Essentials of Behavioral Science Research:

A First Course in Research Methodology

Howard Lee
California State University

Distributed by
WWW.LULU.COM
Morrisville, NC 27560

Essentials of Behavioral Science Research:
A First Course in Research Methodology

Copyright © 2007 by Howard Lee.
All rights reserved.

Published and Distributed by
WWW.LULU.COM
3131 RDU Center Drive, Suite 210
Morrisville, NC 27560

Printed in the United States of America

Table of Contents

Chapter 1 1
Science and the Scientific Approach 1
Science and Common Sense 1
Four Methods of Knowing 3
Science and Its Functions 4
The Aims of Science, Scientific Explanation, and Theory 5
Scientific Research a Definition 7
The Scientific Approach 8
Problem-Obstacle-Idea 8
Hypothesis 8
Reasoning-Deduction 8
Observation-Test-Experiment 10
Chapter Outline-Summary 12

Chapter 2 13
Problems and Hypotheses 13
Criteria of Problems and Problem Statements 14
Hypotheses 14
The Importance of Problems and Hypotheses 15
Virtues of Problems and Hypotheses 15
Problems, Values, and Definitions 17
Generality and Specificity of Problems and Hypotheses 17
Concluding Remarks-the Special Power of Hypotheses 19
Chapter Summary 21

Chapter 3 23
Constructs, Variables and Definitions 23
Concepts and Constructs 23
Constitutive and Operational Definitions of Constructs and Variables 24
Types of Variables 28
Independent and Dependent Variables 28
Active and Attribute Variables 30
Continuous and Categorical Variables 32
Constructs, Observables, and Latent Variables 33
Examples of Variables and Operational Definitions 34

Chapter 4 39

Calculation of Means and Variances 39
Kinds of Variance 40
Population and Sample Variances 41
Between-Groups (Experimental) Variance 41
Error Variance 43
Components of Variance 48
Covariance 49
Study Suggestions 50
Chapter Summary 51

Chapter 5 53

Sampling and Randomness 53
Sampling, Random Sampling, and Representativeness 53
randomness 55
An Example of Random Sampling 55
Randomization 56
Sample Size 60
Kinds of Samples 62
Some books on sampling. 67
Chapter Summary 68

Chapter 6 69

Ethical Considerations in Conducting Behavioral Science Research 69
A Beginning? 70
General Considerations 72
Deception 73
Freedom from Coercion 73
Debriefing 73
Protection of Participants 73
Confidentiality 73
Ethics of Animal Research 74
Study Suggestions 74
tChapter Summary 75

Chapter 7 77

Research Design: Purpose and Principles 77
Purposes of Research Design 77

A Stronger Design 78
Research Design As Variance Control 80
A Controversial Example 81
Maximization of Experimental Variance 82
Control of Extraneous Variables 83
Minimization of Error Variance 85
Chapter Summary 86

Chapter 8 87
Inadequate Designs and Design Criteria 87
Experimental and Nonexperimental Approaches 87
Symbolism and Definitions 88
Faulty Designs 89
Measurement, History, Maturation 90
The Regression Effect 90
Criteria of Research Design 91
Answer Research Questions? 91
Control of Extraneous Independent Variables 92
Generalizability 93
Internal and External Validity 93

Chapter 9 97

General Designs of Research 97
Conceptual Foundations of Research Design 97
A Preliminary Note: Experimental Designs and Analysis Of Variance 99
The Designs 99
The Notion of the Control Group and Extensions of Design 9.1 100
Matching versus Randomization 102
Some ways of matching groups. 102
Matching by Equating Participants 102
The Frequency Distribution Matching Method 103
Matching by Holding Variables Constant 104
Matching by incorporating the Nuisance Variable into the Research Design 104
Participant as Own Control 104
Additional Design Extensions: Design 9.3 using a Pretest 105
Difference Scores 106
Concluding Remarks 108
Study Suggestions 108
Chapter Summary 109

Chapter 10 111

Research Design Applications: 111
simple Randomized Subjects Design 111
Factorial Designs 113
Factorial Designs with More than Two Variables 114
Research Examples of Factorial Designs 114
Flowers: Groupthink 114
Correlated Groups 120
the General Paradigm 120
One group Repeated Trials Design 121
Two group, Experimental Group-Control Group Designs 122
Research Examples of Correlated-groups Designs 122
Multigroup Correlated-groups Designs 125
Units Variance 125
Factorial Correlated Groups 125
Analysis of Covariance 127
Chapter Summary 133

Chapter 11 135

Quasi Experimental and N = 1 Designs of Research 135
Variants of Basic Designs 135
Compromise Designs a.k.a. Quasi Experimental Designs 135
Nonequivalent Control Group Design 135
No-treatment Control Group Design 136
Time Designs 140

Chapter 1

Science and the Scientific Approach

TO UNDERSTAND any complex human activity one must grasp the language and approach of the individuals who pursue it. So it is with understanding science and scientific research. One must know and understand, at least in part, scientific language and the scientific approach to problem-solving.

One of the most confusing things to the student of science is the special way scientists use ordinary words. To make matters worse, they invent new words. There are good reasons for this specialized use of language; they will become evident later. For now, suffice it to say that we must understand and learn the language of social scientists. When investigators tell us about their independent and dependent variables, we must know what they mean. When they tell us that they have randomized their experimental procedures, we must not only know what they mean-we must understand why they do as they do.

Similarly, the scientist's approach to problems must be clearly understood. It is not so much that this approach is different from the layperson. It *is* different, of course, but it is not strange and esoteric. This is quite the contrary. When understood, it will seem natural and almost inevitable what the scientist does. Indeed, we will probably wonder why much more human thinking and problem-solving are not consciously structured along such lines.

The purpose of Chapters 1 and 2 of this book is to help the student learn and understand the language and approach of science and research. In the chapters of this part many of the basic constructs of the social, behavioral and educational scientist will be studied. In some cases it will not be possible to give complete and satisfactory definitions. This is due to the lack of background at this early point in our development. In such cases an attempt will be made to formulate and use reasonably accurate first approximations to later, more satisfactory definitions. Let us begin our study by considering how the scientist approaches problems and how this approach differs from what might be called a common sense approach.

SCIENCE AND COMMON SENSE

Whitehead (1911/1992, p. 157) at the beginning of the 20th Century pointed out that in creative thought common sense is a bad master. “Its sole criterion for judgment is that the new ideas shall look like the old ones.” This is well said. Common sense may often be a bad master for the evaluation of knowledge. But how are science and common sense alike and how are they different? From one viewpoint, science and common sense are alike. This view would say that science is a systematic and controlled extension of common sense. James Bryant Conant (1951) states that common sense is a series of concepts and conceptual schemes[1] satisfactory for the practical uses of humanity. However, these concepts and conceptual schemes may be seriously misleading in modern science-and particularly in psychology and education. To many educators in the 1800’s, it was common sense to use punishment as a basic tool of pedagogy. However, in the mid 1900’s evidence emerged to show that this older commonsense view of motivation may be quite erroneous. Reward seems more effective than punishment in aiding learning. However, recent findings suggest that different forms of punishment are useful in classroom learning (Tingstrom, et al., 1997; Marlow, et al., 1997).

Science and common sense differ sharply in five ways. These disagreements revolve around the words "systematic" and “controlled.” First, the uses of conceptual schemes and theoretical structures are strikingly different. The common person may use "theories" and concepts, but usually does so in a loose fashion. This person often blandly accepts fanciful explanations of natural and human phenomena. An

---

1. *A concept* is a word that expresses an abstraction formed by generalization from particulars. "Aggression" is a concept, an abstraction that expresses a number of particular actions having the similar characteristic of hurting people or objects. *A conceptual scheme* is a set of concepts interrelated by hypothetical and theoretical propositions. A construct is a concept with the additional meaning of having been created or appropriated for special scientific purposes. "Mass," "energy," "hostility," "introversion," and "achievement" are constructs. They might more accurately be called "constructed types" or "constructed classes," classes or sets of objects or events bound together by the possession of common characteristics defined by the scientist. The term "variable" will be defined in a later chapter. For now let it mean a symbol or name of a characteristic that takes on different numerical values.

illness, for instance, may be thought to be a punishment for sinfulness (Klonoff & Landrine, 1994). Jolliness is due to being overweight. Scientists, on the other hand, systematically build theoretical structures, test them for internal consistency, and put aspects of them to empirical test. Furthermore, they realize that the concepts they use are man-made terms that may or may not exhibit a close relation to reality.

Second, scientists systematically and empirically test their theories and hypotheses. Nonscientists test "hypotheses," too, but they test them in a selective fashion. They often "select" evidence simply because it is consistent with the hypotheses. Take the stereotype: Asians are science and math oriented. If people believe this, they can easily "verify" the belief by noting that many Asians are engineers and scientists (see Tang, 1993). Exceptions to the stereotype, the non-science Asian or the mathematically challenged Asian, for example, are not perceived. Sophisticated social and behavioral scientists knowing this "selection tendency" to be a common psychological phenomenon, carefully guard their research against their own preconceptions and predilections and against selective support of hypotheses. For one thing, they are not content with armchair or fiat exploration of relations; they must test the relations in the laboratory or in the field. They are not content, for example, with the presumed relations between methods of teaching and achievement, between intelligence and creativity, between values and administrative decisions. They insist upon systematic, controlled, and empirical testing of these relations.

A third difference lies in the notion of control. In scientific research, control means several things. For the present, let it mean that the scientist tries systematically to rule out variables that are possible "causes" of the effects under study other than the variables hypothesized to be the "causes." Laypeople seldom bother to control systematically their explanations of observed phenomena. They ordinarily make little effort to control extraneous sources of influence. They tend to accept those explanations that are in accord with their preconceptions and biases. If they believe that slum conditions produce delinquency, they tend to disregard delinquency in nonslum neighborhoods. The scientist, on the other hand, seeks out and "controls" delinquency incidence in different kinds of neighborhoods. The difference, of course, is profound.

Another difference between science and common sense is perhaps not so sharp. It was said earlier that the scientist is constantly preoccupied with relations among phenomena. The layperson also does this by using common sense for explanations of phenomena. But the scientist consciously and systematically pursues relations. The layperson's preoccupation with relations is loose, unsystematic, and uncontrolled. The layperson often seizes, for example, on the fortuitous occurrence of two phenomena and immediately links them indissolubly as cause and effect.

Take the relation tested in the classic study done many years ago by Hurlock (1925). In more recent terminology, this relation may be expressed: Positive reinforcement (reward) produces greater increments of learning than does punishment. The relation is between reinforcement (or reward and punishment) and learning. Educators and parents of the nineteenth century often assumed that punishment was the more effective agent in learning. Educators and parents of the present often assume that positive reinforcement (reward) is more effective. Both may say that their viewpoints are "only common sense." It is obvious, they may say, that if you reward (or punish) a child, he or she will learn better. The scientist, on the other hand, while personally espousing one or the other or neither of these viewpoints, would probably insist on systematic and controlled testing of both (and other) relations, as Hurlock did. Using the scientific method Hurlock found incentive to be substantially related to arithmetic achievement. The group receiving praise scored higher than the reproof or ignored groups.

A final difference between common sense and science lies in different explanations of observed phenomena. The scientist, when attempting to explain the relations among observed phenomena, carefully rules out what have been called "metaphysical explanations." A metaphysical explanation is simply a proposition that cannot be tested. To say, for example, that people are poor and starving because God wills it, or that it is wrong to be authoritarian, is to talk metaphysically.

None of these propositions can be tested; thus they are metaphysical. As such, science is not concerned with them. This does not mean that scientists would necessarily spurn such statements, say they are not true, or claim they are meaningless. It simply means that *as scientists* they are not concerned with them. In short, science is concerned with things that can be publicly observed and tested. If propositions or questions do not contain implications for such public observation and testing, they are not scientific propositions or questions.

FOUR METHODS OF KNOWING

Charles Sanders Peirce as reported in Buchler (1955) said that there are four general ways of knowing or, as he put it, fixing belief. In the ensuing discussion, the authors are taking some liberties with Peirce's original formulation in an attempt to clarify the ideas and to make them more germane to the present discussion. The first is the *method of tenacity.* Here people hold firmly to the truth, the truth that they know to be true because they hold firmly to it, because they have always known it to be true. Frequent repetition of such "truths" seems to enhance their validity. People often cling to their beliefs in the face of clearly conflicting facts. And they will also infer "new" knowledge from propositions that may be false.

A second method of knowing or fixing belief is the *method of authority.* This is the method of established belief. If the Bible says it, it is so. If a noted physicist says there is a God, it is so. If an idea has the weight of tradition and public sanction behind it, it is so. As Peirce points out, this method is superior to the method of tenacity, because human progress, although slow, can be achieved using this method. Actually, life could not go on without the method of authority. Dawes (1994) states that as individuals, we cannot know everything. We accept the authority of the U.S. Drug and Food Administration in determining what we eat and drink are safe. Dawes states that the completely open mind that questions all authority does not exist. We must take a large body of facts and information on the basis of authority. Thus, it should not be concluded that the method of authority is unsound; it is unsound only under certain circumstances.

The *a priori method is* the third way of knowing or fixing belief. Graziano and Raulin (1993) call it the *method of intuition.* It rests its case for superiority on the assumption that the propositions accepted by the "a priorist" are self-evident. Note that a priori propositions "agree with reason" and not necessarily with experience. The idea seems to be that people, through free communication and intercourse, can reach the truth because their natural inclinations tend toward truth. The difficulty with this position lies in the expression "agree with reason." Whose reason? Suppose two honest and well-meaning individuals, using rational processes, reach different conclusions, as they often do. Which one is right? Is it a matter of taste, as Peirce puts it? If something is self-evident to many people-for instance, that learning hard subjects trains the mind and builds moral character, that American education is inferior to Asian and European education-does this mean it is so? According to the a priori method, it does-it just "stands to reason."

The fourth method is the *method of science.* Peirce says:

"To satisfy our doubts, . . . therefore, it is necessary that a method should be found by which our beliefs may be determined by nothing human, but by some external permanency-by something upon which our thinking has no effect. . . . The method must be such that the ultimate conclusion of every man shall be the same. Such is the method of science. Its fundamental hypothesis. . . is this: "There are real things, whose characters are entirely independent of our opinions about them. . . ." (Buchler 1955, p.18)

The scientific approach has a characteristic that no other method of attaining knowledge has: self correction. There are built-in checks all along the way to scientific knowledge. These checks are so conceived and used that they control and verify scientific activities and conclusions to the end of attaining dependable knowledge. Even if a hypothesis seems to be supported in an experiment, the scientist will test alternative plausible hypotheses that, if also supported, may cast doubt on the first hypothesis. Scientists do not accept statements as true, even though the evidence at first looks promising. They insist upon testing them. They also insist that any testing procedure be open to *public* inspection. One interpretation of the scientific method and scientific method is that there is no one scientific method as such. Rather, there are a number of methods that scientists can and do use, but it can probably be said that there is one scientific approach.

As Peirce says, the checks used in scientific research are anchored as much as possible in reality lying outside the scientist's personal beliefs, perceptions, biases, values, attitudes, and emotions. Perhaps the best single word to express this is "objectivity." Objec*tivity is* agreement among "expert" judges on what is observed or what is to be done or has been done in research (see Kerlinger, 1979 for a discussion of objectivity, its meaning and its controversial character.). According to Sampson (1991, p.12) objectivity " refers to those statements about the world that we currently can justify and defend using the standards of argument and proof employed within the community to which we belong - for example, the community of scientists." But, as we shall see later, the scientific approach involves more than both of these statements. The point is that more dependable knowledge is attained because science ultimately appeals to evidence: propositions are subjected to empirical test. An objection may be

raised: Theory, which scientists use and exalt, comes from people, the scientists themselves. But, as Polanyi (1958/1974, p. 4) points out, "A theory is something other than myself." Thus a theory helps the scientist to attain greater objectivity. In short, scientists systematically and consciously use the self-corrective aspect of the scientific approach.

SCIENCE AND ITS FUNCTIONS

What is science? The question is not easy to answer. Indeed, no definition of science will be directly attempted. We shall, instead, talk about notions and views of science and then try to explain the functions of science.

Science is a badly misunderstood word. There seem to be three popular stereotypes that impede understanding of scientific activity. One is the white coat-stethoscope-laboratory stereotype. Scientists are perceived as individuals who work with facts in laboratories. They use complicated equipment, do innumerable experiments, and pile up facts for the ultimate purpose of improving the lot of humanity. Thus, while somewhat unimaginative grubbers after facts, they are redeemed by noble motives. You can believe them when, for example, they tell you that such-and-such toothpaste is good for you or that you should not smoke cigarettes.

The second stereotype of scientists is that they are brilliant individuals who think, spin complex theories, and spend their time in ivory towers aloof from the world and its problems. They are impractical theorists, even though their thinking and theory occasionally lead to results of practical significance like atomic energy.

The third stereotype equates science with engineering and technology. The building of bridges, the improvement of automobiles and missiles, the automation of industry, the invention of teaching machines, and the likes are thought to be science. The scientist's job, in this conception, is to work at the improvement of inventions and artifacts. The scientist is conceived to be a sort of highly skilled engineer working to make life smooth and efficient.

These notions impede student understanding of science, the activities and thinking of the scientist, and scientific research in general. In short, they make the student's task harder than it would otherwise be. Thus they should be cleared away to make room for more adequate notions.

There are two broad views of science: the static and the dynamic. According to Conant (1951, pp. 23-27) the *static view,* the view that seems to influence most laypeople and students, is that science is an activity that contributes systematized information to the world. The scientist's job is to discover new facts and to add them to the already existing body of information. Science is even conceived to be a body of facts. In this view, science is also a way of explaining observed phenomena. The emphasis, then, is on the *present state of knowledge* and *adding to it* and on the present set of laws, theories, hypotheses, and principles.

The *dynamic view,* on the other hand, regards science more as an activity, what scientists do. The present state of knowledge is important, of course. But it is important mainly because it is a base for further scientific theory and research. This has been called a *heuristic view.* The word "heuristic," meaning serving to discover or reveal, now has the notion of self-discovery connected with it. A heuristic method of teaching, for instance, emphasizes students' discovering things for themselves. The heuristic view in science emphasizes theory and interconnected conceptual schemata that are fruitful for further research. A heuristic emphasis is a discovery emphasis.

It is the heuristic aspect of science that distinguishes it in good part from engineering and technology. On the basis of a heuristic hunch, the scientist takes a risky leap. As Polanyi (1958/1974, p. 123) says, "It is the plunge by which we gain a foothold at another shore of reality. On such plunges the scientist has to stake bit by bit his entire professional life." Michel (1991, p. 23) adds "anyone who fears being mistaken and for this reason study a "safe" or "certain" scientific method, should never enter upon any scientific enquiry." Heuristic may also be called problem-solving, but the emphasis is on imaginative and not routine problem-solving. The heuristic view in science stresses problem-solving rather than facts and bodies of information. Alleged established facts and bodies of information are important to the heuristic scientist because they help lead to further theory, further discovery, and further investigation.

Still avoiding a direct definition of science-but certainly implying one-we now look at the function of science. Here we find two distinct views. The practical person, the nonscientist generally, thinks of science as a discipline or activity aimed at improving things, at making progress. Some scientists, too, take this position. The function of science, in this view is to make discoveries, to learn facts, to advance knowledge in order to improve things. Branches of

science that are clearly of this character receive wide and strong support. Witness the continuing generous support of medical and meteorological research. The criteria of practicality and "payoff" are preeminent in this view, especially in educational research (see Kerlinger, 1977; Bruno, 1972).

A very different view of the function of science is well expressed by Braithwaite (1953/1996, p. 1):
"The function of science. . . is to establish general laws covering the behaviors of the empirical events or objects with which the science in question is concerned, and thereby to enable us to connect together our knowledge of the separately known events, and to make reliable predictions of events as yet unknown. " The connection between this view of the function of science and the dynamic-heuristic view discussed earlier is obvious, except that an important element is added: the establishment of general laws-or theory, if you will. If we are to understand modern behavioral research and its strengths and weaknesses, we must explore the elements of Braithwaite's statement. We do so by considering the aims of science, scientific explanation, and the role and importance of theory.

Sampson (1991) discusses two opposing views of science. There is the conventional or traditional perspective and then there is the sociohistorical perspective. The conventional view perceives science as a mirror of nature or a windowpane of clear glass that presents nature without bias or distortion. The goal here is to describe with the highest degree of accuracy what the world really looks like. Here Sampson states that science is an objective referee. Its job is to "resolve disagreements and distinguish what is true and correct from what is not." When the conventional view of science is unable to resolve the dispute, it only means that there is insufficient data or information to do so. Conventionalists, however, feel it is only a matter of time before the truth is apparent.

The sociohistorical view sees science as a story. The scientists are storytellers. Here the idea is that reality can only be discovered by the stories that can be told about it. Here, this approach is unlike the conventional view in that there is no neutral arbitrator. Every story will be flavored by the storyteller's orientation. As a result there is no single true story. Sampson's table comparing these two are reproduced in Table 1.

Even though Sampson gives these two views of science in light of social psychology, his presentation has applicability in all areas of the behavioral science.

THE AIMS OF SCIENCE, SCIENTIFIC EXPLANATION, AND THEORY

The basic aim of science is theory. Perhaps less cryptically, the basic aim of science is to explain natural phenomena. Such explanations are called theories. Instead of trying to explain each and every separate behavior of children, the scientific psychologist seeks general explanations that encompass and link together many different behaviors. Rather than try to explain children's methods of solving arithmetic problems, for example, the scientist seeks general explanations of all kinds of problem-solving. It might be called a general theory of problem-solving.

This discussion of the basic aim of science as theory may seem strange to the student, who has probably been inculcated with the notion that human activities have to pay off in practical ways.
If we said that the aim of science is the betterment of humanity, most readers would quickly read the words and accept them. But the *basic* aim of science is not the betterment of humanity. It is theory. Unfortunately, this sweeping and really complex statement is not easy to understand. Still, we must try because it is important. More on this point is given in chapter 16 of Kerlinger (1979).

Other aims of science that have been stated are: explanation, understanding, prediction, and control. If we accept theory as the ultimate aim of science, however, explanation and understanding become subaims of the ultimate aim. This is because of the definition and nature of theory: *A theory is a set of interrelated constructs (concepts), definitions, and propositions that present a systematic view of phenomena by specifying relations among variables, with the purpose of explaining and predicting the phenomena.*This definition says three things. One, a theory is a set of propositions consisting of defined and interrelated constructs. Two, a theory sets out the interrelations among a set of variables (constructs), and in so doing, presents a systematic view of the phenomena described by the variables. Finally, a theory explains phenomena. It does so by specifying what variables are related to what variables and how they are related, thus enabling the researcher to predict from certain variables to certain other variables. One might, for example, have a theory of school failure. One's variables might be intelligence, verbal and numerical aptitudes, anxiety, social class membership, nutrition, and achievement motivation.

Table 1. Sampson's Two Views of the Science of Social Psychology.

	Conventional	Sociohistorical
Primary Goal	To describe the world of human social experience and activity as it really is and as it really functions	To describe the various accounts of human social experience and activity; to understand both their social and historical bases and the role they play in human life.
Governing belief	There is a place from which to see reality that is independent of that reality; thus, there can be a nonpositioned observer who can grasp reality as it is without occupying any particular biasing standpoint.	We can only encounter reality from some standpoint; thus, the observer is always standing somewhere and is thereby necessarily a positioned observer
Guiding Metaphor	Science is like a mirror designed to reflect things as they really are.	Science is like a storyteller proposing accounts and versions of reality.
Methodological Priorities	Methods are designed to control those factors that would weaken the investigator's ability to discern the true shape of reality.	Broad social and historical factors always frame the investigator's understanding; the best we can achieve is a richer and deeper understanding based on encountering the historically and culturally diverse accounts that people use in making their lives sensible.

The phenomenon to be explained, of course, is school failure or, perhaps more accurately, school achievement. That is, school failure could be perceived as being at one end of the school achievement continuum with school success being at the other end. School failure is explained by specified relations between each of the seven variables and school failure, or by combinations of the seven variables and school failure. The scientist, successfully using this set of constructs, then "understands" school failure. He is able to "explain" and, to some extent at least, "predict" it.

It is obvious that explanation and prediction can be subsumed under theory. The very nature of a theory lies in its explanation of observed phenomena. Take reinforcement theory. A simple proposition flowing from this theory is: If a response is rewarded (reinforced) when it occurs, it will tend to be repeated. The psychological scientist whom first formulated some such proposition did so as an explanation of the observed repetitious occurrences of responses. Why did they occur and reoccur with dependable regularity? Because they were rewarded. This is an explanation, although it may not be a satisfactory explanation to many people. Someone else may ask why reward increases the likelihood of a response's occurrence. A full-blown theory would have the explanation. Today, however, there is no really satisfactory answer. All we can say is that, with a high degree of probability, the reinforcement of a response makes the response more likely to occur and reoccur (see Nisbett & Ross, 1980). In other words, the propositions of a theory, the statements of relations, constitute the explanation, as far as that theory is concerned, of observed natural phenomena.

Now, on prediction and control, it can be said that scientists do not really have to be concerned with explanation and understanding. Only prediction and control are necessary. Proponents of this point of view may say that the adequacy of a theory is its predictive power. If by using the theory we are able to predict successfully, then the theory is confirmed and this is enough. We need not necessarily look for further underlying explanations. Since we can predict reliably, we can control because control is deducible from prediction.

The prediction view of science has validity. But as far as this book is concerned, prediction is considered to be an aspect of theory. By its very nature, a theory predicts. That is, when from the primitive propositions of a theory we deduce more complex ones, we are in essence "predicting." When we explain observed phenomena, we are always stating a relation between, say, the class A and the class B. Scientific explanation inheres in specifying the relations between one class of empirical events and another, under certain conditions. We say: If A, then B, A and B referring to classes of objects or events.[1] But this is prediction, prediction from A to B. Thus a theoretical explanation implies prediction. And we come back to the idea that theory is the ultimate aim of science. All else flows from theory.

There is no intention here to discredit or denigrate research that is not specifically and consciously theory-oriented. Much valuable social scientific and

educational research is preoccupied with the shorter-range goal of finding specific relations; that is, merely to discover a relation is part of science. The ultimately most usable and satisfying relations, however, are those that are the most generalized, those that are tied to other relations in a theory.

The notion of generality is important. Theories, because they are general, apply to many phenomena and to many people in many places. A specific relation, of course, is less widely applicable. If, for example, one finds that test anxiety is related to test performance. This finding, though interesting and important, is less widely applicable and less understood than finding a relation in a network of interrelated variables that are parts of a theory. Modest, limited, and specific research aims, then, are good. Theoretical research aims are better because, among other reasons, they are more general and can be applied to a wide range of situations. Additionally, when both a simple theory and a complex one exist and both account for the facts equally well, the simple explanation is preferred. Hence in the discussion of generalizability, a good theory is also parsimonious. However, a number of incorrect theories concerning mental illness persist because of this parsimony feature. Some still believe that individuals are possessed with demons. Such an explanation is simple when compared to psychological and/or medical explanations.

Theories are tentative explanations. Each theory is evaluated empirically to determine how well it predicts new findings. Theories can be used to guide one's research plan by generating testable hypotheses and to organize facts obtained from the testing of the hypotheses. A good theory is one that cannot fit all observations. One should be able to find an occurrence that would contradict it. Blondlot's theory of N-Rays is an example of a poor theory. Blondlot claimed that all matter emitted N-Rays (Weber, 1973). Although N-Rays was later demonstrated to be nonexistent, Barber (1976) reported that nearly 100 papers were published in a single year on N-Rays in France. Blondlot even developed elaborate equipment for the viewing of N-Rays. Scientists claiming they saw N-Rays only added support to Blondlot's theory and findings. However, when a person did not see N-Rays, Blondlot claimed that the person's eyes were not sensitive enough or the person did not set up the instrument correctly. No possible outcome was taken as evidence against the theory. In more recent times, another faulty theory that took over 75 years to debunk concerned the origin of peptic ulcers. In 1910 Schwartz (as reported in Blaser, 1996) had claimed that he had firmly established the cause of ulcers. He stated that peptic ulcers were due to stomach acids. In the years that followed, medical researchers devoted their time and energy toward treating these ulcers by developing medication to either neutralize the acids or block them. These treatments were never totally successful and they were expensive. However in 1985, J. Robin Warren and Barry Marshall (as reported in Blaser, 1996) discovered that the helioc bacter pylori was the real culprit for stomach ulcers. Almost all cases of this type of ulcers were successfully treated with anti-biotics and for a considerably lower. For 75 years no possible outcome was taken as evidence against this stress-acid theory of ulcers.

1. Statements of the form "If p, then q," called conditional statements in logic, are the core of scientific inquiry. They and the concepts or variables that go into them are the central ingredient of theories. The logical foundation of scientific inquiry that underlies much of the reasoning in this book is outlined in Kerlinger(1977).

SCIENTIFIC RESEARCH A DEFINITION

It is easier to define scientific research than it is to define science. It would not be easy, however, to get scientists and researchers to agree on such a definition. Even so, we attempt one here:

Scientific research is systematic, controlled, empirical, amoral, public and critical investigation of natural phenomena. It is guided by theory and hypotheses about the presumed relations among such phenomena.

This definition requires little explanation since it is mostly a condensed and formalized statement of much that was said earlier or that will be said soon. Two points need emphasis, however. First, when we say that scientific research is systematic and controlled, we mean, in effect, that scientific investigation is so ordered that investigators can have critical confidence in research outcomes. As we shall see later, scientific research observations are tightly disciplined. Moreover, among the many alternative explanations of a phenomenon, all but one are systematically ruled out. One can thus have greater confidence that a tested relation is as it is than if one had not controlled the observations and ruled out alternative possibilities. In some instances a cause-and-effect relationship can be established.

