available and those being rapidly developed, and the increasing sophistication of graduate students of psychology, sociology, and education have set new high standards that should be healthy stimuli to the imaginations of both research workers and developers of scientific measurement.

**Chapter Summary**

1. Validity deals with accuracy. Does the instrument measure what it is supposed to measure?
2. There are three types of validity
   • content
   • criterion-related
   • construct
3. Content validity is concerned with the representativeness or sampling adequacy of the test's content.
4. Face validity is similar to content validity, but it is nonquantitative and involves merely a visual inspection of the test by sophisticated or unsophisticated reviewers.
5. Under criterion-related validity there are two methods: concurrent and predictive.
6. The distinguishing characteristic between concurrent and predictive validities is the temporal relationship between the instrument and the criterion.
7. An instrument high in criterion-related validity helps test users make better decisions in terms of placement, classification, selection, and assessment.
8. Construct validity seeks to explain individual differences in test scores. It deals with abstract concepts that may contain two or more dimensions.
9. Construct validity requires both convergence and discriminability.
10. Convergence states that instruments purporting to measure the same thing should be highly correlated.
11. Discriminability is shown when instruments that supposedly measure different things have a low correlation.
12. A method used to show both convergence and discriminability is Campbell and Fiske's (1959) multitrait–multimethod matrix.
13. We can show the relationship between validity and reliability mathematically.
14. Knowledge on how measurements are interpreted is important to research studies.
15. Two less traditional topics concerning interpretation and validity are: criterion-referenced testing and information referenced testing (or admissible probability measurement).

**Study Suggestions**

1. The measurement literature is vast. The following references have been chosen for their particular excellence or their relevance to important measurement topics. Some of the discussions, however, are technical and difficult. The student will find elementary discussions of reliability and validity in most measurement texts.


   Tryon, R. (1957). Reliability and behavior domain validity: A reformulation and historical critique. *Psychological Bulletin*, 54, 229–249. [This is an excellent and important article on reliability. It contains a good worked example.]

   The following anthologies of measurement articles are valuable sources of the classics in the field. This is especially true of the Mehrens and Ebel and the Jackson and Messick volumes.


2. An important method in validity studies is cross-validation. Advanced students can profit from Mosier's essay in the Chase and Ludlow book listed above. A brief summary of Mosier's essay can be found in Guilford (1954, p. 406).
3. The more advanced student will also want to know something about response sets—a threat to validity, particularly to the validity of personality, attitude, and value items and instruments. Response sets are tendencies to respond to items in certain ways—high, low, approve, disapprove, extreme, and so on—regardless of the content of the items. The resulting scores are therefore systematically biased. The literature is extensive and cannot be cited here. An excellent exposition, however, can be found in Numally, (1978), chap. 16, especially pp. 653ff. Advocates of the effects of response sets on measurement instruments are quite strong in their statements. Rorer (1965), has thrown a considerable dash of salt on the response-set tail.

   The position taken in this book is that response sets certainly operate and sometimes have considerable effect, but that the strong claims of advocates are exaggerated. Most of the variance in well-constructed measures seems to be due to variables being measured, and relatively little to response sets. Investigators must be aware of response sets and their possible deleterious effects on measurement instruments, but should not be afraid to use the
instruments. If one were to take too seriously the schools of thought on response sets and on what has been called the experimenter effect (in education, the Pygmalion effect) discussed earlier, one would have to abandon behavioral research except, perhaps, research that can be done with so-called unobtrusive measures.

4. Imagine that you have given a test of six items to six persons. The scores of each person on each item are given below. Say that you have also given another test of six items to six persons. These scores are also given below. The scores of the first test, I, are given on the left; the scores of the second test, II, are given on the right.

<table>
<thead>
<tr>
<th>Persons</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>Persons</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>7</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

The scores in II are the same as those in I, except that the orders of the scores of items (b), (c), (d), and (f) have been changed.

a. Do a two-way analysis of variance of each set of scores. Compare and interpret the F-ratios. Pay special attention to the F-ratio for Persons (Individuals).

b. Compute \( r_{rt} = (V_{od} - V_{od}) / V_{od} \) for I and II. Interpret the two \( r_{rt} \)s. Why are they so different?

c. Add the odd items across the rows; add the even items. Compare the rank orders and the ranges of the odd totals, the even totals, and the totals of all six items. The coefficients of correlation between odd and even items, corrected, are .98 and .30. Explain why they are so different. What do they mean?

d. Assume that there were 100 persons and 60 items. Would this have changed the procedures and the reasoning behind them? Would the effect of changing the orders of, say, five to 10 items have affected the \( r_{rt} \) as much as in these examples? If not, why not?

[Answers: (a) I: \( F_{item} = 3.79 (.05) \); \( F_{person} = 20.44 (.001) \). II: \( F_{item} = 1.03 \) (n.s); \( F_{person} = 1.91 \) (n.s). (b) I: \( r_{rt} = .95 \); II: \( r_{rt} = .48 \).]