Second, scientific investigation is empirical. If the scientist believes something is so, that belief must somehow or other put to an outside independent test. Subjective belief, in other words, must be checked against objective reality. Scientists must always subject their notions to the court of empirical inquiry and test. Scientists are hypercritical of the results of their own and others' research. Every scientist writing a research report has other scientists reading what one writes while he or she writes it. Though it is easy to err, to exaggerate, to overgeneralize when writing up one's own work, it is not easy to escape the feeling of scientific eyes constantly peering over one's shoulder.

In science there is peer review. This means that others of equal training and knowledge are called upon to evaluate another scientists work before it is published in scientific journals. There are both positive and negative points concerning this. It is through peer review that fraudulent studies have been exposed. The essay written by R. W. Wood (1973) on his experiences with Professor Blondlot of France concerning the nonexistence of N-rays gives a clear demonstration of peer review. Peer review works very well for science and it promotes quality research. The system however is not perfect. There are occasions where peer review worked against science. This is documented throughout history with people such as Kepler, Galileo, Copernicus, Jenner, and Semelweiss. The ideas of these individuals were just not popular with their peers. More recently in psychology, the works of John Garcia on the biological constraints on learning went contrary to his peers. Garcia managed to publish his findings in a journal (Bulletin of the Psychonomic Society) that did not have peer review. Others who read Garcia's work and replicated it found Garcia's work to be valuable. In the large majority of cases, peer review of science is beneficial.

Thirdly, knowledge obtained scientifically is not subject to moral evaluation. The results are neither considered "bad" or "good," but in terms of its validity and reliability. The scientific method is however subject to issues of morality. That is, scientists are held responsible for the methods used in obtaining scientific knowledge. In psychology, codes of ethics are enforced to protect those under study. Science is a cooperative venture. Information obtained from science is available to all. Plus the scientific method is well known and available to all that chooses to use it.

THE SCIENTIFIC APPROACH

The scientific approach is a special systematized form of all-reflective thinking and inquiry. Dewey (1933/1991), in his influential *How We Think,* outlined a general paradigm of inquiry. The present discussion of the scientific approach is based largely on Dewey's analysis.

Problem-Obstacle-Idea

The scientist will usually experience an obstacle to understanding, a vague unrest about observed and unobserved phenomena, a curiosity as to why something is as it is. The first and most important step is to get the idea out in the open, to express the problem in some reasonably manageable form. Rarely or never will the problem spring full-blown at this stage. The scientist must struggle with it, try it out, and live with it. Dewey (1933/1991, p. 108) says, "There is a troubled, perplexed, trying situation, where the difficulty is, as it were, spread throughout the entire situation, infecting it as a whole." Sooner or later, explicitly or implicitly, the scientist states the problem, even if the expression of it is inchoate and tentative. Here the scientist intellectualizes, as (1933/1991 p. 109) puts it, "what at first is merely *an emotional* quality of the whole situation."' In some respects, this is the most difficult and most important part of the whole process. Without some sort of statement of the problem, the scientist can rarely go further and expect the work to be fruitful. With some researchers, the idea may come from speaking to a colleague or observing a curious phenomenon. The idea here is that the problem usually begins with vague and/or unscientific thoughts or unsystematic hunches. It then goes through a step of refinement.

Hypothesis

After intellectualizing the problem, after referring to past experiences for possible solutions, after observing relevant phenomena, the scientist may formulate a hypothesis. A hypothesis is a conjectural statement, a tentative proposition about the relation between two or more phenomena or variables. Our scientist will say, "If such-and-such occurs, then so-and-so results."

Reasoning-Deduction

This step or activity is frequently overlooked or underemphasized. It is perhaps the most important part of Dewey's analysis of reflective thinking. The scientist deduces the consequences of the hypothesis he has formulated. Conant (1951), in talking about the

rise of modern science, said that the new element added in the seventeenth century was the use of deductive reasoning. Here is where experience, knowledge, and perspicacity are important.

Often the scientist, when deducing the consequences of a formulated hypothesis, will arrive at a problem quite different from the one started with. On the other hand, the deductions may lead to the belief that the problem cannot be solved with present technical tools. For example, before modern statistics was developed, certain behavioral research problems were insoluble. It was difficult, if not impossible, to test two or three interdependent hypotheses at one time. It was next to impossible to test the interactive effect of variables. And we now have reason to believe that certain problems are insoluble unless they are tackled in a multivariate manner. An example of this is teaching methods and their relation to achievement and other variables. It is likely that teaching methods, per se, do not differ much if we study only their simple effects. Teaching methods probably work differently under different conditions, with different teachers, and with different pupils. It is said that the methods "interact" with the conditions and with the characteristics of teachers and of pupils. Simon (1987) stated another example of this. Simon (1987) pointed out that a research on pilot training proposed by Williams and Adelson in 1954 could not be carried out using traditional experimental research methods. The study proposed to study 34 variables and their influence on pilot training. Using traditional research methods, the number of variables under study was too overwhelming. Over 20 years later, Simon (1976, 1984) showed that such studies could be effectively studied using economical multifactor designs.

An example may help us understand this reasoning-deduction step. Suppose an investigator becomes intrigued with aggressive behavior. The investigator wonders why people are often aggressive in situations where aggressiveness may not be appropriate. Personal observation leads to the notion that aggressive behavior seems to occur when people have experienced difficulties of one kind or another. (Note the vagueness of the problem here.) After thinking for some time, reading the literature for clues, and making further observations, the hypothesis is formulated: Frustration leads to aggression. "Frustration" is defined as prevention from reaching a goal and "aggression" as behavior characterized by physical or verbal attack on other persons or objects.

What follows from this is a statement like: If frustration leads to aggression, then we should find a great deal of aggression among children who are in schools that are restrictive, schools that do not permit children much freedom and self-expression. Similarly, in difficult social situations assuming such situations are frustrating, we should expect more aggression than is "usual." Reasoning further, if we give experimental subjects interesting problems to solve and then prevent them from solving them, we can predict some kind of aggressive behavior. In a nutshell, this process of moving from a broader picture to a more specific one is called deductive reasoning.

Reasoning may, as indicated above, change the problem. We may realize that the initial problem was only a special case of a broader, more fundamental and important problem. We may, for example start with a narrower hypothesis: Restrictive school situations lead to negativism in children. Then we can generalize the problem to the form: Frustration leads to aggression. While this is a different form of thinking from that discussed earlier, it is important because of what could almost be called its heuristic quality. Reasoning can help lead to wider, more basic, and thus more significant problems, as well as provide operational (testable) implications of the original hypothesis. This type of reasoning is called inductive reasoning. It starts from particular facts and moves to a general statement or hypothesis. If one is not careful, this method could lead to faulty reasoning. This is due to the method's natural tendency to exclude other data that do not fit the hypothesis. The inductive reasoning method is inclined to look for supporting data rather than refuting evidence.

Consider the classical study by Peter Wason (Wason and Johnson-Laird, 1972) that has been a topic of much interest (Hoch, 1986; Klayman and Ha, 1987). In this study, students were asked to discover a rule the experimenter had in mind that generated a sequence of numbers. One example was to generate a rule for the following sequence of numbers: "3, 5, 7." Students were told that they could ask about other sequences. Students would receive feedback on each sequence proposed as to whether it fits or does not fit the rule the experimenter had in mind. When the students felt confident, they could put forth the rule. Some students would offer "9, 11, 13." They are told that this sequence fits the rule. They may then follow with "15, 17, 19." And again they are told that this sequence fits. The students then may offer as their answer: "The rule is three consecutive odd numbers." They would be told that this is *not* the rule. Others that would be offered after some more proposed sequences are "increasing numbers in increments of two, or odd numbers in increments of two." In each of

these, they are told that it is not the rule that the experimenter was thinking of. The actual rule in mind was “any three increasing positive numbers.” Had the students’ proposed the sequences “8, 9, 10” or “1, 15, 4500” they would have been told that these also fit the rule. Where the students made their error was in testing only the cases that fitted their first proposed sequence that confirmed their hypothesis.

Although oversimplified, the Wason study demonstrated what could happen in actual scientific investigations. A scientist could easily be locked into repeating the same type of experiment that always supports the hypothesis.

Observation-Test-Experiment

It should be clear by now that the observation-test-experiment phase is only part of the scientific enterprise. If the problem has been well stated, the hypothesis or hypotheses adequately formulated, and the implications of the hypotheses carefully deduced, this step is almost automatic-assuming that the investigator is technically competent.

The essence of testing a hypothesis is to test the *relation* expressed by the hypothesis. We do not test variables, as such; we test the relation between the variables. Observation, testing, and experimentation are for one large purpose: putting the problem relation to empirical test. To test without knowing at least fairly well what and why one is testing is to blunder. Simply to state a vague problem, like ‘ How does Open Education affect learning?’ and then to test pupils in schools presumed to differ in "openness," or to ask: What are the effects of cognitive dissonance? and then, after experimental manipulations to create dissonance, to search for presumed effects could lead only to questionable information. Similarly, to say one is going to study attribution processes without really knowing why one is doing it or without stating relations between variables is research nonsense.

Another point about testing hypotheses is that we usually do not test hypotheses directly. As indicated in the previous step on reasoning, we test deduced implications of hypotheses. Our test hypothesis may be: "Subjects told to suppress unwanted thoughts will be more preoccupied with them than subjects who are given a distraction." This was deduced from a broader and more general hypothesis: “Greater efforts to suppress an idea leads to greater preoccupation with the idea." We do not test "suppression of ideas" nor "preoccupation." We test the relation between them, in this case the relation between suppression of unwanted thoughts and the level of preoccupation. (see Wegner, Schneider, Carter and White, 1987; Wegner, 1989)

Dewey emphasized that the temporal sequence of reflective thinking or inquiry is not fixed. We can repeat and reemphasize what he says in our own framework. The steps of the scientific approach are not neatly fixed. The first step is not neatly completed before the second step begins. Further, we may test before adequately deducing the implications of the hypothesis. The hypothesis itself may seem to need elaboration or refinement as a result of deducing implications from it. Hypotheses and their expression will often be found inadequate when implications are deduced from them. A frequent difficulty occurs when a hypothesis is so vague that one deduction is as good as another-that is, the hypothesis may not yield to precise test.

Feedback to the problem, the hypotheses, and, finally, the theory of the results of research is highly important. Learning theorists and researchers, for example, have frequently altered their theories and research as a result of experimental findings (see Malone, 1991; Schunk, 1996; Hergenhahn, 1996). Theorists and researchers have been studying the effects of early environment and training on later development. Kagan and Zentner (1996) reviewed the results of 70 studies concerned with the relation between early life experiences and psychopathology in adulthood. They found that juvenile delinquency could be predicted by the amount of impulsivity detected at preschool age. Lynch, Short and Chua (1995) found that musical processing was influenced by perceptual stimulation an infant experienced at age 6 months to 1 year. These and other research have yielded varied evidence converging on this extremely important theoretical and practical problem. Part of the essential core of scientific research is the constant effort to replicate and check findings, to correct theory on the basis of empirical evidence, and to find better explanations of natural phenomena. One can even go so far as to say that science has a cyclic aspect. A researcher finds, say, that A is related to B in such-and-such a way. Then more research is conducted to determine under what other conditions A is similarly related to B. Other researchers challenge this theory and this research, offering explanations and evidence of their own. The original researcher, it is hoped, alters one’s work in the light of new data. The process never ends.

Let us summarize the so-called scientific approach to inquiry. First there is doubt, a barrier, an

indeterminate situation crying out to be made determinate. The scientist experiences vague doubts, emotional disturbance, and inchoate ideas. There is a struggle to formulate the problem, even if inadequately. The scientist then studies the literature, scans one's own experience and the experience of others. Often the researcher simply has to wait for an inventive leap of the mind. Maybe it will occur; maybe not. With the problem formulated, with the basic question or questions properly asked, the rest is much easier. Then the hypothesis is constructed, after which its empirical implications are deduced. In this process the original problem, and of course the original hypothesis, may be changed. It may be broadened or narrowed. It may even be abandoned. Last, but not finally, the relation expressed by the hypothesis is tested by observation and experimentation. On the basis of the research evidence, the hypothesis is supported or rejected. This information is then fed back to the original problem, and the problem is kept or altered as dictated by the evidence. Dewey pointed out that one phase of the process may be expanded and be of great importance, another may be skimped, and there may be fewer or more steps involved. Research is rarely an orderly business anyway. Indeed, it is much more disorderly than the above discussion may imply. Order and disorder, however, are not of primary importance. What is much more important is the controlled rationality of scientific research as a process of reflective inquiry, the interdependent nature of the parts of the process, and the paramount importance of the problem and its statement.

Study Suggestion

Some of the content of this chapter is highly controversial. The views expressed are accepted by some thinkers and rejected by others. Readers can enhance understanding of science and its purpose, the relation between science and technology, and basic and applied research by selective reading of the literature. Such reading can be the basis for class discussions. Extended treatment of the controversial aspects of science, especially behavioral science, is given in the first author's book, *Behavioral Research: A Conceptual Approach.* New York: Holt, Rinehart and Winston, 1979, chaps. 1, 15, and 16. Many fine articles on science and research have been published in science journals and philosophy of science books. Here are some of them. Included is also a special report in *Scientific American*. All are pertinent to this chapter's substance.

Barinaga, M. (1993). Philosophy of science: Feminists find gender everywhere in science. *Science, 260*, 392-393. Discusses the difficulty of separating cultural views of women and science. Talks about science as a predominantly male field.

Editors of *Scientific American*. Science versus antiscience. (Special report). January 1997, 96-101. Presents three different anti-science movements: creationist, feminist, and media.

Hausheer, J. & Harris, J. (1994). In search of a brief definition of science. *The Physics Teacher, 32(5),* 318. Mentions that any definition of science must include guidelines for evaluating theory and hypotheses as either science or nonscience.

Holton, G. (1996). The controversy over the end of science. *Scientific American, 273(10),* 191. This is an article concerned with the development of two camps of thought: the linearists and the cyclists. The Linearists take a more conventional perspective of science. The cyclists see science as degenerating within itself.

Horgan, J. (1994). Anti-omniscience: An eclectic gang of thinkers pushes at knowledge's limits. *Scientific American, 271*, 20-22. Discusses the limits of science. Also see Horgan, J. (1997). *The end of science*. New York: Broadway Books.

Miller, J. A (1994). Postmodern attitude toward science. *Bioscience, 41(6),* 395.

Discues the reasons some educators and scholars in the humanities have adopted a hostile attitude toward science.

Smith, B. (1995). Formal ontology, common sense and cognitive science. *International Journal of Human-Computer Studies, 43(5-6),* 641-667. An article examining common sense and cognitive science.

Timpane, J. (1995). How to convince a reluctant scientist. *Scientific American, 272*, 104. This article warns that too much originality in science would lead to non-acceptance and difficulty of understanding. It also discusses how scientific acceptance is governed by both

old and new data and the reputation of the scientist.

Chapter Outline-Summary

1. To understand complex human behavior, one must understand the scientific language and approach.
2. Science is a systematic and controlled extension of common sense. There are 5 differences between science and common sense
   a) science uses conceptual schemes and theoretical structures
   b) science systematically and empirically tests theories and hypotheses
   c) science attempts to control possible extraneous causes
   d) science consciously and systematically pursues relations
   e) science rules out metaphysical (untestable) explanations
3. Peirce's Four Methods of Knowing
   a) method of tenacity -- influenced by established past beliefs
   b) method of authority -- influenced by the weight of tradition or public sanction
   c) a priori method -- also known as the method of intuition. Natural inclination toward the truth.
   d) method of science - self-correcting. Notions are testable and objective
4. The stereotype of science has hindered understanding of science by the public
5. Views, functions of science
   a) static view sees science contributing scientific information to world. Science adds to the body of information and present state of knowledge.
   b) dynamic view is concerned with the activity of science. What scientists do. With this comes the heuristic view of science. This is one of self-discovery. Science takes risks and solves problems.
6. Aims of science
   a) develop theory and explain natural phenomenon
   b) promote understanding and develop predictions
7. A theory has 3 elements
   a) set of properties consisting of defined and interrelated constructs
   b) systematically sets the interrelations among a set of variables.
   c) explains phenomenon
8. Scientific research is systematic, controlled, empirical and critical investigation of natural phenomenon. It is guided by theory and hypotheses about presumed relations among such phenomenon. It is also public and amoral.
9. The scientific approach according to Dewey is made up of the following:
   a) Problem-Obstacle-Idea -- formulate the research problem or question to be solved.
   b) Hypothesis -- formulate a conjectural statement about the relation between phenomenon or variables
   c) Reasoning-Deduction -- scientist deduces the consequences of the hypothesis. This can lead to a more significant problem and provide ideas on how the hypothesis can be tested in observable terms.
   d) Observation-Test-Experiment- this is the data collection and analysis phase. The results of the research conducted are related back to the problem.

Chapter 2

Problems and Hypotheses

MANY PEOPLE think that science is basically a fact-gathering activity. It is not. As M. R. Cohen (1956/ 1997, p. 148) says:

> There is.. no genuine progress in scientific insight through the Baconian method of accumulating empirical facts without hypotheses or anticipation of nature. Without some guiding idea we do not know what facts to gather... we cannot determine what is relevant and what is irrelevant.

The scientifically uninformed person often has the idea that the scientist is a highly objective individual who gathers data without preconceived ideas. Poincare (1952/1996, p. 143) long ago pointed out how wrong this idea is. He said: "It is often said that experiments should be made without preconceived ideas. That is impossible. Not only would it make every experiment fruitless, but even if we wished to do so, it could not be done."

PROBLEMS

It is not always possible for a researcher to formulate the problem simply, clearly, and completely. The research may often have only a rather general, diffuse, even confused notion of the problem. This is in the nature of the complexity of scientific research. It may even take an investigator years of exploration, thought, and research before the researcher can clearly state the questions. Nevertheless, adequate statement of the research problem is one of the most important parts of research. The difficulty of stating a research problem satisfactorily at a given time should not allow us to lose sight of the ultimate desirability and necessity of doing so.

Bearing this difficulty in mind, a fundamental principle can be stated: If one wants to solve a problem, one must generally know what the problem is. It can be said that a large part of the solution lies in knowing what it is one is trying to do. Another part lies in knowing what a problem is and especially what a scientific problem is.

What is a good problem statement? Although research problems differ greatly and there is no one "right" way to state a problem, certain characteristics of problems and problem statements can be learned and used to good advantage. To start, let us take two or three examples of published research problems and study their characteristics. First, take the problem of the study by Hurlock (1925)[1] mentioned in Chapter 1: What is the effects on pupil performance of different types of incentives? Note that the problem is stated in question form. The simplest way is here the best way. Also note that the problem states a relation between variables, in this case between the variables incentives and pupil performance (achievement). ("Variable" will be formally defined in Chapter 3. For now, a variable is the name of a phenomenon, or a construct, that takes a set of different numerical values.)

A problem, then, is an interrogative sentence or statement that asks: What relation exists between two or more variables? The answer is what is being sought in the research. A problem in most cases will have two or more variables. In the Hurlock example, the problem statement relates incentive to pupil performance. Another problem, studied in an influential experiment by Bahrick (1984, 1992), is associated with the age-old questions: How much of what you are studying now will you remember 10 years from now? How much of it will you remember 50 years from today? How much of it will you remember later if you never used it? Formally Bahrick asks: Does semantic memory involve separate processes? One variable is the amount of time since the material was first learned, a second would be the quality of original learning and the other variable is remembering (or forgetting). Still another problem, by Little, Sterling and Tingstrom (1996), is quite different: Does geographic and racial cues influence attribution (perceived blame)? One variable geographical cues, a second would be racial information and the other is attribution. Not all research problems clearly have two or more variables. For example, in experimental psychology, the research focus is often on psychological processes like memory and categorization. In her justifiably well known and influential study of perceptual categories, Rosch(1973) in effect asked the question: Are there nonarbitrary ("natural") categories of color and form? Although the relation between two or more variables is not apparent in this problem statement, in the actual research the categories were related to learning.

---

1. When citing problems and hypotheses from the literature, we have not always used the words of the authors. In fact, the statements of many of the problems are ours and not those of the cited authors. Some authors use only problem statements; some use only hypotheses; others use both.

Toward the end of this book we will see that factor analytic research problems also lack the relation form discussed above. In most behavioral research problems, however, the relations among two or more variables are studied, and we will therefore emphasize such relation statements.

Criteria of Problems and Problem Statements

There are three criteria of good problems and problem statements. One, the problem should express a relation between two or more variables. It asks, in effect, questions like: Is A related to B? How are A and B related to C? How is A related to B under conditions C and D? The exceptions to this dictum occur mostly in taxonomic or methodological research.

Two, the problem should be stated clearly and unambiguously in question form. Instead of saying, for instance, "The problem is..." or "The purpose of this study is... ", ask a question. Questions have the virtue of posing problems directly. The purpose of a study is not necessarily the same as the problem of a study. The purpose of the Hurlock study, for instance, was to throw light on the use of incentives in school situations. The problem was the question about the relation between incentives and performance. Again, the simplest way is the best way: ask a question.

The third criterion is often difficult to satisfy. It demands that the problem and the problem statement should be such as to *imply* possibilities of empirical testing. A problem that does not contain implications for testing its stated relation or relations is not a scientific problem. This means not only that an actual relation is stated, but also that the variables of the relation can somehow be measured. Many interesting and important questions are not scientific questions simply because they are not amenable to testing. Certain philosophic and theological questions, while perhaps important to the individuals who consider them, cannot be tested empirically and are thus of no interest to the scientist as a scientist. The epistemological question, "How do we know?," is such a question. Education has many interesting but nonscientific questions, such as, "Does democratic education improve the learning of youngsters?" "Are group processes good for children?" These questions can be called metaphysical in the sense that they are, at least as stated, beyond empirical testing possibilities. The key difficulties are that some of them are not relations, and most of their constructs are very difficult or impossible to so define that they can be measured.

HYPOTHESES

A hypothesis is a conjectural statement of the relation between two or more variables. Hypotheses are always in declarative sentence form, and they relate, either generally or specifically, variables to variables. There are two criteria for "good" hypotheses and hypothesis statements. They are the same as two of those for problems and problem statements. One, hypotheses are statements about the relations between variables. Two, hypotheses carry clear implications for testing the stated relations. These criteria mean, then, that hypothesis statements contain two or more variables that are measurable or potentially measurable and that they specify how the variables are related.

Let us take three hypotheses from the literature and apply the criteria to them. The first hypothesis from a study by Wegner, et al (1987). seems to defy common sense: The greater the suppression of unwanted thoughts, the greater the preoccupation with those unwanted thoughts (suppress now; obsess later). Here a relation is stated between one variable, suppression of an idea or thought, and another variable, preoccupation or obsession. Since the two variables are readily defined and measured, implications for testing the hypothesis, too, are readily conceived. The criteria are satisfied. In the Wegner, et al study, subjects were asked to *not* think about a "white bear." Each time they did think of the white bear, they would ring a bell. The number of bell rings indicated the level of preoccupation. A second hypothesis is from a study by Ayres and Hughes (1986). This study's hypothesis is unusual. It states a relation in the so-called null form: Levels of noise or music has no effect on visual functioning. The relation is stated clearly: one variable, loudness of sound (like music), is related to another variable, visual functioning, by the words "has no effect on." On the criterion of potential testability, however, we meet with difficulty. We are faced with the problem of defining "visual functioning" and "loudness" so they can be are measured. If we can solve this problem satisfactorily, then we definitely have a hypothesis. Ayres and Hughes did solve this by defining loudness as 107 decibels and visual functioning in terms of a score on a visual acuity task. And this hypothesis did lead to answering a question that people often ask: "Why do we turn down the volume of the car stereo when we are *looking* for a street address?" Ayres and Hughes found a definite drop in perceptual

functioning when the level of music was at 107 decibels.

The third hypothesis represents a numerous and important class. Here the relation is indirect, concealed, as it were. It customarily comes in the form of a statement that Groups A and B will differ on some characteristic. For example: Women more often than men believe they should lose weight even though their weight is well within normal bounds (Fallon & Rozin, 1985). That is, women differ from men in terms of their perceived body shape. Note that this statement is one step removed from the actual hypothesis, which may be stated: Perceived body shape is in part a function of gender. If the latter statements were the hypothesis stated, then the first might be called a subhypothesis, or a specific prediction based on the original hypothesis.

Let us consider another hypothesis of this type but removed one step further. Individuals having the same or similar characteristics will hold similar attitudes toward cognitive objects significantly related to the occupational role (Saal and Moore, 1993). ("Cognitive objects" are any concrete or abstract things perceived and "known" by individuals. People, groups, job or grade promotion, the government, and education are examples.) The relation in this case, of course, is between personal characteristics and attitudes (toward a cognitive object related to the personal characteristic, for example, gender and attitudes toward others receiving a promotion). In order to test this hypothesis, it would be necessary to have at least two groups, each with a different characteristic, and then to compare the attitudes of the groups. For instance, as in the case of the Saal and Moore study, the comparison would be between men and women. They would be compared on their assessment of fairness toward a promotion given to a co-worker of the opposite or same sex. In this example, the criteria are satisfied.

THE IMPORTANCE OF PROBLEMS AND HYPOTHESES

There is little doubt that hypotheses are important and indispensable tools of scientific research. There are three main reasons for this belief. One, they are, so to speak, the working instruments of theory. Hypotheses can be deduced from theory and from other hypotheses. If, for instance, we are working on a theory of aggression, we are presumably looking for causes and effects of aggressive behavior. We might have observed cases of aggressive behavior occurring after frustrating circumstances. The theory, then, might include the proposition: Frustration produces aggression (Dollard, Doob, Miller, Mowrer & Sears, 1939; Berkowitz, 1983; Dill & Anderson, 1995). From this broad hypothesis we may deduce more specific hypotheses, such as: To prevent children from reaching desired goals (frustration) will result in their fighting each other (aggression); if children are deprived of parental love (frustration), they will react in part with aggressive behavior.

The second reason is that hypotheses can be tested and shown to be probably true or probably false. Isolated facts are not tested, as we said before; only relations are tested. Since hypotheses are relational propositions, this is probably the main reason they are used in scientific inquiry. They are, in essence, predictions of the form, "If A, then B," which we set up to test the relation between A and B. We let the facts have a chance to establish the probable truth or falsity of the hypothesis.

Three, hypotheses are powerful tools for the advancement of knowledge because they enable scientists to get outside themselves. Though constructed by humans, hypotheses exist, can be tested, and can be shown to be probably correct or incorrect apart from person's values and opinions. This is so important that we venture to say that there would be no science in any complete sense without hypotheses.

Just as important as hypotheses are the problems behind the hypotheses. As Dewey (1938/1982, pp. 105-107) has well pointed out, research usually starts with a problem. He says that there is first an indeterminate situation in which ideas are vague, doubts are raised, and the thinker is perplexed. Dewey further points out that the problem is not enunciated, indeed cannot be enunciated, until one has experienced such an indeterminate situation.

The indeterminacy, however, must ultimately be removed. Though it is true, as stated earlier, that a researcher may often have only a general and diffuse notion of the problem, sooner or later he or she has to have a fairly clear idea of what the problem is. Though this statement seems self-evident, one of the most difficult things to do is to state one's research problem clearly and completely. In other words, you must know what you are trying to find out. When you finally do know, the problem is a long way toward solution.

VIRTUES OF PROBLEMS AND HYPOTHESES

Problems and hypotheses, then, have important virtues. One, they direct investigation. The relations

expressed in the hypotheses tell the investigator what to do. Two, problems and hypotheses, because they are ordinarily generalized relational statements, enable the researcher to deduce specific empirical manifestations implied by the problems and hypotheses. We may say, following Guida and Ludlow (1989): If it is indeed true that children in one type of culture (Chile) have higher test anxiety than children of another type of culture (White-Americans), then it follows that children in the Chilean culture should do more poorly in academics than children in the American culture. They should perhaps also have a lower self-esteem or more external locus-of-control when it comes to school and academics.

There are important differences between problems and hypotheses. Hypotheses, if properly stated, can be tested. While a given hypothesis may be too broad to be tested directly, if it is a "good" hypothesis, then other testable hypotheses can be deduced from it. Facts or variables are not tested as such. The relations stated by the hypotheses are tested. And a problem cannot be scientifically solved unless it is reduced to hypothesis form, because a problem is a question, usually of a broad nature, and is not directly testable. One does not test the questions: Does the presence or absence of a person in a public restroom alter personal hygiene (Pedersen, Keithly and Brady, 1986)? Does group counseling sessions reduce the level of psychiatric morbidity in police officers? (Doctor, Cutris and Issacs, 1994)? Instead, one tests one or more hypotheses implied by these questions. For example, to study the latter problem, one may hypothesize that police officers who attend stress reduction counseling sessions will take fewer sick days off than those police officers who did not attend counseling sessions. The hypothesis in the former problem could state that the presence of a person in a public restroom will cause the other person to wash their hands.

Problems and hypotheses advance scientific knowledge by helping the investigator confirm or disconfirm theory. Suppose a psychological investigator gives a number of subjects three or four tests, among which is a test of anxiety and an arithmetic test. Routinely computing the intercorrelations between the three or four tests, one finds that the correlation between anxiety and arithmetic is negative. One concludes, therefore that the greater the anxiety the lower the arithmetic score. But it is quite conceivable that the relation is fortuitous or even spurious. If, however, the investigator had hypothesized the relation on the basis of theory, the investigator could have greater confidence in the results. Investigators who do not hypothesize relations in advance, in short, do not give the facts a chance to prove or disprove anything. The words "prove" and "disprove" are not to be taken here in their usual literal sense. A hypothesis is never really proved or disproved. To be more accurate we should probably say something like: The weight of evidence is on the side of the hypothesis, or the weight of the evidence casts doubt on the hypothesis. Braithwaite(1953/1996 p. 14) says:

> "Thus the empirical evidence of its instance never proves the hypothesis: in suitable cases we may say that *it establishes* the hypothesis, meaning by this that the evidence makes it reasonable to accept the hypothesis; but it never *proves* the hypothesis in the sense that the hypothesis is a logical consequence of the evidence."

This use of the hypothesis is similar to playing a game of chance. The rules of the game are set up in advance, and bets are made in advance. One cannot change the rules after an outcome, nor can one change one's bets after making them. That would not be "fair." One cannot throw the dice first and then bet. Similarly, if one gathers data first and then selects a datum and comes to a conclusion on the basis of the datum, one has violated the rules of the scientific game. The game is not "fair" because the investigator can easily capitalize on, say, two significant relations out of five tested. What happens to the other three? They are usually forgotten. But in a fair game every throw of the dice is counted, in the sense that one either wins or does not win on the basis of the outcome of each throw.

Hypotheses direct inquiry. As Darwin pointed out over a hundred years ago, observations have to be for or against some view if they are to be of any use. Hypotheses incorporate aspects of the theory under test in testable or near-testable form. Earlier, an example of reinforcement theory was given in which testable hypotheses were deduced from the general problem. The importance of recognizing this function of hypotheses may be shown by going through the back door and using a theory that is very difficult, or perhaps impossible, to test. Freud's theory of anxiety includes the construct of repression. Now, by repression Freud meant the forcing of unacceptable ideas deep into the unconscious. In order to test the Freudian theory of anxiety it is necessary to deduce relations suggested by the theory. These deductions will, of course, have to include the repression notion,

which includes the construct of the unconscious. Hypotheses can be formulated using these constructs; in order to test the theory, they have to be so formulated. But testing them is another, more difficult matter because of the extreme difficulty of so defining terms such as "repression" and "unconscious" that they can be measured. To the present, no one has succeeded in defining these two constructs without seriously departing from the original Freudian meaning and usage. Hypotheses, then, are important bridges between theory and empirical inquiry.

PROBLEMS, VALUES, AND DEFINITIONS

To clarify further the nature of problems and hypotheses, two or three common errors will now be discussed. First, scientific problems are not moral and ethical questions. Are punitive disciplinary measures bad for children? Should an organization's leadership be democratic? What is the best way to teach college students? To ask these questions is to ask value and judgmental questions that science cannot answer. Many so-called hypotheses are not hypotheses at all. For instance: The small-group method of teaching is better than the lecture method. This is a value statement; it is an article of faith and not a hypothesis. If it was possible to state a relation between the variables, and if it was possible to define the variables so as to permit testing the relation, then we might have a hypothesis. But there is no way to test value questions scientifically.

A quick and relatively easy way to detect value questions and statements is to look for words such as "should," "ought," "better than" (instead of "greater than"). Also one can look for similar words that indicate cultural or personal judgments or preferences. Value statements, however, are tricky. While a "should" statement is obviously a value statement, certain other kinds of statements are not so obvious. Take the statement: Authoritarian methods of teaching lead to poor learning. Here there is a relation. But the statement fails as a scientific hypothesis because it uses two value expressions or words, "authoritarian methods of teaching" and "poor learning." Neither of which can be defined for measurement purposes without deleting the words "authoritarian" and "poor."[1]

Other kinds of statements that are not hypotheses or are poor ones are frequently formulated, especially in education. Consider, for instance: The core curriculum is an enriching experience. Another type, too frequent, is the vague generalization: Reading skills can be identified in the second grade; the goal of the unique individual is self-realization; Prejudice is related to certain personality traits.

Another common defect of problem statements often occurs in doctoral theses: the listing of methodological points or "problems" as subproblems. These methodological points have two characteristics that make them easy to detect: (1) they are not substantive problems that spring from the basic problem; and (2) they relate to techniques or methods of sampling, measuring, or analyzing. They are usually not in question form, but rather contain the words "test," "determine," "measure," and the like. "To determine the reliability of the instruments used in this research," "To test the significance of the differences between the means," and "To assign pupils at random to the experimental groups" are examples of this mistaken notion of problems and subproblems.

GENERALITY AND SPECIFICITY OF PROBLEMS AND HYPOTHESES

One difficulty that the research worker usually encounters and that almost all students working on a thesis find annoying is the generality and specificity of problems and hypotheses. If the problem is too general, it is usually too vague to be tested. Thus, it is scientifically useless, though it may be interesting to read. Problems and hypotheses that are too general or too vague are common. For example: Creativity is a function of the self-actualization of the individual; Democratic education enhances social learning and citizenship; Authoritarianism in the college classroom inhibits the creative imagination of students. These are interesting problems. But, in their present form, they are worse than useless scientifically, because they cannot be tested and because they give one the spurious assurance that they are hypotheses that can "some day" be tested.

Terms such as "creativity," "self-actualization," "democracy," "authoritarianism," and the like have, at the present time at least, no adequate empirical referents.[2] Now, it is quite true that we can define "creativity," say, in a limited way by specifying one or two creativity tests. This may be a legitimate

---

1. An almost classic case of the use of the word "authoritarian" is the statement sometimes heard among educators: The lecture method is authoritarian. This seems to mean that the speaker does not like the lecture method and is telling us that it is bad. Similarly, one of the most effective ways to criticize a teacher is to say that teacher is authoritarian.

procedure. Still, in so doing, we run the risk of getting far away from the original term and its meaning. This is particularly true when we speak of artistic creativity. We are often willing to accept the risk in order to be able to investigate important problems, of course. Yet terms like "democracy" are almost hopeless to define. Even when we do define it, we often find we have destroyed the original meaning of the term. An outstanding exception to this statement is Bollen's (1980) definition and measurement of "democracy." We will examine both in subsequent chapters.

The other extreme is too great specificity. Every student has heard that it is necessary to narrow problems down to workable size. This is true. But, unfortunately, we can also narrow the problem out of existence. In general, the more specific the problem or hypothesis the clearer are its testing implications. But triviality may be the price we pay. Researchers cannot handle problems that are too broad because they tend to be too vague for adequate research operations. On the other hand, in their zeal to cut the problems down to workable size or to find a workable problem, they may cut the life out of it. They may make it trivial or inconsequential. A thesis for instance, on the simple relation between the speed of reading and size of type, while important and maybe even interesting, is too thin by itself for doctoral study. The doctoral student would need to expand on the topic by also recommending a comparison between genders and considering variables such as culture and family background. The researcher could possibly expand the study to look at levels of illumination and font types. Too great specificity is perhaps a worse danger than too great generality. The researcher may be able to answer the specific question but will not be able to generalize the finding to other situations or groups of people. At any rate, some kind of compromise must be made between generality and specificity. The ability effectively to make such compromises is a function partly of experience and partly of critical study of research problems.

Here are a few examples contrasting research problems stated as too general or too specific:

1. Too General: There are gender differences in game playing.

   Too Specific: Tommy's score will be 10 points higher than Carol's on Tetris Professional Gold.

   About Right: Video game playing will result in a higher transfer of learning for boys than girls.

2. Too General: People can read large size letters faster than small size letters.

   Too Specific: Seniors at Duarte High School can read 24 point fonts faster than 12 point fonts.

   About Right: A comparison of three different font sizes and visual acuity on reading speed and comprehension.

THE MULTIVARIABLE NATURE OF BEHAVIORAL RESEARCH AND *PROBLEMS*

Until now the discussion of problems and hypotheses has been pretty much limited to two variables, x and y. We must hasten to correct any impression that such problems and hypotheses are the norm in behavioral research. Researchers in psychology, sociology, education, and other behavioral sciences have become keenly aware of the multivariable nature of behavioral research. Instead of saying: If p, then q, it is often more appropriate to say: If p_1, p_2,..., p_k, then q; or: If p then q, under conditions r, s, and t.

An example may clarify the point. Instead of simply stating the hypothesis: If frustration, then aggression, it is more realistic to recognize the multivariable nature of the determinants and influences of aggression. This can be done by saying, for example: If high intelligence, middle class, male, and frustrated, then aggression. Or: If frustration, then aggression, under the conditions of high intelligence, middle class, and male. Instead of one x, we now have four x's. Although one phenomenon may be the most important in determining or influencing another phenomenon, it is unlikely that most of the phenomena of interest to behavioral scientists are determined simply. It is much more likely that they are determined multiply. It is much more likely that aggression is the result of several influences working in complex ways. Moreover, aggression itself has multiple aspects. After all, there are different kinds of aggression.

Problems and hypotheses thus have to reflect the multivariable complexity of psychological, sociological, and educational reality. We will talk of one x and one y, especially in the early part of the book. However, *it must be* understood that behavioral research, which used to be almost exclusively

---

2. Although many studies of authoritarianism have been done with considerable success, it is doubtful that we know what authoritarianism in the classroom means. For instance, an action of a teacher that is authoritarian in one classroom may not be authoritarian in another classroom. The alleged democratic behavior exhibited by one teacher may even be called authoritarian if exhibited by another teacher. Such elasticity is not the stuff of science.

univariate in its approach, has become more and more multivariable. We have purposely used the word "multivariable" instead of "multivariate" for an important reason. Traditionally, "multivariate" studies are those that have more than one y variable and one or more x variables. When we speak of one y and more than one x variable, we use the more appropriate term "multivariable" to make the distinction. For now, we will use "univariate" to indicate one x and one y. "Univariate," strictly speaking, also applies to y. We will soon encounter multivariate conceptions and problems. And later parts of the book will be especially concerned with a multivariate approach and emphasis. For a clear explanation on the differences between multivariable and multivariate see Kleinbaum, Kupper, Muller and Nizam (1997).

CONCLUDING REMARKS-THE SPECIAL POWER OF HYPOTHESES

One will sometimes hear that hypotheses are unnecessary in research. Some feel that they unnecessarily restrict the investigative imagination, and that the job of science and scientific investigation is to find out things and not to belabor the obvious. Some feel that hypotheses are obsolete, and the like. Such statements are quite misleading. They misconstrue the purpose of hypotheses.

It can almost be said that the hypothesis is one of the most powerful tools yet invented to achieve dependable knowledge. We observe a phenomenon. We speculate on possible causes. Naturally, our culture has answers to account for most phenomena, many correct, many incorrect, many a mixture of fact and superstition, many pure superstition. It is the business of scientists to doubt most explanations of phenomena. Such doubts are systematic. Scientists insist upon subjecting explanations of phenomena to controlled empirical test. In order to do this, they formulate the explanations in the form of theories and hypotheses. In fact, the explanations are hypotheses. Scientists simply discipline the business by writing systematic and testable hypotheses. If an explanation cannot be formulated in the form of a testable hypothesis, then it can be considered to be a metaphysical explanation and thus not amenable to scientific investigation. As such, it is dismissed by scientists as being of no interest.

The power of hypotheses goes further than this, however. A hypothesis is a prediction. It says that if x occurs, y will also occur. That is, y is predicted from x. If, then, x is made to occur (vary), and it is observed that y also occurs (varies concomitantly), then the hypothesis is confirmed. This is more powerful evidence than simply observing, without prediction, the covarying of x and y. It is more powerful in the betting-game sense discussed earlier. The scientist makes a bet that x leads to y. If, in an experiment, x does lead to y, then one has won the bet. A person cannot just enter the game at any point and pick a perhaps fortuitous common occurrence of x and y. Games are not played this way (at least in our culture). This person must play according to the rules, and the rules in science are made to minimize error and fallibility. Hypotheses are part of the rules of the science game.

Even when hypotheses are not confirmed, they have power. Even when y does not covary with x, knowledge is advanced. Negative findings are sometimes as important as positive ones, since they cut down the total universe of ignorance and sometimes point up fruitful further hypotheses and lines of investigation. *But the scientist cannot tell positive from negative evidence unless one uses hypotheses.* It is possible to conduct research without hypotheses, of course, particularly in exploratory investigations. But it is hard to conceive modern science in all its rigorous and disciplined fertility without the guiding power of hypotheses.

Study Suggestions

1. Use the following variable names to write research problems and hypotheses: frustration, academic achievement, intelligence, verbal ability, race, social class (socioeconomic status), sex, reinforcement, teaching methods occupational choice, conservatism, education, income, authority, need for achievement, group cohesiveness, obedience, social prestige, permissiveness.

2. Ten problems from the research literature are given below. Study them carefully, choose two or three, and construct hypotheses based on them.

(a) Do children of different ethnic groups have different levels of test anxiety? (Guida and Ludlow, 1989).

(b) Does cooperative social situations lead to higher levels of intrinsic motivation? (Hom, Berger, et al., 1994).

(c) Are affective responses influenced by people's facial activity? (Strack, Martin and Stepper, 1988).

(d) Will jurors follow prohibitive judicial instructions and information? (Shaw and Skolnick, 1995).

(e) What are the positive effects of using alternating pressure pads to prevent pressure sores in homebound hospice patients? (Stoneberg, Pitcock, and Myton, 1986).

(f) What are the effects of early Pavlovian conditioning on later Pavlovian conditioning? (Lariviere and Spear, 1996).

(g) Does the efficacy of encoding information into long-term memory depend on the novelty of the information? (Tulving and Kroll, 1995).

(h) What is the effect of alcohol consumption on the likelihood of condom use during causal sex? (MacDonald, Zanna and Fong, 1996).

(I) Are there gender differences in predicting retirement decisions? (Talaga and Beehr, 1995).

(j) Is the Good Behavior Game a viable intervention strategy for children in a classroom that require behavior change procedures? (Tingstrom, 1994).

3. Ten hypotheses are given below. Discuss possibilities of testing them. Then read two or three of the studies to see how the authors tested them.

(a) Job applicants who claim a great deal of experience at nonexistent tasks overstate their ability on real tasks. (Anderson, Warner and Spencer, 1984).

(b) In social situations, men misread women's intended friendliness as a sign of sexual interest. (Saal, Johnson and Weber, 1989)

(c) The greater the team success, the greater the attribution of each team member toward one's ability and luck. (Chambers and Abrami, 1991).

(d) Increasing interest in a task will increase compliance. (Rind, 1997).

(e) Extracts from men's perspiration can affect women's menstrual cycle. (Cutler, Preti, Kreiger and Huggins, 1986).

(f) Physically attractive people are viewed with higher intelligence than non-attractive people. (Moran and McCullers, 1984).

(g) One can receive help from a stranger if that stranger is similar to oneself, or if the request is made at a certain distance. (Glick, DeMorest and Hotze, 1988).

(h) Cigarette smoking (nicotine) improves mental performance. (Spilich, June, and Remer, 1992).

(i) People stowing valuable items in unusual locations will have better memory of that location than stowing valuable items in usual locations. (Winograd and Soloway, 1986).

(j) Gay men with symptomatic HIV disease are significantly more distressed than gay men whose HIV status is unknown. (Cochran and Mays, 1994).

4. Multivariate (for now, more than two dependent variables) problems and hypotheses have become common in behavioral research. To give the student a preliminary feeling for such problems, we here append several of them. Try to imagine how you would do research to study them.

(a) Do men and women differ in their perceptions of their genitals, sexual enjoyment, oral sex and masturbation? (Reinholtz and Muehlenhard, 1995).

(b) Are youthful smokers more extroverted while older smokers are more depressed and withdrawn (Stein, Newcomb and Bentler, 1996)?

(c) How much do teacher's ratings of social skills for popular students different from rejected students? (Stuart, Gresham and Elliot, 1991); (Frentz, Gresham, and Elliot, 1991).

(d) Do counselor-client matching on ethnicity, gender and language influence treatment outcomes of school aged children? (Hall, Kaplan and Lee, 1994).

(e) Are there any differences in the cognitive and functional abilities of Alzheimer's patients who reside at a special care unit versus those residing at a traditional care unit? (Swanson, Maas, and Buckwalter, 1994).

(f) Do hyperactive children with attention deficit differ from nonhyperactive children with attention deficit on reading, spelling and written language achievement? (Elbert, 1993).

(g) Will perceivers see women who prefer the courtesy title of Ms. as being higher on instrumental qualities and lower on expressiveness qualities than women who prefer traditional courtesy titles? (Dion and Cota, 1991).

(h) Will an empowering style of leadership increase team member satisfaction and perceptions of team efficacy to increase team effectiveness? (Kumpfer, Turner, Hopkins, and Librett, 1993)

(i) How do ethnicity, gender and socioeconomic background influence psychosis-proneness: perceptual aberration, magical ideation and

schizotypal personality (Porch, Ross, Hanks and Whitman, 1995).

(j) Does stimulus exposure have two effects, one cognitive and one affective, which in turn affect liking, familiarity, and recognition confidence and accuracy (Zajonc, 1980)?

The Last two problems and studies are quite complex because the relations stated are complex. The other problems and studies, though also complex, have only one phenomenon presumably affected by other phenomena, whereas the last two problems have several phenomena influencing two or more other phenomena. Readers should not be discouraged if they find these problems a bit difficult. By the end of the book they should appear interesting and natural.

Chapter Summary

1. Formulating the research problem is not an easy task. Researcher starts with a general, diffused and vague notion and then gradually refines it. Research problems differ greatly and there is no one right way to state the problem.
2. Three criteria of a good problem and problem statement
   a) The problem should be expressed as a relation between two or more variables.
   b) The problem should be put in the form of a question.
   c) The problem statement should imply the possibilities of empirical testing.
3. A hypothesis is a conjectural statement of the relation between two or more variables. It is put in the form of a declarative statement. A criteria for a good hypothesis is the same as a) and b) in criteria of a good problem.
4. Importance of problems and hypotheses.
   a) It is a working instrument of science. It is a specific working statement of theory.
   b) Hypotheses can be tested and be predictive.
   c) Advances knowledge.
5. Virtues of problems and hypotheses
   a) Directs investigation and inquiry
   b) Enables research to deduce specific empirical manifestations
   c) Serves as the bridge between theory and empirical inquiry.
6. Scientific problems are not moral or ethical questions. Science cannot answer value or judgmental questions.
7. Detection of value questions: Look for words such as “better than”, “should”, or “ought.”
8. Another common defect of problem statements is the listing of methodological points as subproblems.
   a) they are not substantive problems that come directly from the basic problem
   b) they relate to techniques or methods of sampling, measuring, or analyzing. Not in question form.
9. On problems, there is a need to compromise between being too general and too specific. The ability to do this comes with experience.
10. Problems and hypotheses need to reflect the multivariate complexity of behavioral science reality.
11. The hypothesis is one of the most powerful tools invented to achieve dependable knowledge. It has the power of prediction. Negative finding for a hypothesis can serve to eliminate one possible explanation and open up other hypotheses and lines of investigation.

Chapter 3

Constructs, Variables and Definitions

SCIENTISTS operate on two levels: theory-hypothesis-construct and observation. More accurately, they shuttle back and forth between these levels. A psychological scientist may say, "Early deprivation produces learning deficiency." This statement is a hypothesis consisting of two concepts, "early deprivation" and "learning deficiency," joined by a relation word, "produces." It is on the theory hypothesis-construct level. Whenever scientists utter relational statements and whenever they use concepts, or constructs as we shall call them, they are operating at this level.

Scientists must also operate at the level of observation. They must gather data to test hypotheses. To do this, they must somehow get from the construct level to the observation level. They cannot simply make observations of "early deprivation" and "learning deficiency." They must so define these constructs that observations are possible. The problem of this chapter is to examine and clarify the nature of scientific concepts or constructs. Also, this chapter will examine and clarify the way in which behavioral scientists get from the construct level to the observation level, how they shuttle from one to the other.

CONCEPTS AND CONSTRUCTS

The terms "concept" and "construct" have similar meanings. Yet there is an important distinction. *A concept* expresses an abstraction formed by generalization from particulars. "Weight" is a concept. It expresses numerous observations of things that are “more or less” and “heavy or light.” “Mass,” “energy,” and “force” are concepts used by physical scientists. They are, of course, much more abstract than concepts such as “weight,” “height,“ and “length.”

A concept of more interest to readers of this book is "achievement." It is an abstraction formed from the observation of certain behaviors of children. These behaviors are associated with the mastery or "learning" of school tasks-reading words, doing arithmetic problems, drawing pictures, and so on. The various observed behaviors are put together and expressed in a word. "Achievement," "intelligence," "aggressiveness," "conformity," and "honesty" are all concepts used to express varieties of human behavior.

A construct is a concept. It has the added meaning, however, of having been deliberately and consciously invented or adopted for a special scientific purpose. "Intelligence" is a concept, an abstraction from the observation of presumably intelligent and nonintelligent behaviors. But as a scientific construct, "intelligence" means both more and less than it may mean as a concept. It means that scientists consciously and systematically use it in two ways. One, it enters into theoretical schemes and is related in various ways to other constructs. We may say, for example, that school achievement is in part a function of intelligence and motivation. Two, "intelligence" is so defined and specified that it can be observed and measured. We can make observations of the intelligence of children by administering an intelligence test to them, or we can ask teachers to tell us the relative degrees of intelligence of their pupils.

VARIABLES

Scientists somewhat loosely call the constructs or properties they study "variables." Examples of important variables in sociology, psychology, political science and education are gender, income, education, social class, organizational productivity, occupational mobility, level of aspiration, verbal aptitude, anxiety, religious affiliation, political preference, political development (of nations), task orientation, racial and ethnic prejudices, conformity, recall memory, recognition memory, achievement. It can be said that a variable is a property that takes on different values. Putting it redundantly, a variable is something that varies. While this way of speaking gives us an intuitive notion of what variables are, we need a more general and yet more precise definition.

A variable is a symbol to which numerals or values are assigned. For instance, x is a variable: it is *a symbol to* which we assign numerical values.

The variable x may take on any justifiable set of values, for example, scores on an intelligence test or an attitude scale. In the case of intelligence we assign to x a set of numerical values yielded by the procedure designated in a specified test of intelligence. This set of values ranges from low to high, from, say, 50 to 150.

A variable, x, however, may have only two values. If gender is the construct under study, then x can be assigned 1 and 0, where 1 standing for one of the genders and 0 standing for the other. It is still a variable. Other examples of two-valued variables are in-out, correct-incorrect, old-young, citizen-noncitizen, middle class-working class, teacher-nonteacher, Republican-Democrat, and so on. Such variables are called dichotomies, dichotomous, or binary variables.

Some of the variables used in behavioral research are true dichotomies, that is, they are characterized by the presence or absence of a property: male-female, home-homeless, employed-unemployed. Some variables are polytomies. A good example is religious preference: Protestant, Catholic, Muslim, Jew, Buddhist, Other. Such dichotomies and polytomies have been called "qualitative variables." The nature of this designation will be discussed later. Most variables, however, are theoretically capable of taking on continuous values. It has been common practice in behavioral research to convert continuous variables to dichotomies or polytomies. For example, intelligence, a continuous variable, has been broken down into high and low intelligence, or into high, medium, and low intelligence. Variables such as anxiety, introversion, and authoritarianism have been treated similarly. While it is not possible to convert a truly dichotomous variable such as gender to a continuous variable, it is always possible to convert a continuous variable to a dichotomy or a polytomy. As we will see later, such conversion can serve a useful conceptual purpose, but is poor practice in the analysis of data because it throws away information.

CONSTITUTIVE AND OPERATIONAL DEFINITIONS OF CONSTRUCTS AND VARIABLES

The distinction made earlier between "concept" and "construct" leads naturally to another important distinction between kinds of definitions of constructs and variables. Words or constructs can be defined in two general ways. First, we can define a word by using other words, which is what a dictionary usually does. We can define "intelligence" by saying it is "operating intellect," "mental acuity," or "the ability to think abstractly." Such definitions use other concepts or conceptual expressions in lieu of the expression being defined.

Second, we can define a word by telling what actions or behaviors it expresses or implies. Defining "intelligence" this way requires that we specify what behaviors of children are "intelligent" and what behaviors are "not intelligent." We may say that a seven-year-old child who successfully reads a story is "intelligent." If the child cannot read the story, we may say the child is "not intelligent." In other words, this kind of definition can be called *a behavioral or observational definition*. Both "other word" and "observational" definitions are used constantly in everyday living.

There is a disturbing looseness about this discussion. Although scientists use the types of definition just described they do so in a more precise manner. We express this usage by defining and explaining Margenau's (1950/1977) distinction between constitutive and operational definitions. *A constitutive definition* defines a construct with other constructs. For instance, we can define "weight" by saying that it is the "heaviness" of objects. Or we can define "anxiety" as "subjectified fear." In both cases we have substituted one concept for another. Some of the constructs of a scientific theory may be defined constitutively. Torgerson (1958/1985), borrowing from Margenau, says that all constructs, in order to be scientifically useful, must possess constitutive meaning. This means that they must be capable of being used in theories.

An operational definition assigns meaning to a construct or a variable by specifying the activities or "operations" necessary to measure it and evaluate the measurement. Alternatively, an operational definition is a specification of the activities of the researcher in measuring a variable or in manipulating it. An operational definition is a sort of manual of instructions to the investigator.

It says, in effect, "Do such-and-such in so-and-so a manner." In short, it defines or gives meaning to a variable by spelling out what the investigator must do to measure it and evaluate that measurement.

Michel (1990) gives an excellent historical account on how operational definitions became popular in the social and behavioral sciences. Michel cites P. W. Bridgeman, a Nobel laureate, for creating the operation definition in 1927. Bridgeman as quoted in Michel (1990, p. 15) states: "In general we mean by any concept nothing more than a set of operations; *the concept is synonymous with the corresponding set of operations*." Each different operation would define a different concept.

A well-known, if extreme, example of an operational definition is: Intelligence (anxiety, achievement, and so forth) is scores on X intelligence test, or intelligence is what X intelligence test measures. Also high scores indicate a greater level of intelligence than lower scores. This definition tells us what to do to measure intelligence. It says nothing about how well intelligence is measured by the specified instrument. (Presumably the adequacy of the test was ascertained prior to the investigator's use of it.). In this usage, an operational definition is an equation where we say, "Let intelligence equal the scores on X test of intelligence and higher scores indicate a higher degree of intelligence than lower scores." We also seem to be saying, "The meaning of intelligence (in this research) is expressed by the scores on X intelligence test."

There are, in general, two kinds of operational definitions: (1) *measured* and (2) ex*perimental.* The definition given above is more closely tied to measured than to experimental definitions. *A measured* operational definition describes how a variable will be measured. For example, achievement may be defined by a standardized achievement test, by a teacher-made achievement test, or by grades. Doctor, Cutris and Isaacs (1994), studying the effects of stress counseling on police officers, operationally defined psychiatric morbidity as scores on the General Health Questionnaire and the number of sick leave days taken. Higher scores and large number of days indicated elevated levels of morbidity. Little, Sterling and Tingstrom (1996) studied the effects of race and geographic origin on attribution. Attribution was operationally defined as a score on the Attributional Style Questionnaire. A study may include the variable *consideration.* It can be defined operationally by listing behaviors of children that are presumably considerate behaviors and then requiring teachers to rate the children on a five-point scale. Such behaviors might be when the children say to each other, "I'm sorry," or "Excuse me." Or when one child yields a toy to another on request (but not on threat of aggression), or when one child helps another with a school task. It can also be defined as counting the number of considerate behaviors. The greater the number, the higher the level of consideration.

An *experimental* operational definition spells out the details (operations) of the investigator's manipulation of a variable. Reinforcement can be operationally defined by giving the details of how subjects are to be reinforced (rewarded) and not reinforced (not rewarded) for specified behaviors. Hom, Berger, Miller, & Belvin (1994) defined reinforcement experimentally in operational terms. In this study, children were assigned to one of four groups. Two of the groups received a cooperative reward condition while the other two groups got an individualistic reward condition. Bahrick (1984) defines long-term memory in terms of at least two processes when in comes to the retention of academically oriented information. One process called "permastore" selectively chooses some information to be stored permanently and is highly resistant to decay (forgetting). The other process seems to select certain information that is apparently less significant and hence is less resistant to forgetting. This definition contains clear implications for experimental manipulation. Strack, Martin, and Stepper (1988) operationally defined smiling as the activation of the muscles associated with the human smile. This was done by having a person hold a pen in a person's mouth in a certain way. This was unobtrusive in that the subjects in the study were not asked to pose in a smiling face. Other examples of both kinds of operational definitions will be given later.

Scientific investigators must sooner or later face the necessity of measuring the variables of the relations they are studying. Sometimes

measurement is easy, sometimes difficult. To measure gender or social class is easy; to measure creativity, conservatism, or organizational effectiveness is difficult. The importance of operational definitions cannot be overemphasized. They are indispensable ingredients of scientific research because they enable researchers to measure variables and because they are bridges between the theory-hypothesis-construct level and the level of observation. There can be no scientific research without observations, and observations are impossible without clear and specific instructions on what and how to observe. Operational definitions are such instructions.

Though indispensable, operational definitions yield only limited meanings of constructs. No operational definition can ever express all of a variable. No operational definition can ever express the rich and diverse aspects of human prejudice, for example. This means that the variables measured by scientists are always limited and specific in meaning. The "creativity" studied by psychologists is not the "creativity" referred to by artists, though there will of course be common elements. A person who thinks of a creative solution for a math problem may show little creativity as a poet (Barron and Harrington, 1981). Some psychologists have operationally defined creativity as performance on the Torrance Test of Creative Thinking (Torrance, 1982). Children who score high on this test are more likely to make creative achievements as adults.

Some scientists say that such limited operational meanings are the only meanings that "mean" anything, that all other definitions are metaphysical nonsense. They say that discussions of anxiety are metaphysical nonsense, unless adequate operational definitions of anxiety are available and are used. This view is extreme, though it has healthy aspects. To insist that every term we use in scientific discourse be operationally defined would be too narrowing, too restrictive, and, as we shall see, scientifically unsound. Northrop (1947/1983 p. 130) says, for example, "The importance of operational definitions is that they make verification possible and enrich meaning. They do not, however, exhaust scientific meaning." Margenau (1950/1977, p. 232) makes the same point in his extended discussion of scientific constructs.

Despite the dangers of extreme operationism, it can be safely said that operationism has been and still is a healthy influence. As Skinner (1945, p. 274) puts it, "The operational attitude, in spite of its shortcomings, is a good thing in any science but especially in psychology because of the presence there of a vast vocabulary of ancient and nonscientific origin." When the terms used in education are considered, clearly education, too, has a vast vocabulary of ancient and nonscientific terms. Consider these: the whole child, horizontal and vertical enrichment, meeting the needs of the learner, core curriculum, emotional adjustment, and curricular enrichment. This is also true in the field of geriatric nursing. Here nurses deal with such terms as the aging process, self-image, attention-span and unilateral neglect (Eliopoulos, 1993; Smeltzer and Bare, 1992).

To clarify constitutive and operational definitions (and theory too) look at Figure 3.1, which has been adapted after Margenau (1950/1977) and Torgerson (1958/1985). The diagram is supposed to illustrate a well-developed theory. The single lines represent theoretical connections or relations between constructs. These constructs, labeled with lower-case letters, are defined constitutively; that is, c_4 is defined somehow by c_3, or vice versa. The double lines represent operational definitions. The C constructs are directly linked to observable data; they are indispensable links to empirical reality. However, not all constructs in a scientific theory are defined operationally. Indeed, it is a rather thin theory that has all its constructs so defined.

Let us build a "small theory" of underachievement to illustrate these notions. Suppose an investigator believes that underachievement is, in part, a function of pupils' self-concepts. The investigator believes that pupils who perceive themselves as inadequate and have negative self-percepts, also tend to achieve less than their potential capacity and aptitude indicate they should achieve. It follows that ego-needs (which we will not define here) and motivation for achievement (call this n-ach, or need for achievement) are tied to underachievement. Naturally, the investigator is also aware of the

relation between aptitude and intelligence and achievement in general. A diagram to illustrate this "theory" might look like Figure 3.2.

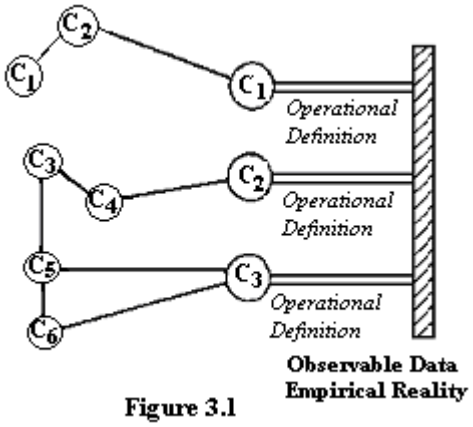


Figure 3.1

The investigator has *no direct* measure of self-concept, but assumes that inferences can be drawn about an individual's self concept from a figure-drawing test. Self-concept is operationally defined, then, as certain responses to the figure-drawing test. This is probably the most common method of measuring psychological (and educational) constructs. The heavy single line between c_1 and C_1 indicates the relatively direct nature of the presumed relation between self-concept and the test. (The double line between C_1 and the level of observation indicates an operational definition, as it did in Figure 3.1.)

Similarly, the construct achievement (c_4) is operationally defined as the discrepancy between measured achievement (C_2) and measured aptitude (c_5). In this model the investigator has no direct measure of achievement motivation, no operational definition of it. In another study, naturally, the investigator may specifically hypothesize a relation between achievement and achievement motivation, in which case one will try to define achievement motivation operationally.

A single solid line between concepts, for example, the one between the construct achievement (c_4) and achievement test (C_2), indicates a relatively well established relation between postulated achievement and what standard achievement tests measure. The single solid lines between C_1 and C_2 and between C_2 and C_3 indicate obtained relations between the test scores of these measures. (The lines between C_1 and C_2 and between C_2 and C_3 are labeled *r* for "relation," or "coefficient of correlation.")

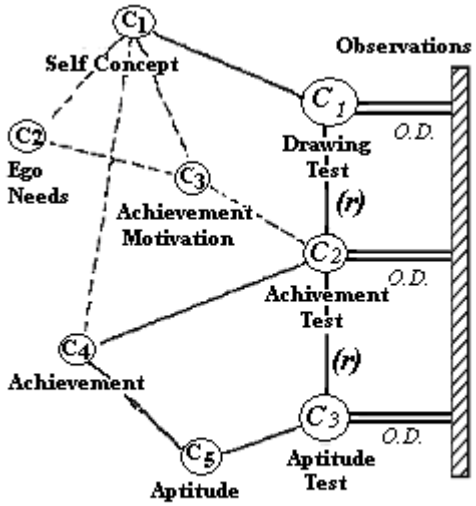


Figure 3.2

The broken single lines indicate postulated relations between constructs that are not relatively well established. A good example of this is the postulated relation between self-concept and achievement motivation. One of the aims of science is to make these broken lines solid lines by bridging the operational definition-measurement gap. In this case, it is quite conceivable that both self-concept and achievement motivation can be operationally defined and directly measured.

In essence, this is the way the behavioral scientist operates. The scientist shuttles back and forth between the level of theory-constructs and the level of observation. This is done by operationally defining the variables of the theory that are amenable to such definition. Then the relations are estimated between the operationally defined and measured variables. From these estimated relations the scientist makes inferences as to the relations between the constructs. In the above example, the behavioral scientist calculates the relation between C_1 (figure-drawing test) and C_2 (achievement test). If the relation is established on this observational level, the scientist infers that a relation exists between c_1 (self-concept) and c_4 (achievement).

TYPES OF VARIABLES

Independent and Dependent Variables

With definitional background behind us, we return to variables. Variables can be classified in several ways. In this book three kinds of variables are very important and will be emphasized: (1) independent and dependent variables, (2) active and attribute variables and (3) continuous and categorical variables.

The most useful way to categorize variables is as independent and dependent. This categorization is highly useful because of its general applicability, simplicity, and special importance in conceptualizing and designing research and in communicating the results of research. *An independent variable is* the *presumed* cause of the *dependent variable,* the *presumed* effect. The independent variable is the antecedent; the dependent is the consequent. Since one of the goals of science is to find relations between different phenomena looking at the relation between independent and dependent variables does this. It is the independent variable that is assumed to influence the dependent variable. In some studies, the independent variable "causes" changes in the dependent variable. When we say: If A, then B, we have the conditional conjunction of an independent variable (A) and a dependent variable (B).

The terms "independent variable" and "dependent variable" come from mathematics, where X is the independent and Y the dependent variable. This is probably the best way to think of independent and dependent variables because there is no need to use the touchy word "cause" and related words, and because such use of symbols applies to most research situations. There is no theoretical restriction on numbers of X's and Y's. When, later, we consider multivariate thinking and analysis, we will deal with several independent and several dependent variables.

In experiments the independent variable is the variable manipulated by the experimenter. Changes in the values or levels of the independent variable produces changes in the dependent variable. When educational investigators studied the effects of different teaching methods on math test performance, they would vary the method of teaching. In one condition they may have lecture only and in the other it would be lecture plus video. Teaching method is the independent variable. The outcome variable, test score, is the dependent variable.

The assignment of participants to different groups based on the existence of some characteristic is an example of where the researcher was not able to manipulate the independent variable. The values of the independent variable in this situation pre-exist. The participant either has the characteristic or not. Here, there is no possibility of experimental manipulation, but the variable is considered to "logically" to have some effect on a dependent variable. Subject characteristic variables make up most of these types of independent variables. One of the more common independent variables of this kind is gender (Female and Male). So if a researcher wanted to determine if females and males differ on math skills, a math test would be given to representatives of both groups and then the test scores are compared. The math test would be the dependent variable. A general rule is that when the researcher manipulates a variable or assigns participants to groups according to some characteristic, that variable is the independent variable. Table 3.1 gives a comparison between the two types of independent variable and its relation to the dependent variable. The independent variable must have at least two levels or values. Notice in Table 3.1 that both situations have two levels for the independent variable.

Table 3.1. Relation of manipulated & Non-manipulated IVs to the dependent variable.

Levels of Independent Variable

Teaching Method		Gender	
Lecture Only	Lecture plus Video	Female	Male
Math Test Scores	Math Test Scores	Math Test Scores	Math Test Scores

Dependent variables

The dependent variable, of course, is the variable predicted to, whereas the independent variable is prdicted *from.* The dependent variable, Y, is the presumed effect, which varies concomitantly with changes or variation in the independent variable, X. It is the variable that is observed for variation as a presumed result of variation in the independent variable. The dependent variable is the outcome measure that the researcher uses to determine if changes in the independent variable had an effect. In predicting from X to Y, we can take any value of X we wish, whereas the value of Y we predict to is "dependent on" the value of X we have selected. The dependent variable is ordinarily the condition we are trying to explain. The most common dependent variable in education, for instance, is achievement or "learning." We want to account for or explain achievement. In doing so we have a large number of possible X's or independent variables to choose from.

When the relation between intelligence and school achievement is studied, intelligence is the independent and achievement is the dependent variable. (Is it conceivable that it might be the other way around?) Other independent variables that can be studied in relation to the dependent variable, achievement, are social class, methods of teaching, personality types, types of motivation (reward and punishment), attitudes toward school, class atmosphere, and so on. When the presumed determinants of delinquency are studied, such determinants as slum conditions, broken homes, lack of parental love, and the like, are independent variables and, naturally, delinquency (more accurately, delinquent behavior) is the dependent variable. In the frustration-aggression hypothesis mentioned earlier, frustration is the independent variable and aggression the dependent variable. Sometimes a phenomenon is studied by itself, and either an independent or a dependent variable is implied. This is the case when teacher behaviors and characteristics are studied. The usual implied dependent variable is achievement or child behavior. Teacher behavior can of course be a dependent variable. Consider an example in nursing science. When cognitive and functional measures of Alzheimer patients are compared between traditional nursing homes and special care units (SCU), the independent variable is the place of care. The dependent variables are the cognitive and functional measures (Swanson, Maas, and Buckwalter, 1994).

The relation between an independent variable and a dependent variable can perhaps be more clearly understood if we lay out two axes at right angles to each other. One axis represents the independent variable and the other axis represents the dependent variable. (When two axes are at right angles to each other, they are called *orthogonal* axes.) Following mathematical custom, X, the independent variable, is the horizontal axis and Y, the dependent variable, the vertical axis. (X is called the *abscissa* and Y the *ordinate.)* X values are laid out on the X-axis and Y values on the Y-axis.

A very common and useful way to "see" and interpret a relation is to plot the pairs of XY values, using the X and Y axes as a frame of reference. Let us suppose, in a study of child development, that we have two sets of measures. The X measures chronological age and the Y measures reading age. *Reading age* is a so-called growth age. Seriatim measurements of individuals growths-in height, weight intelligence, and so forth-are expressed as the average chronological age at which they appear in the standard population.

X: Chronological Age (in Months)	Y: Reading Age (in Months)
72	48
84	62
96	69
108	71
120	100
132	112

These measures are plotted in Figure 3.3.

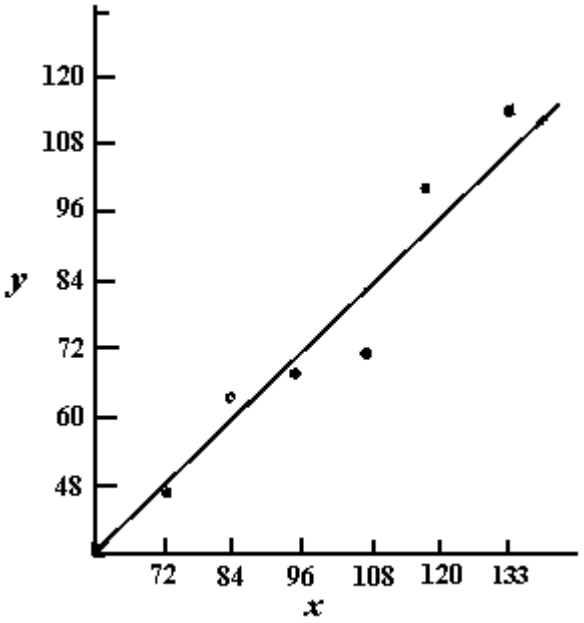


Figure 3.3

The relation between chronological age (CA) and reading age (RA) can now be "seen" and roughly approximated. Note that there is a pronounced tendency, as might be expected for more advanced CA to be associated with higher RA, medium CA with medium RA, and less advanced CA with lower RA. In other words, the relation between the independent and dependent variables, in this case between CA and RA, can be seen from a graph such as this. A straight line has been drawn in to "show" the relation. It is a rough average of all the points of the plot. Note that if one has knowledge of independent variable measures and a relation such as that shown in Figure 3.3, one can predict with considerable accuracy the dependent variable measures. Plots like this can of course be used with any independent and dependent variable measures.

The student should be alert to the possibility of a variable being an independent variable in one study and a dependent variable in another, or even both in the same study. An example is job satisfaction. A majority of the studies involving job satisfaction use it as a dependent variable. Day and Schoenrade (1997) showed the effect of sexual orientation on work attitudes. One of those work attitudes was job satisfaction. Likewise Hodson (1989) studies gender differences in job satisfaction. Scott, Moore and Miceli (1997) found job satisfaction linked to the behavior patterns of workaholics. There have been studies where job satisfaction was used as an independent variable. Meiksins and Watson (1989) showed how much job satisfaction influenced the professional autonomy of engineers. Studies by Somers (1996), Francis-Felsen, Coward, Hogan and Duncan (1996) and Hutchinson and Turner (1988) examined job satisfaction's effect on nursing personnel turnover.

Another example is anxiety. Anxiety has been studied as an independent variable affecting the dependent variable achievement. Oldani (1997) found mother's anxiety during pregnancy influenced the achievement (measured as success in the music industry) of the offspring. Capaldi, Crosby and Stoolmiller (1996) used the anxiety levels of teenage boys to predict the timing of their first sexual intercourse. Onwuegbuzie and Seaman (1995) studied the effects of test anxiety on test performance in a statistics course. It can also readily be conceived and used as a dependent variable. For example, it could be used to study the difference between types of culture, socioeconomic status and gender (see Guida and Ludlow, 1989; Murphy, Olivier, Monson, and Sobol, 1991). In other words, the independent and dependent variable classification is really a classification of uses of variables rather than a distinction between different kinds of variables.

Active and Attribute Variables

A classification that will be useful in our later study of research design is based on the distinction between experimental and measured variables. It is important when planning and executing research to distinguish between these two types of variables. Manipulated variables will be called *active* variables; measured variables will be called *attribute* variables. For example, Colwell, Foreman and Trotter (1993) compared two methods of treating pressure ulcers of bed-ridden patients. The dependent variables were efficacy and cost-effectiveness. The two methods were

moist gauze dressing and hydrocolloid wafer dressing. The researchers had control over which patients got which type of treatment. As such, the treatment or independent variable was an active or manipulated variable.

Any variable that is manipulated, then, is an active variable. "Manipulation" means, essentially, doing different things to different groups of subjects, as we will see clearly in a later chapter where we discuss in depth the differences between experimental and nonexperimental research. When a researcher does one thing to one group, for example, positively reinforces a certain kind of behavior, and does something different to another group or has the two groups follow different instructions, this is manipulation. When one uses different methods of teaching, or rewards the subjects of one group and punishes those of another, or creates anxiety through worrisome instructions, one *is actively* manipulating the variables methods, reinforcement, and anxiety.

Another related classification, used mainly by psychologists, is *stimulus* and *response* variables. *A stimulus variable* is any condition or manipulation by the experimenter of the environment that evokes a response in an organism. *A response variable* is any kind of behavior of the organism. The assumption is made that for any kind of behavior there is always a stimulus. Thus the organism's behavior is a response. This classification is reflected in the well-known equation: $R = f(O, S)$, which is read: "Responses are a function of the organism and stimuli," or "Response variables are a function of organismic variables and stimulus variables."

Variables that cannot be manipulated are attribute variables or subject-characteristic variables. It is impossible, or at least difficult, to manipulate many variables. All variables that are human characteristics such as intelligence, aptitude, gender, socioeconomic status, conservatism, field dependence, need for achievement, and attitudes are attribute variables. Subjects come to our studies with these variables (attributes) ready-made or pre-existing. Early environment, heredity, and other circumstances have made individuals what they are. Such variables are also called *organismic* variables. Any property of an individual, any characteristic or attribute, is an organismic variable. It is part of the organism, so to speak. In other words, organismic variables are those characteristics that individuals have in varying degrees when they come to the research situation. The term *individual differences* imply organismic variables. One of the more common attribute variables in the social and behavioral sciences is gender: female-male. Studies designed to compare gender differences involve an attribute variable. Take for example the study by de Weerth and Kalma (1993). These researchers compared females to males on their response to spousal or partner infidelity. Gender is the attribute variable here. Gender is not a manipulated variable There are studies where a test score or a collection of test scores were used to divide a group of people into two or more groups. In this case the group differences are reflected in an attribute variable. For example, the study by Hart, Forth and Hare (1990) administered a psychopathology test to male prison inmates. Based on their scores inmates were assigned to one of three groups: low, medium and high. They were then compared on their score on a battery of neuropsychological tests. The level of psychopathology pre-exists and is not manipulated by the researcher. If an inmate scored high, he was placed in the high group. Hence psychopathology is an attribute variable in this study. There are some studies where the independent variable could have been manipulated. However, for logistic or legal reasons they were not. An example of where the independent variable could have been manipulated but was not is the study by Swanson, Maas and Buckwalter (1994). These researchers compared different care facilities effect on cognitive and functional measures of Alzheimer patients. The attribute variable was the type of facility. The researchers were not allowed to place patients into the two different care facilities (traditional nursing home versus special care unit). The researchers were forced to study the subjects after they had been assigned to a care facility. Hence the independent variable can be thought of as a non-manipulated one. The researcher inherited groups that were intact.

The word "attribute," moreover, is accurate enough when used with inanimate objects or referents. Organizations, institutions, groups, populations, homes, and geographical areas have attributes. Organizations are variably productive; institutions become outmoded; groups differ in cohesiveness; geographical areas vary widely in resources.

This active-attribute distinction is general, flexible, and useful. We will see that some variables are by their very nature always attributes, but other variables that are attributes can also be active. This latter characteristic makes it possible to investigate the "same" relations in different ways. A good example, again, is the variable anxiety. We can measure the anxiety of subjects. Anxiety is in this case obviously an attribute variable. However, we can manipulate anxiety, too. We can induce different degrees of anxiety, for example, by telling the subjects of one experimental group that the task they are about to do is difficult, that their intelligence is being measured, and that their futures depend on the scores they get. The subjects of another experimental group are told to do their best but to relax. They are told the outcome is not too important and will have no influence on their futures. Actually, we cannot assume that the measured (attribute) and the manipulated (active) "anxieties" are the same. We may assume that both are "anxiety" in a broad sense, but they are certainly not the same.

CONTINUOUS AND CATEGORICAL VARIABLES

A distinction especially useful in the planning of research and the analysis of data between continuous and categorical variables has already been introduced. Its later importance, however, justifies more extended consideration.

A continuous variable is capable of taking on an ordered set of values within a certain range. This definition means, first, that the values of a continuous variable reflect at least a rank order, a larger value of the variable meaning more of the property in question than a smaller value. The values yielded by a scale to measure dependency, for instance, express differing amounts of dependency from high through medium to low. Second, continuous measures in actual use are contained in a range, and each individual obtains a "score" within the range. A scale to measure dependency may have the range 1 through 7. Most scales in use in the behavioral sciences also have a third characteristic: there is a theoretically infinite set of values within the range. (Rank-order scales are somewhat different` they will be discussed later in the book.) That is, a particular individual's score may be 4.72 rather than simply 4 or 5.

Categorical variables, as we will call them, belong to a kind of measurement called nominal. In nominal measurement, there are two or more subsets of the set of objects being measured. Individuals are categorized by their possession of the characteristic that defines any subset. "To categorize" means to assign an object to a subclass or subset of a class or set on the basis of the object's having or not having the characteristic that defines the subset. The individual being categorized either has the defining property or does not have it; it is an all-or-none kind of thing. The simplest examples are dichotomous categorical variables: female-male, Republican-Democrat, right-wrong. Polytomies, variables with more than two subsets or partitions, are fairly common, especially in sociology and economics: religious preference, education (usually), nationality, occupational choice, and so on.

Categorical variables and nominal measurement have simple requirements: all the members of a subset are considered the same and all are assigned the same name (nominal) and the same numeral. If the variable is religious preference, for instance, all Protestants are the same, all Catholics are the same, and all "others" are the same. If an individual is a Catholic (operationally defined in a suitable way) that person is assigned to the category "Catholic" and also assigned a "1" in that category. In brief, that person is counted as a "Catholic." Categorical variables are "democratic. There is no rank order or greater than and less-than among the categories, and all members of a category have the same value.

The expression "qualitative variables" has sometimes been applied to categorical variables, especially to dichotomies, probably in contrast to "quantitative variables" (our continuous variables).

Such usage reflects a somewhat distorted notion of what variables are. They are always quantifiable, or they are not variables. If x has only two subsets and can take on only two values, 1 and 0, these are still values, and the variable varies. If x is a polytomy, like political affiliation, we quantify again by assigning integer values to individuals. If an individual, say, is a Democrat, then put that person in the Democrat subset. That individual is assigned a *1*. All individuals in the Democrat subset would be assigned a value of 1. It is extremely important to understand this because, for one thing, it is the basis of quantifying many variables, even experimental treatments, for complex analysis. In multiple regression analysis, as we will see later in the book, all variables, continuous and categorical, are entered as variables into the analysis. Earlier, the example of gender was given, 1 being assigned to one gender and 0 to the other. We can set up a column of 1's and 0's just as we would set up a column of dependency scores. The column of 1's and 0's is the quantification of the variable gender. There is no mystery here. Such variables have been called "dummy variables." Since they are highly useful and powerful, even indispensable, in modern research data analysis they need to be clearly understood. A deeper explanation of this can be found in Kerlinger and Pedhazur (1973). The method is easily extended to polytomies. *A polytomy is* a division of the members of a group into three or more subdivisions.

CONSTRUCTS, OBSERVABLES, AND LATENT VARIABLES

In much of the previous discussion of this chapter it has been implied, though not explicitly stated, that there is a sharp difference between constructs and observed variables. Moreover, we can say that constructs are nonobservables, and variables, when operationally defined, are observables. The distinction is important, because if we are not always keenly aware of the level of discourse we are on when talking about variables, we can hardly be clear about what we are doing.

An important and fruitful expression, which we will encounter and use a good deal later in the book, is "latent variable." A latent variable is an unobserved "entity" presumed to underlie observed variables. The best-known example of an important latent variable is "intelligence." We can say that three ability tests, verbal, numerical, and spatial, are positively and substantially related. This means, for the most part, that people high on one tend to be high on the others; similarly, persons low on one tend to be low on the others. We believe that something is common to the three tests or observed variables and name this something "intelligence." It is a latent variable.

We have encountered many examples of latent variables in previous pages: achievement, creativity, social class, job satisfaction, religious preference, and so on. Indeed, whenever we utter the names of phenomena on which people or objects vary, we are talking about latent variables. In science our real interest is more in the relations among latent variables than it is in the relations among observed variables because we seek to explain phenomena and their relations. When we enunciate a theory, we enunciate in part systematic relations among latent variables. We are not too interested in the relation between observed frustrated behaviors and observed aggressive behaviors, for example, though we must of course work with them at the empirical level. We are really interested in the relation between the latent variable frustration and the latent variable aggression.

We must be cautious, however, when dealing with nonobservables. Scientists, using such terms as "hostility," "anxiety," and "learning," are aware that they are talking about invented constructs. The "reality" of these constructs is inferred from behavior. If they want to study the effects of different kinds of motivation, they must know that "motivation" is a latent variable, a construct invented to account for presumably "motivated" behavior. They must know that its "reality" is only a postulated reality. They can only judge that youngsters are motivated or not motivated by observing their behavior. Still, in order to study motivation, they must measure it or manipulate it. But they cannot measure it directly because it is an "in-the-head" variable, an unobservable entity, a latent variable, in short. The construct was invented for "something" *presumed to be* inside individuals, "something" prompting them to behave in such-and-such manners. This means that researchers

must always measure presumed indicators of motivation and not motivation itself. They must, in other words, always measure some kind of behavior, be it marks on paper, spoken words, or meaningful gestures, and then make inferences about presumed characteristics-or latent variables.

Other terms have been used to express more or less the same ideas. For example, Tolman (1951 pp. 115-129.) calls constructs intervening variables. *Intervening variable is* a term invented to account for internal, unobservable psychological processes that in turn account for behavior. An intervening variable is an "in-the-head" variable. It cannot be seen, heard, or felt. It is inferred from behavior. "Hostility" is inferred from presumably hostile or aggressive acts. "Anxiety" is inferred from test scores, skin responses, heart beat, and certain experimental manipulations. Another term is "hypothetical construct." Since this expression means much the same as latent variable with somewhat less generality, we need not pause over it. We should mention, however, that "latent variable" appears to be a more general and applicable expression than "intervening variable" and "hypothetical construct." This is because it can be used for virtually any phenomena that presumably influence or are influenced by other phenomena. In other words, "latent variable" can be used with psychological, sociological, and other phenomena. "Latent variable" seems to be the preferable term because of its generality. Also, because it is now possible in the analysis of covariance structures approach to assess the effects of latent variables on each other and on so-called manifest or observed variables. This rather abstract discussion will later be made more concrete and, it is hoped, meaningful. We will then see that the idea of latent variables and the relations between them is an extremely important, fruitful, and useful one that is helping to change fundamental approaches to research problems.

EXAMPLES OF VARIABLES AND OPERATIONAL DEFINITIONS

A number of constructs and operational definitions have already been given. To illustrate and clarify the preceding discussion, especially that in which the distinction was made between experimental and measured variables and between constructs and operationally defined variables, several examples of constructs or variables and operational definitions are given below. If a definition is experimental, it is labeled (E); if it is measured, it is labeled (M).

Operational definitions differ in degree of specificity. Some are quite closely tied to observations. "Test" definitions, like "Intelligence is defined as a score on X intelligence test " are very specific. A definition like "Frustration is prevention from reaching a goal" is more general and requires further specification to be measurable.

Social Class ". . . two or more orders of people who are believed to be, and are accordingly ranked by the members of a community, in socially superior and inferior positions." (M) Warner and Lunt, 1941 p. 82). To be operational, this definition has to be specified by questions aimed at people's beliefs about other people's positions. This is a subjective definition of social class. Social class, or social status, is also defined more objectively by using such indices as occupation, income, and education, or by combinations of such indices. For example, ". . . we converted information about the education, occupation and income of the parents of the NLSY youths into an index of socioeconomic status (SES) in which the highest scores indicate advanced education, affluence and prestigious occupations. Lowest scores indicate poverty, meager education and the most menial jobs." (M) (Herrnstein and Murray, 1996, p. 131).

Achievement (School, Arithmetic, and Spelling) Achievement is customarily defined operationally by citing a standardized test of achievement (for example, Iowa Tests of Basic Skills, Elementary or the Achievement Test of the Kaufman Assessment Battery for Children (K-ABC), by grade-point averages, or by teacher judgments. "Student achievement was measured by the combined test scores of reading and mathematics." (M) (Peng and Wright, 1994). Occasionally achievement is in the form of a performance test. Silverman (1993) examined students on two skills in volleyball: The serve test and the forearm passing test. In the serve test, students received a score between 0 and 4 depending on where the served ball dropped. The forearm passing test involved bouncing the ball off

of one's forearm. The criteria used was to count the number of times the student can pass the ball above a 8 foot line against the wall within a 1 minute period. (M) Also used in some educational studies is an operational definition of the concept *student achievement perception*. Here the students are asked to evaluate themselves. The question used by Shoffner (1990) was "What kind of student do you think you are?" The response choices available were "A student," "B student," and "C student." (M)

Achievement (Academic Performance) "As a result, grades for all students in all sections were obtained and used to determine the section-rank for each student participating in the study. Section percentile rank was computed for each of these students and was used as the dependent measure of achievement in the final data analysis." (M) (Strom, Hocevar, and Zimmer, 1990).

Hom, Berger, et al (1994) operationally define intrinsic Motivation as "The cumulative amount of time that each student played with the pattern blocks with the reward system absent." (M).

Popularity. Popularity is often defined operationally by the number of sociometric choices an individual receives from other individuals (in his class, play group, and so on). Individuals are asked: "With whom would you like to work?," "With whom would you like to play?," and the like. Each individual is required to choose one, two, or more individuals from his group on the basis of such criterion questions. (M)

Task Involvement ". . . each child's behavior during a lesson was coded every 6 sec as being appropriately involved, or deviant. The task involvement scores for a lesson was the percentage of 6-sec units in which the children were coded as appropriately involved." (M) (Kounin and Doyle, 1975).

Reinforcement. Reinforcement definitions come in a number of forms. Most of them involve, in one way or another, the principle of reward. However, both positive and negative reinforcement may be used. The author gives specific experimental definitions of "reinforcement. " For example,

> In the second 10 minutes, every opinion statement S made was recorded by E and reinforced. For two groups, E agreed with every opinion statement by saying: "Yes, you're right," "That's so," or the like, or by nodding and smiling affirmation if he could not interrupt. (E) the model and the child were administered alternately 12 different sets of story items. To each of the 12 items, the model consistently expressed judgmental responses in opposition to the child's moral orientation. . . and the experimenter reinforced the models behavior with verbal approval responses such as "Very good," "That's fine," and "That's good." The child was similarly reinforced whenever he adopted the models class of moral judgments in response to his own set of items. (E) [This is called "social reinforcement."] (Bandura and MacDonald, 1994).

Another example: The teacher gives verbal praise each time the child exhibits the target behavior. The target behaviors are attending to instruction, school work, and responding aloud. The recording is done every 15 seconds. (E) (Martens, Hiralall, and Bradley, 1997).

Attitudes Toward AIDS is defined by an 18-item scale. Each item consisted of a Likert-type format reflecting different attitudes toward AIDS patients. Some sample items are "People with AIDS should not be permitted to use public toilets," and "There should be mandatory testing of all Americans for AIDS." (M) (Lester, 1989).

Comrey (1993) defines borderline Personality as having low scores on three scales of the Comrey Personality Scales. The three scales are Trust versus Defensiveness, Social Conformity versus Rebelliousness and Emotional Stability versus Neuroticism.

Employee delinquency is operationally defined as a combination of three variables. The variables are the number of chargeable accidents, the number of warning letters and the number of suspensions. (M) (Hogan and Hogan, 1989).

Religiosity is defined as a score on the Francis Scale of Attitudes toward Christianity. This scale consists of 24 items. Each item has a Likert-type response scale. Sample items include: "Saying my prayers helps me a lot." and "God helps me to lead a better life." (M) (Gillings and Joseph, 1996). Religiosity should not be confused with religious preference. Here religiosity refers to the strength of devotion to one's chosen religion.

Self Esteem. Self-Esteem is a manipulated independent variable in the study by Steele, Spencer and Lynch (1993). Here subjects are given a self-esteem test, but when they are given feedback, the information on the official looking feedback report was bogus. Subjects of the same measured level of self-esteem are divided into 3 feedback groups: positive, negative and none. In the positive feedback condition (positive self-esteem), subjects are described with statements such as "clear thinking." Those in the negative group (negative self-esteem) are given adjectives like "passive in action." The no feedback group are told that their personality profile (self-esteem) were not ready due to a backlog in scoring and interpretation. (E). Most studies on self-esteem use a measured operational definition. In the above example, Steele, Spencer and Lynch used the Janis-Field Feelings of Inadequacy Self-Esteem scale. (M) In another example, Luhtanen and Crocker (1992) defines collective self-esteem as a score on a scale containing 16 Likert-type items. These items ask respondents to think about a variety of social groups and membership such as gender, religion, race and ethnicity. (M).

Race. Race is usually a measured variable. However, in a study by Annis and Corenblum, (1986), 83 Canadian Indian kindergartners and first graders were asked questions on racial preferences and self-identity by either a White or Indian experimenter. (E). The interest here was on whether the race of the experimenter influenced responses.

Loneliness. One definition of this is a score on the UCLA Loneliness Scale. This scale includes items such as "No one really knows me well," or "I lack companionship." There is also the Loneliness Deprivation Scale that has items such as "I experience a sense of emptiness," or "There is no one who shows a particular interest in me." (M) (Oshagan and Allen, 1992).

Halo. There have been many operational definitions of the halo effect. Balzer and Sulsky (1992) summarize them. They found 108 definitions that fit into 6 categories. One of the definitions states that halo is " . . . the average within-ratee variance or standard deviation of ratings." Another would be "comparing obtained ratings with true ratings provided by expert raters." (M).

Memory: Recall and Recognition ". . . *recall,* is to ask the subject to recite what he remembers of the items shown him, giving him a point for each item that matches one on the stimulus list. (M) (Norman, 1976 p. 97.). "The *recognition* test consisted of 62 sentences presented to all subjects. . . . . subjects were instructed to rate each sentence on their degree of confidence that the sentence had been presented in the acquisition set." (M) (Richter and Seay, 1987).

Social Skills. This can be operationally defined as a score on the Social Skills Rating Scale (Gresham and Elliot, 1990). There is the possibility of input from the student, parent and teacher. Social behaviors are rated in terms of frequency of occurrence and also on the level of importance. Some social skill items include: "Gets along with people who are different, (Teacher)," "Volunteers to help family members with tasks, (Parent)" and "I politely question rules that may be unfair. (Student)" (M).

Ingratiation. One of many impression management techniques (see Orpen, 1996 and Gordon, 1996). Ingratiation is operationally defined as a score on the Kumar and Beyerlein (1991) scale. This scale consisted of 25 Likert-type items and designed to measure the frequency that subordinates in a superior-subordinate relationship use ingratiatory tactics. (M). Strutton, Pelton and Lumpkin (1995) modified the Kumar-Beyerlein scale. Instead of measuring ingratiation between and employee and employer-supervisor, it measured ingratiation behavior between a salesperson and a customer. (M).

Feminism. This is defined by a score on the Attitudes Toward Women Questionnaire. This instrument consists of 18 statements to which the respondent registers agreement on a 5-point scale. Items include 'Men have held power for too long'; 'Beauty contests are degrading to women,' 'Children of working mothers are bound to suffer.' " (Wilson and Reading, 1989).

Values. "Rank the ten goals in the order *of their importance* to you. (1) financial success; (2) being liked; (3) success in family life; (4) being intellectually capable; (5) living by religious principles; (6) helping others; (7) being normal,

well-adjusted; (8) cooperating with others; (9) doing a thorough job; (10) occupational success." (M) (Newcomb, 1978).

Democracy (Political Democracy) "The index [of political democracy] consists of three indicators of popular sovereignty and three of political liberties. The three measures of popular sovereignty are: (1) fairness of elections, (2) effective executive selection, and (3) legislative selection. The indicators of political liberties are: (4) freedom of the press, (5) freedom of group opposition, and (6) government sanctions." (M). Bollen (1979) gives operational details of the six social indicators in an appendix (pp. 585-586). This is a particularly good example of the operational definition of a complex concept. Moreover, it is an excellent description of the ingredients of democracy.

The benefits of operational thinking have been great. Indeed, operationism has been and is one of the most significant and important movements of our times. Extreme operationism, of course, can be dangerous because it clouds recognition of the importance of constructs and constitutive definitions in behavioral science, and because it can also restrict research to trivial problems. There can be little doubt, however, that it is a healthy influence. It is the indispensable key to achieving objectivity (without which there is no science) because its demand that observations must be public and replicable helps to put research activities outside of and apart from researchers and their predilections. And, as Underwood (1957 p. 53) has said in his classical text on psychological research:

> . . . I would say that operational thinking makes better scientists. The operationist is forced to remove the fuzz from his empirical concepts. . . . . operationism facilitates communication among scientists because the meaning of concepts so defined is not easily subject to misinterpretation.

Study Suggestions

1. Write operational definitions for five or six of the following constructs. When possible, write two such definitions: an experimental one and a measured one.

reinforcement	punitiveness
achievement	reading ability
underachievement	needs
leadership	interests
transfer of training	delinquency
level of aspiration	need for affiliation
organizational conflict	conformity
political preference	marital satisfaction

Some of these concepts or variables-for example, needs and transfer of training-may be difficult to define operationally. Why?

2. Can any of the variables in 1, above, be both independent and dependent variables? Which ones?

3. It is instructive and broadening for specialists to read outside their fields. This is particularly true for students of behavioral research. It is suggested that the student of a particular field read two or three research studies in one of the best journals of another field. If you are in psychology, read a sociology journal, say the *American Sociological Review.* If you are in education or sociology, read a psychology journal say the *Journal! of Personality and Social Psychology* or the *Journal of Experimental Psychology.* Students not in education can sample the *Journal of Educational* Psy*chology* or the *American Educational Research Journal.* When you read, jot down the names of the variables and compare them to the variables in your own field. Are they primarily active or attribute variables? Note, for instance, that psychology's variables are more "active" than sociology's. What implications do the variables of a field have for its research?

4. Reading the following articles is useful in learning and developing operational definitions.

Kinnier, R. T. (1995). A reconceptualization of values clarification: Values conflict resolution. *Journal of Counseling and Development, 74(1),* 18-24.

Lego, S. (1988). Multiple disorder: An interpersonal approach to etiology, treatment and nursing care. *Archives of Psychiatric Nursing, 2(4),* 231-235.

Lobel, M. (1994). Conceptualizations, measurement, and effects of prenatal maternal stress on birth

outcomes. *Journal of Behavioral Medicine, 17(3*), 225-272.

Navathe, P. D. & Singh, B. (1994). An operational definition for spatial disorientation. *Aviation, Space & Environmental Medicine, 65(12),* 1153-1155.

Sun, K.(1995). The definition of race. *American Psychologist, 50(1),* 43-44.

Talaga, J. A. & Beehr, T. A. (1995). Are there gender differences in predicting retirement decisions? *Journal of Applied Psychology, 80(1),* 16-28.

Woods, D. W., Miltenberger, R. G. & Flach, A. D. (1996). Habits, tics, and stuttering: Prevalence and relation to anxiety an somatic awareness. *Behavior Modification, 20(2),* 216-225.

Chapter Summary

1. A *concept* is an expression of an abstraction formed from generalization of particulars, e.g., weight. This expression is from observations of certain behaviors or actions.
2. A *construct* is a concept that has been formulated so that it can be used in science. It is used in theoretical schemes. It is defined so that it can be observed and measured.
3. A *variable* is defined as a property that can take on different values. It is a symbol to which values are assigned.
4. Constructs and words can be defined by
   a) other words or concepts
   b) description of an implicit or explicit action or behavior
5. A *constitutive definition* is one where constructs are defined by other constructs.
6. An *operational definition* is one where meaning is assigned by specifying the activities or operations necessary to measure and evaluate the construct. Operational definitions can give only limited meaning of constructs. They cannot completely describe a construct or variable. There are two types of operational definitions:
   a) measured - tells us how the variable or construct will be measured
   b) experimental - lays out the details of how the variable (construct) is manipulated by the experimenter.
7. Types of variables
   a) The *independent* variable is a variable that is varied and has a presumed cause on another variable, the dependent variable. In an experiment, it is the manipulated variable. It is the variable under the control of the experimenter. In a non-experimental study, it is the variable that has a logical effect on the dependent variable.
   b) *dependent* variable is a variable whose effect varies concomitantly with changes or variations in the independent variable.
   c) An *active* variable is a variable that is manipulated. Manipulation means that the experimenter has control over how the values change.
   d) An *attributive* variable is one that is measured and cannot be manipulated. A variable that cannot be manipulated is one where the experimenter has no control over the values of the variable.
   e) A *continuous* variable is one capable of taking on an ordered set of values within a certain range. These variables reflect at least a rank order.
   f) *Categorical* variables belong to a kind of measurement where objects are assigned to a subclass or subset The subclasses are distinct and non-overlapping. All objects put into the same category are considered to have the same characteristic.
   g) *Latent* variables are unobservable entities. They are assumed to underlie observed variables.
   h) *Intervening* variables are constructs that account for internal unobservable psychological processes that account for behavior. It is a variable that cannot be seen but inferred from behavior.

Chapter 4

Variance and Covariance

TO STUDY scientific problems and to answer scientific questions, we must study differences among phenomena. In Chapter 5, we examined relations among variables; in a sense, we were studying similarities. Now we concentrate on differences because without differences and without variation, there is no technical way to determine the relations among variables. If we want to study the relation between race and achievement, for instance, we are helpless if we have only achievement measures of White-American children. We must have achievement measures of children of more than one race. In short, race must vary; it must have variance. It is necessary to explore the variance notion analytically and in some depth. To do so adequately, it is also necessary to skim some of the cream off the milk of statistics.

Studying sets of numbers as they are is unwieldy. It is usually necessary to reduce the sets in two ways: by calculating averages or measures of central tendency, and by calculating measures of variability. The measure of central tendency used in this book is the mean. The measure of variability most used is the *variance.* Both kinds of measures epitomize sets of scores, but in different ways. They are both "summaries" of whole sets of scores, "summaries" that express two important facets of the sets of scores: their central or average tendency and their variability. Solving research problems without these measures is extremely difficult. We start our study of variance, then, with some simple computations.

CALCULATION OF MEANS AND VARIANCES

Take the set of numbers $X = \{1, 2, 3, 4, 5\}$. The mean is defined:

$$M = \frac{\sum X}{n} \qquad (4.1)$$

n = the number of cases in the set of scores; Σ means "the sum of" or "add them up." X stands for any one of the scores, that is, each score is an X. The formula then, says, "Add the scores and divide by the number of cases in the set." Thus:

$$M = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

The mean of the set X is 3. "M" will be used to represent the mean in this book. Other symbols that are commonly used are and m.

Calculating the variance, while not as simple as calculating the mean, is still simple. The formula is:

$$V = \frac{\sum x^2}{n} \qquad (4.2)$$

V means variance; n and Σ are the same as in Equation 4.1. Σx^2 is called the sum of squares; it needs some explanation. The scores are listed in a column:

	X	x	x^2
	1	-2	4
	2	-1	1
	3	0	0
	4	1	1
	5	2	4
ΣX:	15		
M:	3		
Σx^2:			10

In this calculation x is a deviation from the mean. It is defined:

$$x = X - M \qquad (4.3)$$

Thus, to obtain x, simply subtract from X the mean of all the scores. For example, when $X = 1$, $x = 1 - 3 = -2$; when $X = 4$, $x = 4 - 3 = 1$; and so on. This has been done above. Equation 4.2, however, says to square each x. This has also been done above. (Remember, that the square of a negative number is always positive.) In other words, Σx^2 tells us to subtract the mean from each score to get x, square each x to get x^2, and then add up the

x^2's. Finally, the average of the x^2's is taken by dividing Σx^2 by n, the number of cases. Σx^2, the *sum of squares,* is a very important statistic that we will use often.

The variance, in the present case, is

$$V = \frac{(-2)^2 + (-1)^2 + (0)^2 + (1)^2 + (2)^2}{5} =$$

$$\frac{4 + 1 + 0 + 1 + 4}{5} = \frac{10}{5} = 2$$

"V" will be used for variance in this book. Other symbols commonly used are s^2 and s^2. The former is a so-called population value; the latter is a sample value. N is used for the total number of cases in a total sample or in a population. ("Sample" and "population" will be defined in a later chapter.) n is used for a subsample or subset of U of a total sample. Appropriate subscripts will be added and explained as necessary. For example, if we wish to indicate the number of elements in a set A, a subset of U, we can write nA or na. Similarly we attach subscripts to x, V, and so on. When double subscripts are used, such as r_{xy}, the meaning will usually be obvious.

The variance is also called the *mean square* (when calculated in a slightly different way). It is called this because, obviously, it is the mean of the x^2 's. Clearly it is not difficult to calculate the mean and the variance.[1]

The question is: Why calculate the mean and the variance? The rationale for calculating the mean is easily explained. The mean expresses the general level, the center of gravity, of a set of measures. It is a good representative of the level of a group's characteristics or performance. It also has certain desirable statistical properties, and is the most ubiquitous statistic of the behavioral sciences. In much behavioral research, for example, means of different experimental groups are compared to study relations, as pointed out in Chapter 5. We may be testing the relation between organizational climates and productivity, for instance. We may have used three kinds of climates and may be interested in the question of which climate has the greatest effect on productivity. In such cases means are customarily compared. For instance, of three groups, each operating under one of three climates, A_1, A_2, and A_3, which has the greatest mean on, say, a measure of productivity?

The rationale for computing and using the variance in research is more difficult to explain. In the usual case of ordinary scores the variance is a measure of the dispersion of the set of scores. It tells us how much the scores are spread out. If a group of pupils is very heterogeneous in reading achievement, then the variance of their reading scores will be large compared to the variance of a group that is homogeneous in reading achievement. The variance, then, is a measure of the spread of the scores; it describes the extent to which the scores differ from each other. For descriptive purposes, the square root of the variance is ordinarily used. It is called the *standard deviation.* Certain mathematical properties, however, make the variance more useful in research. It is suggested that the student supplement this topic of study with the appropriate sections of an elementary statistics text (see Comrey and Lee, 1995). It will not be possible in this book to discuss all the facets of meaning and interpretation of means, variances, and standard deviations. The remainder of this chapter and later parts of the book will explore other aspects of the use of the variance statistic.

KINDS OF VARIANCE

Variances come in a number of forms. When you read the research and technical literature, you will frequently come across the term, sometimes with a qualifying adjective, sometimes not. To understand the literature, it is necessary to have a good idea of the characteristics and purposes of these different variances. And to design and do research, one must have a rather thorough understanding of the variance concept as well as considerable mastery of statistical variance notions and manipulations.

1. The method of calculating the variance used in this chapter differs from the methods ordinarily used. In fact, the method given above is impracticable in most situations. Our purpose is not to learn statistics, as such. Rather, we are pursuing basic ideas. Methods of computation, examples, and demonstrations have been constructed to aid this pursuit of basic ideas.

Population and Sample Variances

The *population variance* is the variance of *U*, a universe or population of measures. Greek symbols are usually used to represent population parameters or measures. For the population variance the symbol s^2 (sigma squared) is used. The symbol s is used for the population standard deviation. The population mean is m (mu). If all the measures of a defined universal set, *U*, are known, then the variance is known. More likely, however, all the measures of *U* are not available. In such cases the variance is estimated by calculating the variance of one or more samples of *U*. A good deal of statistical energy goes into this important activity. A question may arise: How variable is the intelligence of the citizens of the United States? This is a *U* or population question. If there were a complete list of all the millions of people in the United States—and there were also a complete list of intelligence test scores of these people—the variance could be simply if wearily computed. No such list exists. So samples, representative samples, of Americans are tested and means and variances computed. The samples are used to estimate the mean and variance of the whole population. These estimated values are called statistics (in the population they are called parameters). The sample mean is denoted by the symbol M and the sample variance is denoted by $S.D.^2$ or s^2. A number of statistics textbooks use the $\overline{X}$ (X-bar) to represent the sample mean.

Sampling variance is the variance of statistics computed from samples. The means of four random samples drawn from a population will differ. If the sampling is random and the samples are large enough, the means should not vary too much. That is, the *variance of the means* should be relatively small. [1]

Systematic Variance

Perhaps the most general way to classify variance is as *systematic variance* and *error variance. Systematic variance* is the variation in measures due to some known or unknown influences that "cause" the scores to lean in one direction more than another. Any natural or man-made influences that cause events to happen in a certain predictable way are systematic influences. The achievement test scores of the children in a wealthy suburban school will tend to be systematically higher than the scores of the children in a city slum area school. Adept teaching may systematically influence the achievement of children, as compared to the achievement of children who are ineptly taught.

There are many, many causes of systematic variance. Scientists seek to separate those in which they are interested from those in which they are not interested. They also try to separate from systematic variance, variance that is random. Indeed, research may narrowly and technically be defined as the controlled study of variances.

Between-Groups (Experimental) Variance

One important type of systematic variance in research is between-groups or experimental variance. *Between-groups or experimental variance,* as the name indicates, is the variance that reflects systematic differences between *groups* of measures. The variance discussed previously as score variance reflects the differences between individuals in a group. We can say, for instance, that, on the basis of present evidence and current tests, the variance in intelligence of a random sample of eleven-year-old children is about 225 points. (This is obtained by squaring the standard deviation reported in a test manual. The standard deviation of the California Test of Mental Maturity for 11-year-old children, for instance, is about 15, and $15^2 = 225$.) This figure is a statistic that tells us how much the

1. Unfortunately, in much actual research only one sample is usually available—and this one sample is frequently small. We can, however, estimate the sampling variance of the means by using what is called the standard *variance of the* mean. (The term "standard error of the mean" is usually used. The standard error of the mean is the square root of the standard variance of the mean.) The formula is $V_M = V_S / n_S$ where V_M is the standard variance of the mean, V_S the variance of the sample, and n_S, the size of the sample. Notice an important conclusion that can be reached from this equation. If the size of the sample is increased, V_M is decreased. In other words, to be more confident that the sample is close to the population mean, make *n* large. Conversely, the smaller the sample, the riskier the estimate. (See Study Suggestions 5 and 6 at the end of the chapter.)

individuals differ from each other. Between-groups variance, on the other hand, is the variance due to the differences between *groups* of individuals. If the achievement of northern region and southern region children in comparable schools is measured, there would be differences between the northern and southern groups. Groups as well as individuals differ or vary, and it is possible and appropriate to calculate the variance between these groups.

Between-groups variance and experimental variance are fundamentally the same. Both arise from differences between groups. Between-groups variance is a term that covers all cases of systematic differences between groups, experimental and nonexperimental. Experimental variance is usually associated with the variance engendered by active manipulation of independent variables by experimenters.

Here is an example of between-groups variance—in this case experimental variance. Suppose an investigator tests the relative efficacies of three different kinds of reinforcement on learning. After differentially reinforcing the three groups of subjects, the experimenter calculates the means of the groups. Suppose that they are 30, 23, and 19. The mean of the three means is 24, and we calculate the variance *between the means or between the groups:*

	X	x	x^2
	30	6	36
	23	-1	1
	19	-5	25
ΣX:	72		
M:	24		
Σx^2:			62

In the experiment just described, presumably the different methods of reinforcement tend to "bias" the scores one way or another. This is, of course, the experimenter's purpose. The goal of Method A is to increase all the learning scores of an experimental group. The experimenter may believe that Method B will have no effect on learning, and that Method C will have a depressing effect. If the experimenter is correct, the scores under Method A should all tend to go up, whereas under Method C they should all tend to go down. Thus the scores of the groups, as wholes –and, of course, their means differ systematically. Reinforcement is *an active* variable. It is a variable deliberately manipulated by the experimenter with the conscious intent to "bias" the scores differentially. Prokasy (1987) for example, helps solidify this point by summarizing the number of variations of reinforcement within the Pavlovian paradigm in the study of skeletal responses. Thus any experimenter-manipulated variables are intimately associated with systematic variance. When Camel, Withers and Greenough (1986) gave their experimental group of rats different degrees of early experience-environmental (enriched experiences such as a large cage with other rats and opportunities for exploration), and the control group a condition of reduced experience (isolation, kept in individual cages), they were deliberately attempting the build systematic variance into their outcome measures (pattern and number of dendrite branching. Dendrites are the branching structures of a neuron.) The basic idea behind the famous "classical design" of scientific research, in which experimental and control groups are used, is that, through careful control and manipulation, the experimental group's outcome measures (also called "criterion measures") are made to vary systematically, to all go up or down together, while the control group's measures are ordinarily held at the same level. The variance, of course, is between the two groups, that is, the two groups are made to differ. For example, Braud and Braud (1972) manipulated experimental groups in a most unusual way. They trained the rats of an experimental group to choose the larger of two circles in a choice task; the control group rats received no training. Extracts from the brains of the animals of both groups were injected into the brains of two new groups of rats. Statistically speaking, they were trying to increase the between-groups variance. They succeeded: the new "experimental group" animals exceeded the new "control group" animals in choosing the larger circle in the same choice task!

This is clear and easy to see in experiments. In research that is not experimental, in research where already existing differences between groups are studied, it is not always so clear and easy to see that one is studying between-groups variance. But the idea is the same. The principle may be stated in a somewhat different way: The greater the differences between groups, the more an independent variable or variables can be presumed to have operated. If there is little difference between groups, on the other hand, then the presumption must be that an independent variable or

variables have not operated. That is, their effects are too weak to be noticed, or that different influences have canceled each other out. We judge the effects of independent variables that have been manipulated or that have worked in the past, then, by between-groups variance. Whether the independent variables have or have not been manipulated, the principle is the same.

To illustrate the principle, we use the well-studied problem of the effect of anxiety on school achievement. It is possible to manipulate anxiety by having two experimental groups and inducing anxiety in one and not in the other. This can be done by giving each group the same test with different instructions. We tell the members of one group that their grades depend wholly on the test. We tell the members of the other group that the test does not matter particularly, that its outcome will not affect grades. On the other hand, the relation between anxiety and achievement may also be studied by comparing groups of individuals on whom it can be assumed that different environmental and psychological circumstances have acted to produce anxiety. (Of course, the experimentally induced anxiety and the already existing anxiety—— the stimulus variable and the organismic variable, are not assumed to be the same.) A study to test the hypothesis that different environmental and psychological circumstances act to produce different levels test anxiety was done by Guida and Ludlow (1989). These investigators hypothesized that students in the United States culture would show a lower level of test anxiety than students from the Chilean culture. Using the language of this chapter, the investigators hypothesized a larger between-groups variance than could be expected by chance because of the difference between Chilean and American environmental, educational and psychological conditions. (The hypothesis was supported. Chilean students exhibited a higher level of test anxiety than students from the United States. However, when considering only the lower socio-economic groups of each culture, the United States students had higher test anxiety than the Chilean students.)

Error Variance

Error variance is the fluctuation or varying of measures that is unaccounted for. The fluctuations of the measurements in the dependent variable in a research study where all participants were treated equally is considered error variance. Some of these fluctuations are due to chance. In this case, error variance is random variance. It is the variation in measures due to the usually small and self-compensating fluctuations of measures-now here, now there; now up, now down. The sampling variance discussed earlier in the chapter, for example, is random or error variance.

To digress briefly before continuing, it is necessary in this chapter and the next to use the notion of "random" or "randomness." Ideas of randomness and randomization will be discussed in considerable more detail in a ;ater chapter. For the present, however, *randomness* means that there is no known way that can be expressed in language of correctly describing or explaining events and their outcomes. In different words, random events cannot be predicted. A random sample is a subset of a universe. Its members are so drawn that each member of the universe has an equal chance of being selected. This is another way of saying that, if members are randomly selected, there is no way to predict which member will be selected on any one selection-other things equal.

However, one should not think that random variance is the only possible source of error variance. Error variance can be also consist of other components as pointed out by Barber (1976). What gets “pooled” into the term called error variance can include measurement errors within the measuring instrument, procedural errors by the researcher, misrecording of responses, and the researcher’s outcome expectancy. It is possible that “equal” subjects differ on the dependent variable because one may be experiencing a different physiological and psychological functioning at the time the measurements were taken.

Returning to our main discussion, it can be said that error variance is the variance in measurements due to ignorance. Imagine a great dictionary in which everything in the world—every occurrence, every event, every little thing, every great thing—is given in complete detail. To understand any event that has occurred, that is now occurring, or that will occur, all one needs to do is look it up in the dictionary. With this dictionary there are obviously no random or chance occurrences. Everything is accounted for. In brief, there is no error variance; all is systematic variance. Unfortunately, (or more likely, fortunately) we do not have such a dictionary. Many, many events and occurrences cannot be explained. Much variance eludes

identification and control. This is error variance as long as identification and control elude us.

While seemingly strange and even a bit bizarre, this mode of reasoning is useful, provided we remember that some of the error variance of today may not be the error variance of tomorrow. Suppose that we do an experiment on teaching problem solving in which we assign pupils to three groups at random. After we finish the experiment, we study the differences between the three groups to see if the teaching has had an effect. We know that the scores and the means of the groups will always show minor fluctuations, now plus a point or two or three, now minus a point or two or three, which we can probably never control. Something or other makes the scores and the means fluctuate in this fashion. According to the view under discussion, they do not just fluctuate for any reason; there is probably no "absolute randomness." Assuming determinism, there must be some cause or causes for the fluctuations. True, we can learn some of them and possibly control them. When we do this, however, we have systematic variance.

We find out, for instance, that gender "causes" the scores to fluctuate, since males and females are mixed in the experimental groups. (We are, of course, talking figuratively here. Obviously gender does not make scores fluctuate.) So we do the experiment and control gender by using, say, only males. The scores still fluctuate, though to a somewhat lesser extent. We remove another presumed cause of the perturbations: intelligence. The scores still fluctuate, though to a still lesser extent. We go on removing such sources of variance. We are controlling systematic variance. We are also gradually identifying and controlling more and more unknown variance.

Now note that before we controlled or removed these systematic variances, before we "knew" about them, we would have to label all such variance as “error variance”—partly through ignorance and partly through inability to do anything about such variance. We could go on and on doing this and there will still be variance left over. Finally we give in; we "know" no more; we have done all we can. There will still be variance. A practical definition of error variance, then, would be: *Error variance* is the variance left over in a set of measures after all known sources of systematic variance have been removed from the measures. This is so important it deserves a numerical example.

An Example of Systematic and Error Variance

Suppose we are interested in knowing whether politeness in the wording of instructions for a task affects memory of the polite words. Call "politeness" and "impoliteness" the variable A partitioned into A_1 and A_2. (this idea was from Holtgraves, 1997.) The students are assigned at random to two groups. Treatments A_1 and A_2 are assigned at random to the two groups. In this experiment the students of A_1 received instructions that were worded impolitely, such as, “You must write out the full name for each state you remember.” The students of A_2, on the other hand, received instructions that were of the same meaning as those received by A_1 students. However, the wording of the instructions was in a polite form, such as, “It would help if you write out the full name for each state you recall.” After reading the instructions, the subjects are given a distracter task. This task involved recalling the 50 states of the United States. The students are subsequently given a recognition memory test. This test was used to determine the overall memory of the polite words. The scores are as follows:

	A_1	A_2
	3	6
	5	5
	1	7
	4	8
	2	4
M	3	6

The means are different; they vary. There is between-groups variance. Taking the difference between the means at face value—later we will be more precise—we may conclude that vagueness in lecturing had an effect. Calculating the between-groups variance just as we did earlier, we get:

		x	x^2
	3	1.5	2.25
	6	1.5	2.25
M	4.5		
Σx^2			4.50

$$V = \frac{4.5}{2} = 2.25$$

In other words, we calculate the between-groups variance just as we earlier calculated the variance of the five scores 1, 2, 3, 4 and 5. We simply treat the two means as though they were individual scores, and go ahead with an ordinary variance calculation. The between-groups variance, V_b, is, then, 2.25. An appropriate statistical test would show that the difference between the means of the two groups is what is called a "statistically significant" difference. (The meaning of this will be taken up in another chapter.) [1] Evidently, using polite words in instructions helped increase the memory scores of the students.

If we put the 10 scores in a column and calculate the variance we get:

X	x	x^2
3	-1.5	2.25
5	.5	.25
1	-3.5	12.25
4	-.5	.25
2	-2.5	6.25
6	1.5	2.25
5	.5	.25
7	2.5	6.25
8	3.5	12.25
4	-.5	.25
M:	4.5	
Σx^2		42.50

$$V_t = \frac{42.5}{10} = 4.25$$

This is the total variance, V_t . V_t = 4.25 contains all sources of variation in the scores. We already know that one of these is the between-groups variance, V_b = 2.25.

Let us calculate still another variance. We do this by calculating the variance of A_1 alone and the variance of A_2 alone and then averaging the two:

A_1	x	x^2	A_2	x	x^2
3	0	0	6	0	0
5	2	4	5	-1	1
1	-2	4	7	1	1
4	1	1	8	2	4
2	-1	1	4	-2	4
ΣX	15		30		
M	3		6		
Σx^2		10			10

$$V_{A_1} = \frac{10}{5} = 2 \qquad V_{A_2} = \frac{10}{5} = 2$$

The variance of A_1 is 2, and the variance of A_2 is 2. The average is 2. Since each of these variances was calculated *separately* and then averaged, we call the average variance calculated from them the "within-groups variance." We label this variance V_w meaning within variance, or within-groups variance. Thus V_w = 2. *This variance is unaffected by the difference between the* two means. This is easily shown by subtracting a constant of 3 from the scores of A_2. This makes the mean of A_2 equal to 3. Then, if the variance of A_2 is calculated, it will be the same as before: 2. Obviously the within-groups variance will be the same: 2.

Now write an equation: $V_t = V_b + V_w$. This equation says that the total variance is made up of the variance between the groups and the variance within the groups. Is it? Substitute the numerical values: 4.25 = 2.25 + 2.00. Our method works—it shows us, too, that these variances are additive (as calculated).

The variance ideas under discussion can perhaps be clarified with a diagram. In Figure 4.1, a circle broken up into two parts has been drawn. Let the area of the total circle represent the total variance of the 10 scores, or V_t. The larger shaded portion represents the between-groups variance, or V_b. The smaller unshaded portion represents the error variance, or V_w or V_e. From the diagram one can see that $V_t = V_b + V_e$. (Note the similarity to set thinking and the operation of union.)

1. The method of computation used here is not what would be used to test statistical significance. It is used here purely as a pedagogical device. Note, too, that the small numbers of cases in the examples given and the small size of the numbers are used only for simplicity of demonstration. Actual research data, of course, are usually more complex, and many more cases are needed. In actual analysis of variance the correct expression for the between sum of squares is: $SS_b = n\Sigma x_b^2$. For pedagogical simplicity, however, we retain Σx_b^2, later replacing it with SS_b

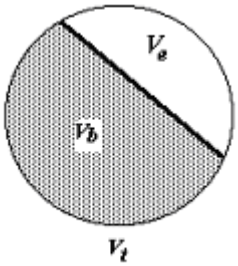


Figure 4.1

A measure of all sources of variance is represented by V_t and a measure of the between-groups variance (or a measure of the effect of the experimental treatment) by V_b. But what is V_w, the within-groups variance? Since, of the total variance, we have accounted for a known source of variance, via the between-groups variance, we assume that the variance remaining is due to chance or random factors. We call it *error variance*. But, you may say, surely there must be other sources of variance? How about individual differences in intelligence, gender, and so on? Since we assigned the students to the experimental groups at random, assume that these sources of variance are equally, or approximately equally, distributed between A_1 and A_2. And because of the random assignment we cannot isolate and identify any other sources of variance. So we call the variance remaining error variance, knowing full well that there are probably other sources of variance but assuming, and hoping our assumption is correct, that they have been equally distributed between the two groups.

A Subtractive Demonstration: Removing Between-Groups Variance from Total Variance

Let us demonstrate all this another way by removing from the original set of scores the between-groups variance, using a simple subtractive procedure. First, we let each of the means of A_1 and A_2 be equal to the total mean; we remove the between-groups variance. The total mean is 4.5. (See above where the mean of all 10 scores was calculated.) Second, we adjust each individual score of A_1 and A_2 by subtracting or adding, as the case may be, an appropriate constant. Since the mean of A_1 is 3, we add 4.5 -3 = 1.5 to each of the A_1 scores. The mean of A_2 is 6, and 6 - 4.5 = 1.5 is the constant to be *subtracted* from each of the A_2 scores.

Study the "corrected" scores. Compare them with the original scores. Note that they vary less than they did before. Naturally. We removed the between-groups variance, a sizable portion of the total variance. The variance that remains is that portion of the total variance due, presumably, to chance. We calculate the variance of the "corrected" scores of A_1, A_2, and the total, and note these surprising results:

Correction:	+ 1.5	–1.5
	A_1	A_2
	3 + 1.5 = 4.5	5 – 1.5 = 3.5
	5 + 1.5 = 6.5	5 – 1.5 = 3.5
	1 + 1.5 = 2.5	7 – 1.5 = 5.5
	4 + 1.5 = 5.5	8 – 1.5 = 6.5
	2 + 1.5 = 3.5	4 – 1.5 = 2.5
ΣX	22.5	22.5
M	4.5	4.5

	A_1	x	x^2	A_2	x	x^2
	4.5	0	0	4.5	0	0
	6.5	2	4	3.5	–1	1
	2.5	–2	4	5.5	1	1
	5.5	1	1	6.5	2	4
	3.5	–1	1	2.5	–2	4
ΣX	22.5			22.		
M	4.5			5		
Σx^2			10	4.5		10

$$V_{A_1} = \frac{10}{5} = 2 \qquad V_{A_2} = \frac{10}{5} = 2$$

The within-groups variance is the same as before. It is unaffected by the correction operation. Obviously the between-groups variance is now zero. What about the total variance, V_t? Calculating it, we obtain $Sx_t^2 = 20$, and $V_t = 20 \div 10 = 2$. Thus the within-groups variance is now equal to the total variance. The reader should study this example carefully until one has a firm grasp of what has happened and *why*.

Although the previous example is perhaps sufficient to make the essential points, it may solidify the student's

understanding of these basic variance ideas if we extend the example by putting in and pulling out another source of variance. The reader may recall that we knew that the within-groups variance contained variation due to individual differences. Now assume that, instead of randomly assigning the students to the two groups, we had matched them on intelligence–and intelligence is related to the dependent variable. That is, we put pair members with approximately equal intelligence test scores into the two groups. The outcome of the experiment might be:

	A_1	A_2
	3	6
	1	5
	4	7
	2	4
	5	8
M	3	6

Note carefully that the only difference between this setup and the previous one is that the matching has caused the scores to covary. The A_1 and A_2 measures now have nearly the same rank order. In fact, the coefficient of correlation between the two sets of scores is 0.90. We have here another source of variance: that due to individual differences in intelligence which is reflected in the rank order of the pairs of criterion measures. (The precise relation between the rank order and matching ideas and their effects on variance will be taken up in another chapter. The student should take it on faith for the present that matching produces systematic variance.)

This variance can be calculated and extracted as before, except that there is an additional operation. First equalize the $A_1$and A_2 means and "correct" the scores as before. This yields:

Correction :	+ 1.5	–1.5
	4.5	4.5
	2.5	3.5
	5.5	5.5
	3.5	2.5
	6.5	6.5
M	4.5	4.5

Second, by equalizing the rows (making each row mean equal to 4.5 and "correcting" the row scores accordingly) we find the following data:

Correction	A_1	A_2
0	4.5 + 0 = 4.5	4.5 + 0 = 4.5
+1.5	2.5 + 1.5 = 4.0	3.5 + 1.5 = 5.0
–1.0	5.5 – 1.0 = 4.5	5.5 – 1.0 = 4.5
+1.5	3.5 + 1.5 = 5.0	2.5 + 1.5 = 4.0
–2.0	6.5 – 2.0 = 4.5	6.5 – 2.0 = 4.5
M	4.5	4.5

Original Means	Corrected Means
4.5	4.5
3.0	4.5
5.5	4.5
3.0	4.5
6.5	4.5
4.5	4.5

The doubly corrected measures now show very little variance. The variance of the 10 doubly corrected scores is 0.10, very small indeed. There is no between-groups (columns) or between-individuals (rows) variance left in the measures, of course. After double correction, all of the total variance is error variance. (As we will see later, when the variances of both columns and rows are extracted like this—although with a quicker and more efficient method—there is no within-groups variance.)

This has been a long operation. A brief recapitulation of the main points may be useful. Any set of measures has a total variance. If the measures from which this variance is calculated have been derived from the responses of human beings, then there will always be at least two sources of variance. One will be due to systematic sources of variation like individual differences of the subjects whose characteristics or accomplishments have been measured and differences between the groups or subgroups involved in research. The other will be due to chance or random error, fluctuations of measures that cannot currently be accounted for. Sources of systematic variance tend to make scores lean in one direction or another. This is reflected in differences in means, of course. If gender is a systematic source of variance in a study of school

achievement, for instance, then the gender variable will tend to act in such a manner that the achievement scores of females will tend to be higher than those of males. Sources of random error, on the other hand, tend to make measures fluctuate now this way now that way. Random errors, in other words, are self-compensating; they tend to balance each other out.

In any experiment or study, the independent variable (or variables) is a source of systematic variance—at least it should be. The researcher "wants" the experimental groups to differ systematically. The researcher usually seeks to maximize such variance while controlling or minimizing other sources of variance, both systematic and error. The experimental example given above illustrates the additional idea that these variances are additive, and that because of this additive property, it is possible to analyze a set of scores into systematic and error variances.

COMPONENTS OF VARIANCE

The discussion so far may have convinced the student that any total variance has "components of variance." The case just considered, however, included one experimental component due to the difference between A_1 and A_2, one component due to individual differences, and a third component due to random error. We now study the case of two components of systematic experimental variance. To do this, we synthesize the experimental measures, creating them from *known* variance components. In other words, we go backwards. We start from "known" sources of variance, because there will be no error variance in the synthesized scores.

We have a variable X that has three values. Let $X = \{0,1, 2\}$. We also have another variable Y, which has three values. Let $Y = \{0, 2, 4\}$. X and Y, then, are *known* sources of variance. We assume an ideal experimental situation where there are two independent variables acting *in concert to* produce effects on a dependent variable, Z. That is, each score of X operates with each score of Y to produce a dependent variable score Z. For example, the X score, 0, has no influence. The X score, 1, operates with Y as follows: $\{(1 + 0), (1 + 2), (1 + 4)\}$. Similarly, the X score, 2, operates with Y: $\{(2 + 0), (2 + 2), \text{and } (2 + 4)\}$. All this is easier to see if we generate Z in clear view.

The set of scores in the 3 ¥ 3 matrix (a matrix is any rectangular set or table of numbers) is the set of Z scores. The purpose of this example will be lost unless the reader remembers that in practice we do not know the X and Y scores; we only know the Z scores. In actual experimental situations we manipulate or set up X and Y. But we only hope they are effective. They may not be. In other words, the sets $X = \{0, 1, 2\}$ and $Y = \{0, 2, 4\}$ can never be known like this. The best we can do is to estimate their influence by estimating the amount of variance in Z due to X and to Y.

		Y		
		0	2	4
X	0	0 + 0	0 + 2	0 + 4
	1	1 + 0	1 + 2	1 + 4
	2	2 + 0	2 + 2	2 + 4

=

	Z		
	0	2	4
0	0	2	4
1	1	3	5
2	2	4	6

The sets X and Y have the following variances:

	X	x	x^2
	0	–1	1
	1	0	0
	2	1	1
ΣX:	3		
M:	1		
Σx^2:			2

	Y	y	y^2
	0	–2	4
	2	0	0
	4	2	4
ΣY:	6		
M:	2		
Σy^2:			8

The set Z has variance as follows:

$$V_x = \frac{2}{3} = .67 \qquad V_y = \frac{8}{3} = 2.67$$

	Z	z	z^2
	0	−3	9
	2	−1	1
	4	1	1
	1	−2	4
	3	0	0
	5	2	4
	2	−1	1
	4	1	1
	6	3	9
ΣZ:	27		
M:	3		
Σz^2			20

Now, 67 + 2.67 = 3.34, or $V_z = V_x + V_y$, within errors of rounding.

This example illustrates that, under certain conditions, variances operate additively to produce the experimental measures we analyze. While the example is "pure" and therefore unrealistic, it is not unreasonable. It is possible to think of X and Y as independent variables. They might be level of aspiration and pupil attitudes. And Z might be verbal achievement, a dependent variable. That real scores do not behave exactly this way does not alter the idea. They behave approximately this way. We plan research to make this principle as true as possible, and we analyze data as though it were true. And it works!

COVARIANCE

Covariance is really nothing new. Recall, in an earlier discussion of sets and correlation that we talked about the relation between two or more variables being analogous to the intersection of sets. Let X be {0, I, 2, 3}, a set of attitude measures for four children. Let Y be {1, 2, 3, 4}, a set of achievement measures of the same children, but not in the same order. Let R be a set of ordered pairs of the elements of X and Y, the rule of pairing being: each individual's attitude and achievement measures are paired, with the attitude measure placed first. Assume that this yields R = {(0, 2), (1, 1), (2, 3), (3, 4)}. By our previous definition of relation, this set of ordered pairs is a relation, in this case the relation between X and Y. The results of the calculations of the variance of X and the variance of Y are:

	X	x	x^2	Y	y	y^2
	0	−1.5	2.25	2	−.5	.25
	1	−.5	.25	1	−1.5	2.2
	2	.5	.25	3	.5	.25
	3	1.5	2.25	4	1.5	2.2
ΣX:	6			10		
M:	1.5			2.5		
Σx^2			5			5
:						

We now set ourselves a problem. (Note carefully in what follows that we are going to work with deviations from the mean, x's and y's, and not with the original raw scores.) We have calculated the variances of X and Y above by using the x's and y's, that is, the deviations from the respective means of X and Y. If we can calculate the variance of any set of scores, is it not possible to calculate the relation *between* any two sets of scores in a similar way? Is it conceivable that we can calculate the variance of the two sets simultaneously? And if we do so, will this be a measure of the variance of the two sets together? Will this variance also be a measure of the relation between the two sets?

What we want to do is to use some statistical operation analogous to the set operation of intersection, $X \cap Y$. To calculate the variance of X or of Y, we squared the deviations from the mean, the x's or the y's, and then added and averaged them. A natural answer to our problem is to perform an analogous operation on the x's and y's *together*. To calculate the variance of X, we did this first: $(x_1 \cdot x_1), \ldots, (x_4 \cdot x_4) = x_1^2, \ldots, x_4^2$. Why, then not follow this through with both x's and y's, multiplying the ordered pairs like this: $(x_1 \cdot y_1), \ldots, (x_4 \cdot y_4)$? Then, instead of writing Sx^2 or Sy^2, we write Sxy, as follows:

x	*y*	=	*xy*
−1.5	−.5	=	.75
−.5	−1.5	=	.75
.5	.5	=	.25
1.5	1.5	=	2.25

$\Sigma xy = 4.00$

$$V_{xy} = CoV_{xy} = \frac{4}{4} = 1.00$$

Let use give names to S*xy* and V_{xy}. S*xy* is called the *cross product,* or the sum of the cross products. V_{xy} is called the *covariance.* We will write it *CoV* with suitable subscripts. If we calculate the variance of these products-symbolized as V_{xy}, or CoV_{xy}—we obtain 1.00, as indicated above. This 1.00, then, can be taken as an index of the relation between two sets. But it is an unsatisfactory index because its size fluctuates with the ranges and scales of different *X*'s and *Y*'s. That is, it might be 1.00 in this case and 8.75 in another case, making comparisons from case to case difficult and unwieldy. We need a measure that is comparable from problem to problem. Such a measure —an excellent one, too–is obtained simply by writing a fraction or ratio. It is the covariance, CoV_{xy} divided by an average of the variances of *X* and *Y.* The average is usually in the form of a square root of the product of *Vx* and *Vy.* The whole formula for our index of relation, then, is

$$R = \frac{CoV_{xy}}{\sqrt{V_x \cdot V_y}}$$

This is one form of the well-known product-moment coefficient of correlation. Using it with our little problem, we obtain:

$$R = \frac{CoV_{xy}}{\sqrt{V_x \cdot V_y}} = \frac{1.00}{1.25} = .80$$

This index, usually written r, can range from + 1.00 through 0 to -1.00, as we learned in Chapter 5.

So we have another important source of variation in sets of scores, provided the set elements, the *X*'s and *Y*'s, have been ordered into pairs after conversion into deviation scores. The variation is aptly called *covariance* and is a measure of the relation between the sets of scores.

It can be seen that the definition of relation as a set of ordered pairs leads to several ways to define the relation of the above example:

R = {(*x*, y); *x* and y are numbers, *x* always coming first} *xRy* = the same as above or "*x* is related to y"
R = {(0, 2), (1, 1), (2, 3), (3, 4)}
R = { (-1.5,-.5), (-.5,-1.5), (.5, .5), (1.5, 1.5)}

$$R = \frac{CoV_{xy}}{\sqrt{V_x \cdot V_y}} = \frac{1.00}{1.25} = .80$$

Variance and covariance are concepts of the highest importance in research and in the analysis of research data. There are two main reasons. One, they summarize, so to speak, the variability of variables and the relations among variables. This is most easily seen when we realize that correlations are covariances that have been standardized to have values between -1 and +1. . But the term also means the covarying of variables in general. In much or most of our research we literally pursue and study covariation of phenomena. Two, variance and covariance form the statistical backbone of multivariate analysis, as we will see toward the end of the book. Most discussions of the analysis of data are based on variances and covariances. Analysis of variance, for example, studies different sources of variance of observations, mostly in experiments, as indicated earlier. Factor analysis is in effect the study of covariances, one of whose major purposes is to isolate and identify common sources of variation. The contemporary ultimate in analysis, the most powerful and advanced multivariate approach yet devised, is called analysis of covariance structures because the system studies complex sets of relation by analyzing the covariances among variables. Variances and covariances will obviously be the core of much of our discussion and preoccupation from this point on.

Study Suggestions

1. A social psychologist has done an experiment in which one group, A_1, was given a task to do in the

presence of an audience, and another group, A_2, was given the same task to do without an audience. The scores of the two groups on the task, a measure of digital skill, were:

A_1	A_2
5	3
5	4
9	7
8	4
3	2

(a) Calculate the means and variances of A_1 and A_2, using the method described in the text.

(b) Calculate the between-groups variance, V_b, and the within-groups variance, V_w.

(c) Arrange all ten scores in a column, and calculate the total variance, V_t.

(d) Substitute the calculated values obtained in (b) and (c), above, in the equation: $V_t = V_b + V_w$. Interpret the results.

[Answers: (a) V_{a1}, = 4.8; V_{a2} = 2.8; (b) V_b = 1.0; V_w = 3.8; (c) V_t = 4.8.]

2. Add 2 to each of the scores of A_1 in 1, above, and calculate V_t, V_b, and V_w. Which of these variances changed? Which stayed the same? Why? [Answers: V_t = 7.8; V_b = 4.0; V_w = 3.8.]
3. Equalize the means of A_1 and A_2, in 1, above, by adding a constant of 2 to each of the scores of A_2. Calculate V_t, V_b, and V_w. What is the main difference between these results and those of 1, above? Why?
4. Suppose a sociological researcher obtained measures of conservatism (A), attitude toward religion (B), and anti-Semitism (C) from 100 individuals. The correlations between the variables were: r_{ab} = .70; r_{ac} = .40; r_{bc} = .30. What do these correlations mean? [Hint: Square the *r*'s before trying to interpret the relations. Also, think of ordered pairs.]
5. The purpose of this study suggestion and Study Suggestion 6 is to give the student an intuitive feeling for the variability of sample statistics, the relation between population and sample variances, and between-groups and error variances. Appendix C contains 40 sets of 100 random numbers 0 through 100, with calculated means, variances, and standard deviations. Draw 10 sets of 10 numbers each from 10 different places in the table.

   a. Calculate the mean, variance, and standard deviation of each of the 10 sets. Find the highest and lowest means and the highest and lowest variances. Do they differ much from each other? What value "should" the means be? (50) While doing this, save the 10 totals and calculate the mean of all 100 numbers. Do the 10 means differ much from the total mean? Do they differ much from the means reported in the table of means, variances, and standard deviations given after the random numbers?
   b. Count the odd and even numbers in each of the 10 sets. Are they what they "should be"? Count the odd and even numbers of the 100 numbers. Is the result "better" than the results of the 10 counts? Why should it be?
   c. Calculate the variance of the 10 means. This is, of course, the between-groups variance, V_b. Calculate the error variance, using the formula: $V_e = V_t - V_b$.
   d. Discuss the meaning of your results after reviewing the discussion in the text.
6. As early as possible in their study, students of research should start to understand and use the computer. Study Suggestion 5 can be better and less laboriously accomplished with the computer. It would be better, for example, to draw 20 samples of 100 numbers each. Why? In any case, students should learn how to do simple statistical operations using existing computer facilities and programs at their institutions. All institutions have programs for calculating means and standard deviations (variances can be obtained by squaring the standard deviations)[1] and for generating random numbers. If you can use your institution's facilities, use them for Study Suggestion 5, but increase the number of samples and their *n*'s.

Chapter Summary

1. Differences between measurements are needed in order to study the relations between variables.
2. A statistical measure used in studying differences is the variance.
3. The variance along with the mean is used to solve research problems.
4. Kinds of variance.
   (a) The variability of a variable or characteristic in the universe or population is called the population variance
   (b) A subset of the universe is called a sample and that sample also has variability. That variability is referred to as sample variance.
   (c) Since the statistic computed from sample-to-sample differs, this difference is referred to as sampling variance.
   (d) Systematic variance is the variation that can be accounted for. It can be explained. Any natural or human made influences that causes events to happen in a predictable way is systematic variance
   (e) A type of systematic variance is called between-groups variance. When there are differences between groups of subjects, and the cause of that difference is known, it is referred to as between-group variance.
   (f) A type of systematic variance is called experimental variance. Experimental variance is a bit more specific than between-groups variance in that it is associated with variance engendered by active manipulation of the independent variable.
   (g) Error variance is the fluctuation or varying of measures in the dependent variable that cannot be directly explained by the variables under study. One part of error variance is due to chance. This is also known as random variance. The source of this fluctuation is generally unknown. Other possible sources for error variance include the procedure of the study, the measuring instrument and the researcher's outcome expectancy.
5. Variances can be broken down into components. In this case, the term variance is referred to as total variance. The partitioning of total variance into components of systematic and error variances plays an important role in statistical analyses of research data.
6. Covariance is the relationship between two or more variables.
   (a) it is an unstandardized correlation coefficient
   (b) covariance and variance are the statistical foundations of multivariate statistics (to be presented in later chapters).

---

1. There may be small discrepancies between your hand-calculated standard deviations and variances and those of the computer because existing programs and built-in routines of hand-held calculators usually use a formula with $N - 1$ rather than N in the denominator of the formula. The discrepancies will be small, however, especially if N is large. (The reason for the different formulas will be explained later when we take up sampling and other matters.)

Chapter 5

Sampling and Randomness

IMAGINE the many situations in which we want to know something about people, about events, about things. To learn something about people, for instance, we take some few people whom we know—or do not know—and study them. After our "study," we come to certain conclusions, often about people in general. Some such method is behind much folk wisdom. Commonsensical observations about people, their motives, and their behaviors derive, for the most part, from observations and experiences with relatively few people. We make such statements as: "People nowadays have no sense of moral values"; "Politicians are corrupt"; and "Public school pupils are not learning the three R's." The basis for making such statements is simple. People, mostly through their limited experiences, come to certain conclusions about other people and about their environment. In order to come to such conclusions, they must *sample* their "experiences" of other people. Actually, they take relatively small samples of all possible experiences. The term "experiences" here has to be taken in a broad sense. It can mean direct experience with other people—for example, first-hand interaction with, say, Moslems or Asians. Or it can mean indirect experience: hearing about Moslems or Asians from friends, acquaintances, parents, and others. Whether experience is direct or indirect, however, does not concern us too much at this point. Let us assume that all such experience is direct. An individual claim to "know" something about Asians and says " I 'know' they are clannish because I have had direct experience with a number of Asians." Or, "Some of my best friends are Asians, and I know that. . . " The point is that this person's conclusions are based on a sample of Asians, or a sample of the behaviors of Asians, or both. This individual can never "know" all Asians; and must depend, in the last analysis, on samples. Indeed, most of the world's knowledge is based on samples, most often on inadequate samples.

SAMPLING, RANDOM SAMPLING, AND REPRESENTATIVENESS

Sampling refers to taking a portion of a population or universe as representative of that population or universe. This definition does not say that the sample taken-or drawn, as researchers say—is representative. It says, rather, taking a portion of the population and *considering* it to be representative. When a school administrator visits certain classrooms in the school system "to get the feel of the system," that administrator is sampling classes from all the classes in the system. This person might assume that by visiting, say, eight to ten classes out of forty "at random," he or she will get a fair notion of the quality of teaching going on in the system. Or another way would be to visit one teacher's class two or three times to sample teaching performance. By doing this, the administrator is now sampling behaviors, in this case teaching behaviors, from the universe of all possible behaviors of the teacher. Such sampling is necessary and legitimate. However, you can come up with some situations where the entire universe could be measured. So why bother with samples? Why not just measure every element of the universe? Take a census? One major reason is one of economics. The second author (HBL) worked in a marketing research department of a large grocery chain in southern California. This chain consisted of 100 stores. Research on customers and products occasionally take the form of a controlled-store test. These are studies conducted in a real day-to-day grocery store operation. Perhaps one might be interested in testing a new dog food. Certain stores would be chosen to receive the new product while another set of stores will not carry the new product. Secrecy is very important in such studies. If a competing manufacture of dog food received information that a marketing test was being done at such-and-such store, they can contaminate the results. To conduct controlled stored tests of cents-off coupons, or new products, or shelf allocation space, a research study could be conducted with two groups of 50 stores each. However, the labor and administrative costs alone would be prohibitive. It would make a lot more sense to use samples that are representative of the population. Choosing 10 stores in each group to conduct the research would cut down the costs of doing the study. Smaller studies are more manageable and controllable. The study using samples can be completed on a timely manner. In some disciplines, such as quality control and education-instructional evaluation, sampling is essential. In quality control there is a procedure called destructive testing. One way to determine whether a product meets specification is to put it through an actual performance test. When the product is destroyed (fails), the product can be evaluated. In testing tires, for example, it would make little sense to destroy every tire just to determine if the manufacturer has adequate quality

control. Likewise the teacher who wants to determine if the child has learned the material will give an examination. It would be difficult to write an examination that covers every aspect of instruction and knowledge retention of the child.

Random sampling is that method of drawing a portion (or sample) of a population or universe so that each member of the population or universe has an equal chance of being selected. This definition has the virtue of being easily understood. Unfortunately, it is not entirely satisfactory because it is limited. A better definition is by Kirk (1990, p. 8).

> The method of drawing samples from a population such that every possible sample of a particular size has an equal chance of being selected is called *random sampling*, and the resulting samples are *random samples*.

This definition is general and thus more satisfactory than the earlier definition.

Define a universe to be studied as all fourth-grade children in X school system. Suppose there are 200 such children. They comprise the population (or universe). We select one child at random from the population. His (or her) chance of being selected is 1/200, if the sampling procedure is random. Likewise, a number of other children are similarly selected. Let us assume that after selecting a child, the child is returned (or the symbol assigned to the child) to the population. Then the chance of selecting any second child is also 1/200. (If we do not return this child to the population, then the chance each of the remaining children has, of course, is 1/199. This is called *sampling without replacement.* When the sample elements are returned to the population after being drawn, the procedure is called *sampling with replacement.)*

Suppose from the population of the 200 fourth-grade children in X school system we decide to draw a random sample of 50 children. This means, if the sample is random, that all possible samples of 50 have the same probability of being selected—a very large number of possible samples. To make the ideas involved comprehensible, suppose a population consists of four children, a, b, c, and d, and we draw a random sample of two children. Then the list of all the possibilities, or the *sample* space, is: (a, b), (a, c), (a, d), (b, c), (b, d), (c, d). There are six possibilities. If the sample of two is drawn at random, then its probability is 1/6. Each of the pairs has the same probability of being drawn. This sort of reasoning is needed to solve many research problems, but we will usually confine ourselves to the simpler idea of sampling connected with the first definition. The first definition, then, is a special case of the second general definition-the special case in which $n = 1$.

Unfortunately, we can never be sure that a random sample is representative of the population from which it is drawn. Remember that any particular sample of size n has the same probability of being selected as any other sample of the same size. Thus, a particular sample may not be representative at all. We should know what "representative" means. Ordinarily, "representative" means to be typical of a population, that is, to exemplify the characteristics of the population. From a research point of view, "representative" must be more precisely defined, though it is often difficult to be precise. We must ask: What characteristics are we talking about? So, in research, a "representative sample" means that the sample has approximately the characteristics of the population relevant to the research in question. If sex and socioeconomic class are variables (characteristics) relevant to the research, a representative sample will have approximately the same proportions of men and women and middle-class and working class individuals as the population. When we draw a random sample, *we hope* that it will be representative. We hope that the relevant characteristics of the population will be present in the sample in approximately the same way they are present in the population. But we can never be sure. There is no guarantee.

What we rely on is the fact, as Stilson (1966) points out, that the characteristics typical of a population are those that are the most frequent and therefore most likely to be present in any particular random sample. When sampling is random, the sampling variability is predictable. We learned in Chapter 7, for example, that if we throw two dice a number of times, the probability of a 7 turning up is greater than that of a 12 turning up. (See Table 7.1.)

A sample drawn at random is unbiased in the sense that no member of the population has any more chance of being selected than any other member. We have here a democracy in which all members are equal before the bar of selection. Rather than using coins or dice, let's use a research example. Suppose we have a population of 100 children. The children differ in intelligence, a variable relevant to our research. We want to know the mean intelligence score of the population, but for some reason we can only sample 30 of the 100 children. If we sample randomly, there are a large number of possible samples of 30 each. The samples have equal probabilities of being selected.

The means of most of the samples will be relatively close to the mean of the population. A few will not be close. If the sampling has been random, the probability of selecting a sample with a mean close to the population mean is greater than the probability of selecting a sample with a mean not close to the population mean.

If we do not draw our sample at random, however, some factor or factors unknown to us may predispose us to select a biased sample. In this case perhaps one of the samples with a mean not close to the population mean. The mean intelligence of this sample will then be a biased estimate of the population mean. If we knew the 100 children, we might unconsciously tend to select the more intelligent children. It is not so much that *we would do* so; it is that our method *allows* us to do so. Random methods of selection do not allow our own biases or any other systematic selection factors to operate. The procedure is objective, divorced from our own predilections and biases.

The reader may be experiencing a vague and disquieting sense of uneasiness. If we can't be sure that random samples are representative, how can we have confidence in our research results and their applicability to the populations from which we draw our samples? Why not select samples systematically so that they are representative? The answer is complex. First—and again—we cannot ever be sure. Second, random samples are more likely to include the characteristics typical of the population if the characteristics are frequent in the population. In actual research, we draw random samples whenever we can and hope and assume that the samples are representative. We learn to live with uncertainty. We try to cut it down whenever we can—just as we do in ordinary day-to-day living but more systematically and with considerable knowledge of and experience with random sampling and random outcomes. Fortunately, our lack of certainty does not impair our research functioning.

RANDOMNESS

The notion of randomness is at the core of modern probabilistic methods in the natural and behavioral sciences. But it is difficult to define "random." The dictionary notion of haphazard, accidental, without aim or direction, does not help us much. In fact, scientists are quite systematic about randomness; they carefully select random samples and plan random procedures.

The position can be taken that nothing happens at random, that for any event there is a cause. The only reason, this position might say, that one uses the word random is that human beings do not know enough. To omniscience nothing is random. Suppose an omniscient being has an omniscient newspaper. It is a gigantic newspaper in which every event down to the last detail—for tomorrow, the next day, and the next day, and on and on into indefinite time—is carefully inscribed. (see Kemeny, 1959, p. 39). There is nothing unknown. And, of course, there is no randomness. Randomness is, as it were, ignorance, in this view.

Taking a cue from this argument, we define randomness in a backhand way. We say events are random if we cannot predict their outcomes. For instance, there is no known way to win a penny-tossing game. Whenever there is no system for playing a game that ensures our winning (or losing), then the event-outcomes of the game are random. More formally put, *randomness* means that there is no known law, capable of being expressed in language that correctly describes or explains events and their outcomes. In a word, when events are random we cannot predict them individually. Strange to say, however, we can predict them quite successfully in the aggregate. That is, we can predict the outcomes of large numbers of events. We cannot predict whether a coin tossed will be heads or tails. But, if we toss the coin 1000 times, we can predict, with considerable accuracy, the total numbers of heads and tails.

An Example of Random Sampling

To give the reader a feeling for randomness and random samples, we now do a demonstration using a table of random numbers. A table of random numbers contains numbers generated mechanically so that there is no discernible order or system in them. It was said above that if events are random they cannot be predicted. But now we are going to predict the *general nature* of the outcomes of our experiment. We select, from a table of random digits, 10 samples of 10 digits each. Since the numbers are random, each sample "should" be representative of the universe of digits. The universe can be variously defined. We simply define it as the complete set of digits in the Rand Corporation table of random digits.[1] We now draw samples from the table. The means of the 10 samples will, of course, be different. However, they should

fluctuate within a relatively narrow range, with most of them fairly close to the mean of all 100 numbers and to the theoretical mean of the whole population of random numbers. The number of even numbers in each sample of 10 should be approximately equal to the number of odd numbers—though. There will be fluctuations, some of them perhaps extreme but most of them comparatively modest. The samples are given in Table 5. l.

TABLE 5.1 Ten Samples of Random Numbers

	1	2	3	4	5	6	7	8	9	10
	9	0	8	0	4	6	0	7	7	8
	7	2	7	4	9	4	7	8	7	7
	6	2	8	1	9	3	6	0	3	9
	7	9	9	1	6	4	9	4	7	7
	3	3	1	1	4	1	0	3	9	4
	8	9	2	1	3	9	6	7	7	3
	4	8	3	0	9	2	7	2	3	2
	1	4	3	0	0	2	6	9	7	5
	2	1	8	8	4	5	2	1	0	3
	3	1	4	8	9	2	9	3	0	1
M	5.0	3.9	5.3	2.4	5.7	3.8	5.2	4.4	5.0	4.9

Total Mean = 4.56

The means of the samples are given below each sample. The mean of *U*, the theoretical mean of the whole population of Rand random numbers, {0,1, 2, 3, 4, 5, 6, 7, 8, 9}, is 4.5. The mean of all 100 numbers, which can be considered a sample of *U*, is 4.56. This is, of course, very close to the mean of *U*. It can be seen that the means of the 10 samples vary around 4.5, the lowest being 2.4 and the highest 5.7. Only two of these means differ from 4.5 by more than 1. A statistical test (later we will learn the rationale of such tests) shows that the 10 means do not differ from each other significantly. (The expression "do not differ from each other significantly" means that the differences are not greater than the differences that would occur by chance.) And by another statistical test nine of them are "good" estimates of the population mean of 4.5 and one (2.4) is not.

Changing the sampling problem, we can define the universe to consist of odd and even numbers. Let's assume that in the entire universe there is an equal number of both. In our sample of 100 numbers there should be approximately 50 odd and 50 even numbers. There are actually 54 odd and 46 even numbers. A statistical test shows that the deviation of 4 for odd and 4 for even does not depart significantly from chance expectation.[1]

Similarly, if we sample human beings, then the numbers of men and women in the samples should be approximately in proportion to the numbers of men and women in the population, if the sampling is random and the samples are large enough. If we measure the intelligence of a sample, and the mean intelligence score of the population is 100, then the mean of the sample should be close to 100. Of course, we must always bear in mind the possibility of selection of the deviant sample, the sample with a mean, say, of 80 or less or 120 or more. Deviant samples do occur, but they are less likely to occur. The reasoning is similar to that for coin-tossing demonstrations. If we toss a coin three times, it is less likely that 3 heads or 3 tails will turn up than it is that 2 heads and 1 tail or 2 tails and 1 head will turn up. This is because *U* = {*HHH*, *HHT, HTH, HTT, THH, THT, TTH*, *TTT*}. There is only one *HHH* point and one *TTT* point, while there are three points with two H's and three with two T's.

RANDOMIZATION

Suppose an investigator wishes to test the hypothesis that counseling helps underachievers. The test involves using two groups of underachievers, one to be counseled, one not to be counseled. Naturally, the wish is to have the two groups equal in other independent variables that may have a possible effect on achievement. One way this can be done is to assign the children to both groups at random by, say, tossing a coin for each child. The child is assigned to one group if the toss is heads and to the other group if the

1. The source of random numbers used was: Rand Corporation, *A Million Random Digits with 700,000 Normal* Deviates (1955). This is a large and carefully constructed table of random numbers. These numbers were *not* computer generated. There are many other such tables however, that are good enough for most practical purposes. Modem statistics texts have such tables. Appendix C at the end of this book contains 4,000 computer-generated random numbers.

1. The nature of such statistical tests, as well as the reasoning behind them. will be explained in detail in Part Four. The student should not be too concerned if he does not completely grasp the statistical ideas expressed here. Indeed, one of the purposes of this chapter is to introduce some of the basic elements of such ideas.

toss is tails. (Note that if there were three experimental groups coin tossing would probably not be used. A 6-sided die may be used. Outcomes of 1 or 2 would assign that child to group 1. Outcomes of 3 and 4 would put the child in group 2 and the outcomes 5 and 6 would designate the child to be in group 3. Or a table of random numbers can be used to assign the children to groups. If an odd number turns up, assign a child to one group, and if an even number turns up, assign the child to the other group. The investigator can now assume that the groups are approximately equal in all possible independent variables. The larger the groups, the safer the assumption. Just as there is no guarantee, however, of not drawing a deviant sample, as discussed earlier, there is no guarantee that the groups are equal or even approximately equal in all possible independent variables. Nevertheless, it can be said that the investigator has used randomization to equalize the groups, or, as it is said, to control influences on the dependent variable other than that of the manipulated independent variable. Although we will use the term "randomization." a number of researchers prefer to use the words "random assignment." The procedure calls for assigning participants to experimental conditions on a random basis. While some believe that random assignment removes variation, in reality it only distributes it.

An "ideal" experiment is one in which all the factors or variables likely to affect the experimental outcome are controlled. If we knew all these factors, in the first place, and *could* make efforts to control them, in the second place, then we might have an ideal experiment. However, the sad case is that we can never know all the pertinent variables nor could we control them even if we did know them. Randomization, however, comes to our aid.

Randomization is the assignment to experimental treatments of members of a universe in such a way that, for any given assignment to a treatment, every member of the universe has an equal probability of being chosen for that assignment. The basic purpose of random *assignment,* as indicated earlier, is to apportion subjects (objects, groups) to treatments. Individuals with varying characteristics are spread approximately equally among the treatments so that variables that might affect the dependent variable other than the experimental variables have "equal" effects in the different treatments. There is no guarantee that this desirable state of affairs will be attained, but it is more likely to be attained with randomization than otherwise. Randomization also has a statistical rationale and purpose. If random assignment has been used, then it is possible to distinguish between systematic or experimental variance and error variance. Biasing variables are distributed to experimental groups according to chance. The tests of statistical significance that we will discuss later logically depend on random assignment. These tests are used to determine whether the observed phenomenon is statistically different from chance. Without random assignment the significance tests lack logical foundation. The idea of randomization seems to have been discovered or invented by Sir Ronald Fisher (see Cowles, 1989). It was Fisher who virtually revolutionized statistical and experimental design thinking and methods using random notions as part of his leverage. He has been referred to as "the father of analysis-of-variance." In any case, randomization and what can be called the principle of randomization is one of the great intellectual achievements of our time. It is not possible to overrate the importance of both the idea and the practical measures that come from it to improve experimentation and inference.

Randomization can perhaps be clarified in two or three ways: by stating the principle of randomization, by describing how one uses it in practice, and by demonstrating how it works with objects and numbers. The importance of the idea deserves all three.

The *principle of randomization* may be stated as the following. Since, in random procedures, every member of a population has an equal chance of being selected, members with certain distinguishing characteristics—male or female, high or low intelligence, conservative or liberal, and so on and on—will, if selected, probably be offset in the long run by the selection of other members of the population with counterbalancing quantities or qualities of the characteristics. We can say that this is a practical principle of what usually happens; we cannot say that it is a law of nature. It is simply a statement of what most often happens when random procedures are used.

We say that subjects are assigned at random to experimental groups, and that experimental treatments are assigned at random to groups. For instance, in the example cited above of an experiment to test the effectiveness of counseling on achievement, subjects can be assigned to two groups at random by using random numbers or by tossing a coin. When the subjects have been so assigned, the groups can be randomly designated as experimental and control groups using a similar procedure. We will encounter a number of examples of randomization as we go along.

A Senatorial Randomization Demonstration

To show how, if not why, the principle of randomization works, we now set up a sampling and design experiment. We have a population of 100 members of the United States Senate from which we can sample. In this population (in 1993), there are 56 Democrats and 44 Republicans. We have selected two important votes, one (Vote 266) an amendment to prohibit higher grazing fees and the other (Vote 290) an amendment concerning funding for abortions. The data used in this example are from the 1993 Congressional Quarterly. These votes were important since each of them reflected presidential proposals. A Nay vote on 266 and a Yea vote on 290 indicates support of the President. Here we ignore their substance and treat the actual votes, or rather, the senators who cast the votes, as populations from which we sample.

We pretend we are going to do an experiment using three groups of senators, with 20 in each group. The nature of the experiment is not too relevant here. We want the three groups of senators to be approximately equal in all possible characteristics. Using a computer program written in BASIC we generated random numbers between 1 and 100. The first 60 numbers drawn with no repeated numbers (sampling without replacement), were recorded in groups of 20 each. Political party affiliation for Democrats and Republicans are noted with the senator's name. Also included are the senators' votes on the two issues, Y = Yea and N = Nay. These data are listed in Table 5.2

How "equal" are the groups? In the total population of 100 senators, 56 are Democrats and 44 are Republicans, or 56 percent and 44 percent. In the total sample of 60 there are 34 Democrats and 26 Republicans, or 57 percent for Democrats and 43 percent for Republicans. There is a difference of 1 percent from the expectation of 56 percent and 44 percent. The obtained and expected frequencies of Democrats in the three groups and the total sample are given in Table 5.3. The deviations from expectation are small. The three groups are not exactly "equal" in the sense that they have equal numbers of Republican senators and Democrats. The first group had 11 Democrats and 9 Republicans, the second group has 10 Democrats and 10 Republicans and the third group has 13 Democrats and 7 Republicans. This not an "unusual" outcome that happens with random sampling. Later we will see that the discrepancies are not different statistically.

Table 5.2. Senatorial Vote per groups of n = 20 on Senate Vote 266 and 290.

#	Name-party	266	290
73	hatfield-r	y	n
27	coats-r	y	n
54	kerrey-d	y	y
93	murray-d	n	y
6	mccain-r	y	n
26	simon-d	n	y
7	bumpers-d	n	y
81	daschle-d	n	y
76	specter-r	n	y
38	cohen-r	n	y
32	kasselbaum-r	y	n
44	riegel-d	n	y
98	kohl-d	n	y
77	pell-d	n	y
61	bingaman-d	y	n
16	roth-r	n	n
24	kempthorne-r	y	n
100	wallop-r	y	n
15	biden-d	n	n
14	lieberman-d	n	y

#	Name-party	266	290
78	chafee-r	y	y
20	coverdell-r	y	n
25	mosley-brown-d	n	y
42	kerry-d	n	y
68	dorgan-d	y	n
57	gregg-r	y	n
11	campbell-d	y	y
31	dole-r	y	n
37	mitchell-d	n	y
30	grassley-r	y	n
22	inouye-d	y	y
99	simpson-r	y	n
8	pryor-d	n	?
4	stevens-r	y	n
23	craig-r	y	n
12	brown-r	y	n
10	feinstein-d	y	y
87	bennett-r	y	n
19	nunn-d	n	n
45	wellstone-d	n	y

#	Name-party	266	290
58	smith-r	y	n
95	byrd-d	n	n
83	matthews-d	y	n
52	burns-r	y	n
80	thurmond-r	y	n
13	dodd-d	y	y
88	hatch-r	y	n
63	moynihan-d	y	y
89	leahy-d	n	y
75	wofford-d	n	y
92	warner-r	y	n
91	robb-d	n	y
34	mcconnell-r	y	n
96	rockefeller-d	n	y
28	lugar-r	y	n
43	levin-d	n	y
59	bradley-d	n	y
69	glenn-d	n	y
9	boxer-d	n	y
67	conrad-d	y	n

TABLE 5.3 Obtained and Expected Frequencies of Political Party (Democrats) in Random Samples of 20 U.S. Senators [a]

	Groups			Total
	I	1I	III	
Obtained	11	10	13	34
Expected [b]	11.2	11.2	11.2	33.6
Deviation	.2	1.2	−1.8	−.4

[a] only the larger of the two expectations of the Republican-Democrat split, the Democrats (.56), is reported.

[b] The expected frequencies were calculated as follows: 20 ×.56= 11.2. Similarly, the total is calculated:

$$60 \times .56 = 33.6.$$

Remember that we are demonstrating both random sampling and randomization, but especially randomization. We therefore ask whether the random assignment of senators to the three groups has resulted in "equalizing" the groups in all characteristics. We can never test all characteristics, of course; we can only test those available. In the present case we have only political party affiliation, which we tested above, and the votes on the two issues: prohibition to increase grazing fees (Vote 266) and prohibition of funds for certain types of abortions (Vote 290). How did the random assignment work with the two issues? The results are presented in Table 5.4. The original vote on Vote 266 of the 99 senators who voted were 59 Yeas and 40 Nays. These total votes yield expected Yea frequencies in the total group of 59 ÷ 99 = .596, or 60 percent. We therefore expect 20 × .60 = 12, in each experimental group. The original vote of the 99 senators who voted on Vote 290 was 40 Yeas, or 40 percent (40 ÷99 = .404). The expected group Yea frequencies, then, are: 20 × .40 = 8. The obtained and expected frequencies and the deviations from expectation for the three groups of 20 senators and for the total sample of 60 on Vote 266 and 290 are given in Table 5.4.

TABLE 5.4 Obtained and Expected Frequencies on Yea Votes on 266 and 290 in Random Groups of Senators

	Groups					
	I		II		III	
	266	290	266	290	266	290
Obtained	9	10	13	9	11	9
Expected[a]	12	8	12	8	12	8
Deviation	3	−2	−1	−1	1	−1

	Total	
	266	290
Obtained	33	28
Expected[a]	36	24
Deviation	3	−4

[a]The expected frequencies were calculated for Group I, Issue 266, as follows: there were 59 Yeas of a total of 99 votes or 59/99 = .60; 20 × .60 = 12. For the total group, the calculation is: 60 × .60 = 36.

It appears that the deviations from chance expectation are all small. Evidently the three groups are approximately "equal" in the sense that the incidence of the votes on the two issues is approximately the same in each of the groups. The deviations from chance expectation of the Yea votes (and, of course, the Nay votes) are small. So far as we can see, the randomization has been "successful." This demonstration can also be interpreted as a random sampling problem. We may ask, for example, whether the three samples of 20 each and the total sample of 60 are representative. Do they accurately reflect the characteristics of the population of 100 senators? For instance, do the samples reflect the

proportions of Democrats and Republicans in the Senate? The proportions in the samples were .55 and .45 (I), .50 and .50 (II), .65 and .35 (III). The actual proportions are .56 and .44. Although there is a 1, 6 and 9 percent deviations in the samples, the deviations are within chance expectation. We can say, therefore, that the samples are representative insofar as political party membership is concerned. Similar reasoning applies to the samples and the votes on the two issues.

We can now do our experiment believing that the three groups are "equal." They may not be, of course, but the probabilities are in our favor. And as we have seen, the procedure usually works well. Our checking of the characteristics of the senators in the three groups showed that the groups were fairly "equal" in political preference and Yea (and Nay) votes on the two issues. Thus we can have greater confidence that if the groups become unequal, the differences are probably due to our experimental manipulation and not to differences between the groups before we started.

However, no less an expert than Feller (1967), writes: "In sampling human populations the statistician encounters considerable and often unpredictable difficulties, and bitter experience has shown that it is difficult to obtain even a crude image of randomness." Williams (1978) presents a number of

examples where “randomization” does not work in practice. One such example that influenced the lives of a large number of men was the picking the military draft lottery numbers in 1970. Although it was never absolutely proven, the lottery numbers did not appear to be random. In this particular instance, the month and day of birth for all 366 days was each put into a capsule. The capsules went into a rotating drum. The drum was turned a number of times so that the capsules would be well mixed. The first capsule drawn had the highest draft priority, number 1. The second capsule drawn had the next highest and so forth. The results showed that the dates for later months had a lower median than earlier months. Hence men with later birthdates were drafted earlier. If the drawings were completely random, the medians for each month should have been much more equal. The point to made here is that many statistical analyses are dependent on successful randomization. To have one in practice is not such an easy task.

SAMPLE SIZE

A rough-and-ready rule taught to beginning students of research is: Use as large samples as possible. Whenever a mean, a percentage, or other statistic is calculated from a sample, a population value is being estimated. A question that must be asked is: How much error is likely to be in statistics calculated from samples of different sizes? The curve of Figure 5.1 roughly expresses the relations between sample size and error, error meaning deviation from population values. The curve says that the smaller the sample the larger the error, and the larger the sample the smaller the error.

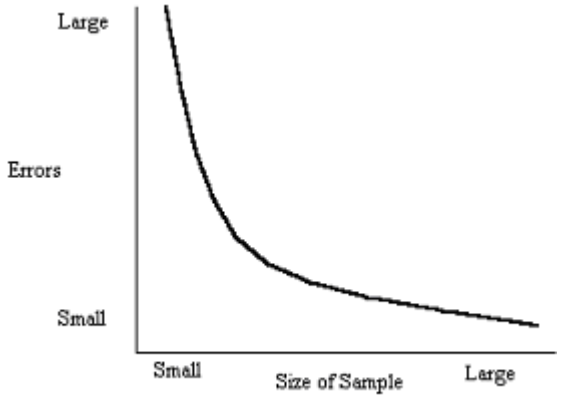


Figure 5.1

Consider the following rather extreme example. Global Assessment Scale (GAS) admission score and total days in therapy of 3166 Los Angeles County children seeking help at Los Angeles County Mental Health facilities from 1983 to 1988, were made available to the second author through the generosity of Dr. Stanley Sue. Dr. Sue is Professor of Psychology and Director of the National Research Center on Asian American Mental Health at the University of California, Davis. Dr. Sue granted the second author permission to use the data. The information of Table 5.5 and Table 5.6 were created from these data. We express our thanks and appreciation to Dr. Stanley Sue. The Global Assessment Scale, hereafter referred to as GAS, is a score assigned by a therapist to each client based on psychological, social and occupational functioning. The GAS score we will use here in our example is the GAS score the client received at the time of admission or first visit to the facility.

From this "population,” 10 samples of two children were randomly selected. The random

selection of these samples and of others were done using the "sample" function in SPSS (Statistical Package for the Social Sciences, Norusis, 1992). The sample means were computed by the "descriptive" routine in SPSS and are given in Table 5.5. The deviations of the means from the means of the population are also given in the table.

TABLE 5.5a Samples (n = 2) of GAS Scores of 3166 Children, Mean of the Samples, and Deviations of the Sample Means from the Population

GAS

Sample			Mean	Dev.
1	61	60	60.5	11.21
2	46	50	48	–1.29
3	65	35	50	0.71
4	50	55	52.5	3.21
5	51	55	53	3.71
6	35	50	42.5	–6.79
7	45	41	43	–6.29
8	44	47	45.4	–3.79
9	43	50	46.5	–2.79
10	60	55	57.5	8.21

Total Mean (20) = 49.9
Population Mean (3166) = 49.29

TABLE 5.5b Samples (n = 2) of Total Days in Therapy Scores of 3166 Children, Mean of the Samples, and Deviations of the Sample Means from the Population

Total Days in Therapy

Sample			Mean	Dev.
1	92	57	74.5	–9.04
2	9	58	33.5	–50.04
3	172	38	105	21.46
4	0	70	35	–48.54
5	3	603	303	219.46
6	151	110	125.5	41.96
7	28	0	15	–69.54
8	189	51	120	36.46
9	28	72	50	-33.54
10	17	398	207.5	123.96

Total Mean (20) = 106.80
Population Mean (3166) = 83.54

The GAS means range from 42.5 to 60.5, and the Total-days means from 14 to 303. The two total means (calculated from the 20 GAS and the 20 Total-days scores) are 49.9 and 106.8. These small sample means vary considerably. The GAS and Total-days means of the population (N = 3166) were 49.29 and 83.54. The deviations of the GAS means range a good deal: from -6.79 to 11.21. The Total-days deviations range from -69.54 to 219.46. With very small samples like these we cannot depend on any one mean as an estimate of the population value. However, we can depend more on the means calculated from all 20 scores, although both have an upward bias.

Four more random samples of 20 GAS and 20 total-days scores were drawn from the population. The four GAS and the four total-days means are given in Table 5.6. The deviations (Dev.) of each of the means of the samples of 20 from the population means are also given in the table, as well as the means of the sample of 80 and of the total population. The GAS deviations range from –.39 to 1.31, and the total-days deviations from –14.14 to 26.41. The mean of the 80 GAS scores is 49.68, and the mean of all 3166 GAS scores is 49.29. The comparable total-days means are 93.08 (n = 80) and 83.54 (N = 3166). These means are quite clearly much better estimates of the population means.

TABLE 5.6a Means and Deviations from Population Means of Four GAS Samples, n = 20, Sue Data

	Samples (n = 20)			
	1	2	3	4
GAS	49.35	48.9	49.85	50.6
Dev.	.06	–.39	.56	1.31
Total-Days	69.4	109.95	89.55	103.45
Dev.	–14.14	26.41	6.01	19.91

TABLE 5.6b Means and Deviations from Population Means of Total Sample, n = 80, and Population, N = 3166, Sue Data

	Total (n = 80)	Population (N = 3166)
GAS	49.68	49.29
Dev.	.385	
Total-Days	93.08	83.54
Dev.	9.54	

We can now draw conclusions. First, statistics calculated from large samples are more accurate (other things equal) than those calculated from small samples. A glance at the deviations of Tables 5.5 and 5.6 will show that the means of the samples of 20 deviated much less from the population mean than did the means of the samples of 2. Moreover, the means from the sample of 80 deviated little from the population means (.39 and 9.54).

It should now be fairly clear why the research and sampling principle is: Use large samples.[1] Large samples are not advocated just because large numbers are good in and of themselves. They are advocated in order to give the principle of randomization, or simply randomness, a chance to "work," to speak somewhat anthropomorphically. With small samples, the probability of selecting deviant samples is a greater than with large samples. For example, in one random sample of 20 senators drawn some years ago, the first 10 senators (of 20) drawn were all Democrats! Such a run of 10 Democrats is most unusual. *But it can and does happen.* Let's say we had chosen to do an experiment with only two groups of 10 each. One of the groups was the one with the 10 Democrats and the other had both Democrats and Republicans. The results could have been seriously biased, especially if the experiment had anything to do with political preference or social attitudes. With large groups, say 30 or more, there is less danger. Many psychology departments at major universities have a research requirement for students enrolled in an introductory psychology class. For such situations, it may be relatively easy to obtain large samples. However, certain research studies such as those found in human engineering or marketing research, the cost of recruiting subjects is high. Remember the Williams and Adelson study discussed by Simon (1987) in Chapter 1? So the rule of getting large samples may not be appropriate for all research situations. In some studies, 30 or more elements, participants or subjects may be too little. This is especially true in studies that are multivariate in nature. Comrey and Lee (1992) for example states that samples of size 50 or less give very inadequate reliability of correlation coefficients. Hence, it may be more appropriate to obtain an approximation to the sample size needed. The statistical determination of sample size will be discussed in Chapter 12 for the various kinds of samples.

1. The situation is more complex than this simple statement indicates. Samples that are too large can lead to other problems; the reasons will be explained in a later chapter.

KINDS OF SAMPLES

The discussion of sampling has until now been confined to simple random sampling. The purpose is to help the student understand fundamental principles; thus the idea of simple random sampling, which is behind much of the thinking and procedures of modern research, is emphasized. The student should realize, however, that simple random sampling is not the only kind of sampling used in behavioral research. Indeed, it is relatively uncommon, at least for describing characteristics of populations and the relations between such characteristics. It is, nevertheless, the model on which all scientific sampling is based.

Other kinds of samples can be broadly classified into probability and nonprobability samples (and certain mixed forms). *Probability samples* use some form of random sampling in one or more of their stages. *Nonprobability samples* do not use random sampling; they thus lack the virtues being discussed. Still, they are often necessary and unavoidable. Their weakness can to some extent be mitigated by using knowledge, expertise, and care in selecting samples and by replicating studies with different samples. It is important for the student to know that probability

sampling is not necessarily superior to nonprobability sampling for all possible situations. Also, probability sampling does not guarantee samples that are more representative of the universe under study. In probability sampling the emphasis is placed in the method and the theory behind it. With nonprobability sampling the emphasis relies on the person doing the sampling and that can bring with it entirely new and complicated batch of concerns. The person doing the sampling must be knowledgeable of the population to be studied and the phenomena under study.

One form of nonprobability sampling *is quota sampling.* Here the knowledge of the strata of the population—sex, race, region, and so on—is used to select sample members that are representative, "typical," and suitable for certain research purposes. A *strata* is the partitioning of the universe or population into two or more non-overlapping (mutually exclusive) groups. From each partition a sample is taken. Quota sampling derives its name from the practice of assigning quotas, or proportions of kinds of people, to interviewers. Such sampling has been used a good deal in public opinion polls. To perform this sampling correctly, the researcher would need to have a very complete set of characteristics for the population. Next, the researcher must know the proportions for each quota. After knowing this, the next step is to collect the data. Since the proportions might be unequal from quota to quota, the sample elements are assigned a weight. Quota sampling is difficult to accomplish because it requires accurate information on the proportions for each quota and such information is rarely available.

Another form of nonprobability sampling *is purposive sampling,* which is characterized by the use of judgment and a deliberate effort to obtain representative samples by including presumably typical areas or groups in the sample. Purposive sampling is used a lot in marketing research. To test the reaction of consumers to a new product, the researcher may give out the new product to people that fits the researcher's notion of what the universe looks like. Political polls are another example where purposive sampling is used. On the basis of past voting results and existing political party registration, in a given region, the researcher purposively selects a group of voting precincts. The researcher feels that this selection will match the characteristics of the entire electorate. A very interesting presentation of how this information was used to help elect a U.S. Senator in California is given in Barkan and Bruno (1972).

So-called *"accidental" sampling*, the weakest form of sampling, is probably also the most frequent. In effect, one takes available samples at hand; classes of seniors in high school, sophomores in college, a convenient PTA, and the like. This practice is hard to defend. Yet, used with reasonable knowledge and care, it is probably not as bad as it has been said to be. The most sensible advice seems to be: Avoid accidental samples unless you can get no others (random samples are usually expensive and in some situations, hard to come by). If you do use them, use extreme circumspection in analysis and interpretation of data.

Probability sampling includes a variety of forms. When we discussed simple random sampling, we were talking about one version of probability sampling. Some of the other common forms of probability sampling are stratified sampling, cluster sampling, two-stage cluster sampling, and systematic sampling. Some more unconventional methods would include the Bayesian approach or the sequential approach. The superiority of one method of sampling over another is usually evaluated in terms of the amount of reduced variability in parameters estimated and in terms of cost. Cost is sometimes interpreted as the amount of labor in data collection and data analysis.

In stratified sampling, the population is first divided into strata such as men and women, African-American and Mexican-American, and the like. Then random samples are drawn from each strata. If the population consists of 52 percent women and 48 percent men, a stratified sample of 100 participants would consist of 52 women and 48 men. The 52 women would be randomly chosen from the available group of women and the 48 men would be randomly selected from the group of men. This is also called proportional allocation. When this procedure is performed correctly it is superior to simple random sampling. When compared to simple random sampling, stratified sampling usually reduces the amount of variability and reduces the cost in data collection and analyses. Stratified sampling capitalizes on the between-strata differences. Figure 5.2 conveys the basic idea of stratified sampling. Stratified sampling adds control to the sampling process by decreasing the amount of sampling error. This design is recommended when the population is composed of sets of dissimilar groups. Randomized stratified sampling allows us to study the stratum differences. It allows us to give special attention to certain groups that would otherwise be ignored because of their size. Stratified random sampling is often accomplished through proportional allocation

procedures (PAP). When using such procedures, the sample's proportional partitioning resembles that of the population. The major advantage in using PAP is that it provides a "self-weighted" sample.

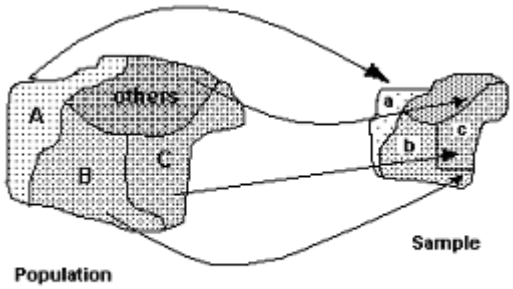


Figure 5.2 Stratified Sampling

Cluster sampling, the most used method in surveys, is the successive random sampling of units, or sets and subsets. A cluster can be defined as a group of things of the same kind. It is a set of sample elements held together by some common characteristic(s). In cluster sampling, the universe is partitioned into clusters. Then the clusters are randomly sampled. Each element in the chosen clusters is then measured. In sociological research, the investigator may use city blocks as clusters. City blocks are then randomly chosen and interviewers then talk to every family in each block selected. This type of cluster sampling is sometimes referred to as *area sampling*. If the researcher was to use simple random sampling or stratified random sampling, that person would need a complete list of families or household to sample from. Such a list may be very difficult to obtain for a large city. Even with such a list, the sampling costs would be high since it would involve measuring households over a wide area of the city. Cluster sampling is most effective if a large number of smaller size clusters are used. In educational research, for example, school districts of a state or county can be used as clusters and a random sample of the school districts is taken. Every school within the school district would be measured. However, school districts may form too large of a cluster. In this case using schools as clusters may be better.

In *two-stage cluster sampling*, we start with a cluster sampling as described above. Then instead of measuring every element of the clusters chosen at random, we select a random sample of the elements and measure those elements. In the educational example given above, we would identify each school district as a cluster. We would then randomly choose k school districts. From these k school districts, instead of measuring every school in the chosen districts (as in regular cluster sampling), we would take another random sample of schools within each district. We would then measure only those schools chosen.

Another kind of probability sampling—if indeed, it can be called probability sampling—is *systematic sampling*. This method is a slight variation of simple random sampling. This method assumes that the universe or population consists of elements that are ordered in some way. If the population consists of N elements and we want to choose a sample of size n, we first need to form the ratio $N \div n$. This ratio is rounded to a whole number, k and then used as the sampling interval. Here the first sample element is randomly chosen from numbers 1 through k and subsequent elements are chosen at every kth interval. For example, if the element randomly selected from the elements 1 through 10 is 6, then the subsequent elements are 16, 26, 36, and so on. The representativeness of the sample chosen in this fashion is dependent upon the ordering of the N elements of the population.

The student who will pursue research further should, of course, know much more about these methods. The student is encouraged to consult one or more of the excellent references on the subject presented at the end of this chapter. Williams (1978) gives an interesting presentation and demonstration of each sampling method using artificial data.

Another related topic to randomness and sampling is randomization or permutation tests. We will give this topic more extensive treatment when we discuss the data analysis for quasi-experimental designs. The proponent of this method in psychology and the behavioral sciences has been Eugene S. Edgington. Edgington (1980, 1996) advocates the use of approximate randomization tests to handle statistical analyses of data from non-random samples and single subject research designs. We can briefly mention here how this procedure works. Take Edgington's example of the correlating the IQ scores of foster parents and their adopted children. If the sample was not randomly selected, the sample may be biased in favor of parents who want to have their IQ and the IQ of their adopted child measured. It is likely that some foster parents will intentionally lower their scores to match those of their adopted child. One way of handling non-random data like this is to first compute the correlation between the parents and child. Then one would randomly pair the parents'' scores with the

child's'. That is, parent 1 may in a random pairing get matched up with the child from parent number 10. After such a random pairing, the correlation is again computed. If the researcher performs 100 such randomization pairings and computed the correlation each time, one can then compare the original correlation to the 100 created from random pairings. If the original correlation is the best (highest) the researcher would have a better idea that the correlation obtained may be credible. These randomizations or permutation tests have been quite useful in certain research and data analysis situations. It has been used to evaluate clusters obtained in a cluster analysis (see Lee and MacQueen, 1980). It has been proposed as a solution to analyzing self-efficacy data that are not independent (see Cervone, 1987).

Randomness, randomization, and random sampling are among the great ideas of science, as indicated earlier. While research can, of course, be done without using ideas of randomness, it is difficult to conceive how it can have viability and validity, at least in most aspects of behavioral scientific research. Modern notions of research design, sampling, and inference, for example, are literally inconceivable without the idea of randomness. One of the most remarkable of paradoxes is that through randomness, or "disorder," we are able to achieve control over the often-obstreperous complexities of psychological sociological and educational phenomena. We impose order, in short, by exploiting the known behavior of sets of random events. One is perpetually awed by what can be called the structural beauty of probability, sampling, and design theory and by its great usefulness in solving difficult problems of research design and planning and the analysis and interpretation of data.

Before leaving the subject, let's return to a view of randomness mentioned earlier. To an omniscient being, there is no randomness. By definition such a being would "know" the occurrence of any event with complete certainty. As Poincare (1952/1996) points out, to gamble with such a being would be a losing venture. Indeed, it would not be gambling. When a coin was tossed ten times, he or she would predict heads and tails with complete certainty and accuracy. When dice were thrown, this being would know infallibly what the outcomes will be. Every number in a table of random numbers would be correctly predicted! And certainly this being would have no need for research and science. What we seem to be saying is that randomness is a term for ignorance. If we, like the omniscient being, knew all the contributing causes of events, then there would be no randomness. The beauty of it, as indicated above, is that we use this "ignorance" and turn it to knowledge. How we do this should become more and more apparent as we go on with our study.

Study Suggestions

A variety of experiments with chance phenomena are recommended: games using coins, dice, cards, roulette wheels, and tables of random numbers. Such games, properly approached, can help one learn a great deal about fundamental notions of modern scientific research, statistics, probability, and, of course, randomness. Try the problems given in the suggestions below. Do not become discouraged by the seeming laboriousness of such exercises here and later on in the book. It is evidently necessary and, indeed, helpful occasionally to go through the routine involved in certain problems. After working the problems given, devise some of your own. If you can devise intelligent problems, you are probably well on your way to understanding.

1. From a table of random numbers draw 50 numbers, 0 through 9. (Use the random numbers of Appendix C, if you wish.) List them in columns of 10 each.

(a) Count the total number of odd numbers; count the total number of even numbers. What would you expect to get by chance? Compare the obtained totals with the expected totals.

(b) Count the total number of numbers 0, 1, 2, 3, 4. Similarly count 5, 6, 7, 8, 9. How many of the first group should you get? The second group? Compare what you do get with these chance expectations. Are you far off?

(c) Count the odd and even numbers in each group of 10. Count the two groups of numbers 0, 1, 2, 3, 4 and 5, 6, 7, 8, 9 in each group of 10. Do the totals differ greatly from chance expectations?

(d) Add the columns of the five groups of 10 numbers each. Divide each sum by 10. (Simply move the decimal point one place to the left.) What would you expect to get as the mean of each group if only chance were "operating"? What did you get? Add the five sums and divide by 50. Is this mean close to the chance

expectation? [Hint: To obtain the chance expectation, remember the population limits.]

2. This is a class exercise and demonstration. Assign numbers arbitrarily to all the members of the class from 1 through N, N being the total number of members of the class. Take a table of random numbers and start with any page. Have a student wave a pencil in the air and blindly stab at the page of the table. Starting with the number the pencil indicates, choose n two-digit numbers between 1 and N (ignoring numbers greater than N and repeated numbers) by, say, going down columns (or any other specified way). n is the numerator of the fraction n/N, which is decided by the size of the class. If N = 30, for instance, let n = 10. Repeat the process twice on different pages of the random numbers table. You now have three equal groups (if N is not divisible by 3, drop one or two persons at random). Write the random numbers on the blackboard in the three groups. Have each class member call out his height in inches. Write these values on the blackboard separate from the numbers, but in the same three groups. Add the three sets of numbers in each of the sets on the blackboard, the random numbers and the heights. Calculate the means of the six sets of numbers. Also calculate the means of the total sets.

   (a) How close are the means in each of the sets of numbers? How close are the means of the groups to the mean of the total group?
   (b) Count the number of men and women in each of the groups. Are the sexes spread fairly evenly among the three groups?
   (c) Discuss this demonstration. What do you think is its meaning for research?

3. In Chapter 6, it was suggested that the student generate 20 sets of 100 random numbers between 0 and 100 and calculate means and variances. If you did this, use the numbers and statistics in this exercise. If you did not, use the numbers and statistics of Appendix C at the end of the book.
   (a) How close to the population mean are the means of the 20 samples? Are any of the means "deviant"? (You might judge this by calculating the standard deviation of the means and adding and subtracting two standard deviations to the total mean.)
   (b) On the basis of (a), above, and your judgment, are the samples "representative"? What does "representative" mean?
   (c) Pick out the third, fifth, and ninth group means. Suppose that 300 subjects had been assigned at random to the three groups and that these were scores on some measure of importance to a study you wanted to do. What can you conclude from the three means, do you think?

4. Most published studies in the behavioral sciences and education have not used random samples, especially random samples of large populations. Occasionally, however, studies based on random samples are done. One such study is:

Osgood, D. W., Wilson, J. K., O'Malley, P. M., Bachman, J. G. & Johnston, L. D. (1996). Routine activities and individual deviant behavior. *American Sociological Review, 61*, 635-655.

This study is worth careful reading, even though its level of methodological sophistication puts a number of its details beyond our present grasp. Try not to be discouraged by this sophistication. Get what you can out of it, especially its sampling of a large population of young men. Later in the book we will return to the interesting problem pursued. At that time, perhaps the methodology will no longer appear so formidable. (In studying research, it is sometimes helpful to read beyond our present capacity-provided we don't do too much of it!)

Another study random samples from a large population is by

Voelkl, K. F. (1995). School warmth, student participation and achievement. *Journal of Experimental Education, 63*, 127-138.

In this study the researcher gives some detail about using a two-stage stratified random sampling

plan to measure student's perception of school warmth.

5. Random assignment of subjects to experimental groups is much more common than random sampling of subjects. A particularly good, even excellent, example of research in which subjects were assigned at random to two experimental groups, is:

Thompson, S. (1980). Do individualized mastery and traditional instructional systems yield different course effects in college calculus? *American Educational Research Journal, 17*, 361-375.

Again, don't be daunted by the methodological details of this study. Get what you can out of it. Note at this time how the subjects were classified into aptitude groups and then assigned at random to experimental treatments. We will also return to this study later. At that time, we should be able to understand its purpose and design and be intrigued by its carefully controlled experimental pursuit of a difficult substantive educational problem: the comparative merits of so-called individualized mastery instruction and conventional lecture-discussion-recitation instruction.

6. Another noteworthy example of random assignment is done in a study by

Glick, P., DeMorest, J. A., & Hotze, C. A. (1988). Keeping your distance: Group membership, personal space, and requests for small favors. Journal of Applied Social Psychology, 18, 315-330.

This study is noteworthy because the study takes place in a real setting outside the laboratories of the university. Also the participants are not necessarily university students. Participants in this study are people in a public eating area within a large indoor shopping mall. Participants were chosen and then assigned to one of six experimental conditions. This article is easy to read and the statistics analysis is not much beyond the level of elementary statistics.

7. Another interesting study that uses yet another variation of random sampling is

Moran, J. D. & McCullers, J. C. (1984). A comparison of achievement scores in physically attractive and unattractive students. *Home Economics Research Journal, 13*, 36-40.

In this study, the researchers randomly selected photographs from school yearbooks. These photographs were then randomly grouped into 10 sets of 16 pictures. Students who were not familiar with the students in the photos were then asked to rate each person in the photo in terms of attractiveness.

Special Note. In some of the above study suggestions and in those of Chapter 6, instructions were given to draw numbers from tables of random numbers or to generate sets of random numbers using a computer. If you have a microcomputer or have access to one, you may well prefer to generate the random numbers using the built-in random number generator (function) of the microcomputer. An outstanding and fun book to read and learn how to do this is Walter's (1997) "The Secret Guide to Computers." Walter shows you how to write a simple computer program using the BASIC language, the language common to most microcomputers. How "good" are the random numbers generated? ("How good?" means "How random?") Since they are produced in line with the best contemporary theory and practice, they should be satisfactory, although they might not meet the exacting requirements of some experts. In our experience, they are quite satisfactory, and we recommend their use to teachers and students.

Some books on sampling.

Babbie, E. R. (1990). *Survey research methods*. (2nd Ed.). Belmont, CA: Wadsworth.

Babbie, E. R. (1995). *The practice of social research, 7th ed*. Belmont, CA: Wadsworth.

Cowles, M. (1989). *Statistics in psychology: A historical perspective.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Deming, W. E. (1966). *Some theory of sampling*. New York: Dover.

Deming, W. E. (1990). *Sampling design in business research.* New York: Wiley.
Kish, L. (1953). Selection of the sample. In Festinger, L. & Katz, D. Eds., *Research methods in the behavioral sciences.* New York: Holt, Rinehart and Winston.
Kish, L. (1995*). Survey sampling.* New York: John Wiley.
Snedecor, G. & Cochran, W. (1989). *Statistical Methods.* (8th Ed). Ames, IA: Iowa State University Press. .
Stephan, F. & McCarthy, P. (1974). *Sampling opinions*. Westport, CT: Greenwood.
Sudman, S. (1976). *Applied sampling*. New York: Academic Press.
Warwick, D. & Lininger, D. (1975). *The sample survey: Theory and practice.* New York: McGraw-Hill.
Williams, B. (1978). *A sampler on sampling*. New York: John Wiley.

Chapter Summary

1. Sampling refers to taking a portion of a population or universe as representative of that population or universe.
2. Studies using samples are economical, manageable and controllable.
3. One of the more popular methods of sampling is *random sampling*.
4. *Random sampling* is that method of drawing a portion (or sample) of a population or universe so that each member of the population or universe has an equal chance of being selected
5. The researcher defines the population or universe. A sample is a subset of the population.
6. We can never be sure that a random sampling is representative of the population.
7. With a random sampling, the probability of selecting a sample with a mean close to the population mean is greater than the probability of selecting a sample with a mean not close to the population mean.
8. Nonrandom sampling may be biased and the chances increase that the sample mean will not be close to the population mean.
9. We say events are random if we cannot predict their outcomes.
10. Random assignment is another term for randomization. This is where participants are assigned to research groups randomly. It is used to control unwanted variances.
11. There are two types of samples: nonprobability and probability samples.
12. Nonprobability samples do not use random assignment, while probability samples do use random sampling.
13. Simple random sampling, stratified random sampling, cluster sampling and systematic sampling are four types of probability sampling.
14. Quota sampling, purposive sampling and accidental sampling are three types of nonprobability sampling.

Chapter 6

Ethical Considerations in Conducting Behavioral Science Research

Fiction and Reality

In the previous chapters we have discussed science and the variables involved in social and behavioral sciences. We have also introduced some of the basic statistical methods used to analyze the data gathered from such research studies. In the chapters following this one, we will be discussing the actual conduct of the research process. Before doing this, we must present a very important topic. This topic involves the issue of research ethics. Some books have placed this topic in the latter part of the book after the research plan and designs have been discussed. We feel that such a topic should be presented earlier. The student of research needs this information in order to design an ethically sound study using the methods given in chapters after this one. It would be ideal if the researcher read this chapter, then read the chapters on research design and then return to re-read the points made in this chapter.

What is research ethics? What is research? These are two terms that are difficult to define. Shrader-Frechette (1994) provides a definition by contrasting "research" from "practice." As we saw in an earlier chapter, research is an activity done to test theories, make inferences and add or update information on a base of knowledge. Professional practice does not usually involve testing theories or hypotheses but rather enhance the welfare of clients using actions and information that have been demonstrated to be successful. Some of these actions were established through earlier scientific research. Even though both "research" and "practice" have ethics, the ethics involved with the research process is directed toward the individuals who do research and their conduct of the research process. . Shrader-Frechette states that research ethics specifies the behavior researchers ought to show during the *entire* process of their investigation. Keith-Spiegel & Koocher (1985) discusses the ethics of psychological practice. Dawes (1994) gives a very critical view of the practice of psychology and psychotherapy. Part of Dawes' discussion concerns ethics of practice.

The discussion, emphasis and practice of research ethics are relatively recent events. Before the 20th century, those scientists who were caught experimenting on people without proper consent were punished. However, there were instances in history where the violations of research ethics yield fruitful results. When one thinks about the ethics of doing research on humans or animals one can not avoid mixed feelings. In examining history, there were those brave individuals like Edward Jenner who injected a child with a weaker form of the smallpox virus and in doing so developed a vaccine for small pox. History has it that Edward Jenner did not get permission from anyone before doing this. Or how about Dr. Barry Marshall who in order to show that the peptic ulcer was caused by bacteria and not acids swallowed the culture of bacteria himself and them successfully treated himself with dosages of anti-biotic drugs. Yet there are documented cases of tragic consequences for researchers who failed to follow proper ethical principles of research and those committing scientific fraud in science. Some of these are noted and discussed in Shrader-Frechette (1994). This is an excellent book worth reading as well as Miller & Hersen (1992) and Erwin, Gendin & Kleiman (1994). Evidence of known or suspected fraud can be traced to research done in ancient Greece.

In conducting research, the sensitive researcher is often confronted with ethical dilemmas. Prior to the 1960's researchers from all fields of science were left to their own conscious in terms of research ethics. The scholarly publications on the appropriate behavior of scientists provided some guidance, but none or few of the guidelines were mandated. An example exemplifying an ethical dilemma can be found in the fictional story of Martin Arrowsmith, the protagonist in Sinclair Lewis' novel *Arrowsmith*. In this novel, Dr. Martin Arrowsmith in a laboratory study discovers by accident a principle that was effective in destroying bacteria. Arrowsmith called it a "phage." When the bubonic plague breaks out in a third world country, Arrowsmith is sent to that country to help the afflicted and to test his phage. Arrowsmith was taught that the true effectiveness of the phage can be determined by giving it to half of the infected population. The other half would be given a placebo or no treatment at all. However, Arrowsmith upon seeing the alarming death rate, including the death of his wife and close friend decides to give the phage to the entire population. If he had followed his experimental plan and his phage was truly effective, the people receiving the phage would survive and those receiving the placebo would not. Arrowsmith's conscience would not allow him to deceive half of the population and let them die in the name of scientific research. He administered the phage to everyone. The plague did end after the natives were inoculated, but Arrowsmith never really knew whether

his phage was effective. Although this was fiction, actual research scientists are at times faced with similar dilemmas.

A Beginning?

It was studies done in the 1960's and 1970's where there was evidence of research fraud and deception of research participants that led to a demand for specific mandated rules for the conduct of research. In 1974, the United States Congress called for the creation of institutional review boards. The purpose of these boards was to review the ethical conduct of those research studies that had received federal research grants. Later in the 1980's federally funded research involving humans and animals were reviewed for both ethical acceptability and research design. By the 1980's many of the major universities in the United States had guidelines for dealing with misconduct in research. Other countries also began to put forth guidelines and rules. The governments of Sweden and the Netherlands required that independent review committees evaluate all biomedical studies. Shrader-Frechette (1994) describes two broad categories in terms of ethical problems in scientific research. The two major categories are (1) processes and (2) products. The research process is deemed harmful if participants do not give informed consent to the procedures used on them. The research process is also considered harmful if the participants were deceived or recruited using deceptive methods. The research product is harmful if the conduct of that research results in a harmful environment for anyone who comes into contact with it. Shrader-Frechette refers to this as "downwinders." The case of radiation poisoning due to scientific tests of nuclear weapons is an example of a research product that is harmful. Shrader-Frechette briefly describes this research and the consequences. Saffer & Kelly (1983) give a more complete account in an informative book titled "Countdown Zero." Saffer and Kelly describe how the atmospheric tests of the atomic bomb in the desert of Nevada in the late 1940's carried over into other parts of the desert. The crew, staff and actors in the movie *The Conqueror* were all exposed to radioactive sand during the filming of the movie in the desert. All of these people developed cancer and later died from cancer related illnesses. Some of the well-known actors and actresses included John Wayne, Susan Hayward and Dick Powell. Saffer and Kelly also describes how the United States military's research on how to fight a nuclear war in the 1950's led to the exposure of many military personnel to radiation fallout. Saffer himself was one of the soldiers who participated in such studies. He noticed after a number of years after leaving the service that fellow soldiers developed cancer.

One of the most infamous case on the unethical use of deception was the *Tuskegee Study* (see Brandt, 1978). In 1932 the U.S. Public Health Service did an experiment on 399 poor, semiliterate African-American males who had contracted syphilis. One purpose of this study was to examine the effects of syphilis on untreated individuals. In order to get afflicted African-American males to participate in the study, they were told they were being treated when in fact they were not. Symptoms of syphilis were periodically measured and recorded. Autopsies were performed on each individual after each death. It would be forty years later before the public became aware of this research. At the time of disclosure, the study was still in progress. The research was clearly unethical; one reason was because treatment was still being withheld from the survivors as late as 1972. They could have been effectively treated with penicillin that became available in the 1940s. One of the major outcries of unethical research behavior has been focused on the use of deception.

Deception is still used in some research studies today. However, the research is critically evaluated before it can be done. All major universities in the United States have a research ethics and human usage committee that screens and evaluates studies for potential deceptions and harmful effects. It is their task to make sure no harm is inflicted on any of the participants.

One of the most noted studies in psychology that used deception was conducted by social psychologist Stanley Milgram, who recruited participants in a "learning" experiment (see Milgrim, 1963). Those who volunteered were told that some would be teachers and the others would be learners. The teachers were in charge of teaching lists of words to the learners. The teachers were told to administer increasingly painful shocks every time the learners made an error. The real purpose of the experiment, however, was not to study learning but to study obedience to authority. Milgrim was particularly interested in whether there was any truth to the claims of Nazi war criminals who said they did the atrocious acts because they were "ordered" by their superiors to do so. Unknown to the participants, all participants served as "teachers." That is, all participants were told that they were teachers. None of the participants served as "learners." The learners were confederates of the experimenter. They pretended to be participants who were randomly chosen to serve as learners. Furthermore, there were actually no shocks administered at anytime. The teachers were tricked into believing that the learners' cries of pain and requests for assistance was real. When instructed to increase the severity of the shocks, some of the participants hesitated. However, when they were instructed by the experimenter to continue, they did so. They even continued "shocking' the learners beyond the point

where the learners "begged" to be released from the experiment. The results were, to Milgram as well to others, almost beyond belief. A great many subjects (the "teachers") unhesitatingly obeyed the experimenter's "Please continue" or "You have no choice, you must go on" and continued to increase the level of the shocks no matter how much the learner pleaded with the "teacher" to stop. What particularly surprised Milgram was that no one ever walked out of the laboratory in disgust or protest. This remarkable obedience was seen time and time again in several universities where the experiment was repeated. Public anger over this experiment centered on the deception that might have caused psychological discomforts and harm to the participants. More than that, some people overgeneralized and thought that many such psychological experiments were being conducted.

For years following the famous study, critics of his study repeatedly dogged Milgrim. There was very little publicity surrounding the fact that Milgrim did a number of follow-up studies on the participants and found that there were no negative effects. In fact, at the conclusion of each experimental session, the participants were debriefed and introduced to the "learner" to show that no dangerous electrical shocks were administered.

Another sensitive area has been one directed at fraud. This includes situations where the researcher altered data from a research study in order to show that a certain hypothesis or theory was true. Other cases of fraud involved the reporting of research findings for research studies that never took place. History shows that there have been a number of prominent individuals who have been involved in fraud (see Erwin, Gendin & Kleiman, 1994). One of the more sensational cases of alleged fraud comes from psychology. The person involved was Sir Cyril Burt, a prominent British psychologist who received knighthood for his work on statistics and the heritability of intelligence. His work was marked by the use of identical twins whose genetic composition was the most alike. Burt supposedly demonstrated that there was a strong genetic component to intelligence by examining the intelligence of twins who were raised together versus those who were separated at birth and hence were reared apart. The intention was to determine how much influence the environment or heredity had on intelligence. In the mid 1970's after the death of Burt, Leon Kamin (1974) reported that a number of the correlations that Burt reported were identical to the third decimal place. By chance alone this was highly improbable. Later it was discovered that a few of Burt's co-authors on research articles published around the Second World War could not be found. Many of Burt's critics felt that Burt created these co-authors in order to mislead the scientific community. Even Leslie Hearnshaw, who was commissioned by Burt's family to write a biography of Burt, claimed to have found evidence of fraud. This particular view of Burt's fraud is detailed in Gould (1981). However, Jensen (1992) presents a different socio-historical view of Burt. Jensen states that the charges against Burt were never adequately proven. Jensen also gives information concerning Burt that was never mentioned in Gould's book or in other publications that were critical of Burt.

Such events as the Tuskegee, Milgram and Burt brought about the creation of laws and regulations to restrict or stop unethical research behavior in the medical, behavioral and social sciences. Professional organizations, such as the American Psychological Association and the American Physiological Society developed commissions to investigate and recommend action on reported cases of unethical research behavior. However, the reported incidence of unethical research by scientists has been minimal. Amongst the cases that have received the most negative publicity in behavioral science research involved Steven Breuning of the University of Pittsburgh. Breuning was convicted in 1988 of fabricating scientific data about drug tests (Ritalin and Dexedrine) on hyperactive children. Breuning's falsified results were widely cited and influenced several states to change their regulations on the treatment of these children. The Breuning case illustrates how dangerous the fraudulent behavior of a scientist can be.

In the physical sciences and medicine, Maurice Buchbinder, a cardiologist, was broached for research problems associated with his testing of the Rotablator. This device is a coronary vessel-cleaning device. Investigation revealed that the device was manufactured by a company where Buchbinder owned millions of dollars in stock. Amongst his ethical violations were (1) the failure to conduct follow-up examinations on about 280 patients, (2) the improper use of the device on patients with severe heart disease, and (3) not properly reporting some of the problems experienced by patients.

Douglas Richman was another research physician who received notoriety in his study of a new hepatitis treatment drug. Richman was cited for failing to report the death of the patients in the study, failing to inform the drug's manufacturer about the serious side effects, and failing to properly explain risks to patients in the study. Even though the reported incidence of fraud and unethical behavior by scientists is scarce, Shrader-Frechette (1994) has pointed out that many unethical behaviors go unnoticed or unreported. Even research journals do not mention anything about requiring the author to present information that the study was done ethically (e.g. with informed consent). It is possible that

when a researcher studies the behavior of humans, that those humans are put at risk through coercion, deception, violation of privacy, breaches of confidentiality, stress, social injury and failure to obtain free informed consent.

Some General Guidelines

The following guidelines are summaries taken from Shrader-Frechette's excellent book. Shrader-Frechette lays down the codes that should be followed by researchers in all areas of study where animals and humans are used as participants. One of the topics centers on situations where the researcher should not perform the research study. There are five general rules to follow when determining that the research should not be done.

- Scientists should not do research that puts people at risk
- Scientists should not do research that violates the norms of free informed consent.
- Scientists should not do research that converts public resources to private gains.
- Scientists should not do research that could seriously damage the environment.
- Scientists ought not do biased research.

In the fifth and last point made by Shrader-Frechette, the implication is toward racial and sexual biases only. One should realize that in all research studies there are biases inherent in the research design itself.

However, one of the major criterion in deciding the execution of a research study is the consequences from that study. Shrader-Frechette states that there are studies that will put humans and animals at risk, but the non-execution of that research may lead to even greater risks to humans and animals. In different words, not all potentially dangerous research should be condemned. Shrader-Frechette states:

"Just as scientists have a duty to do research but to avoid ethically questionable research, so also they have a responsibility not to become so ethically scrupulous about their work that they threaten the societal ends research should serve." (p. 37).

Hence, the researcher must exercise some degree of common sense when deciding to do or not do the research study involving human and animal participants.

Guidelines from the American Psychological Association.

In 1973 the American Psychological Association published ethical guidelines for psychologists. The original guidelines went through a number of revisions since then. The latest guidelines and principles were published in the *American Psychologist* in the March 1990 issue. *The Ethical Principles of Psychologists and Code of Conduct* can be found in the 1994 edition of the *Publication Manual of the American Psychological Association*. The following section will give a brief discussion of the ethical principles and codes that are relevant of behavioral science research. These guidelines are directed toward both human and animal research. All persons working on a research project are bounded by the ethic codes regardless if they are a professional psychologist or a member of the American Psychological Association.

General Considerations

The decision to undertake a research project lies solely with the researcher. The first question that the researcher needs to ask oneself is whether or not the research is worth doing. Will the information obtained from the study be valuable and useful for science and human welfare? Will it help improve the human health and welfare? If the researcher feels that the research is worthwhile, then the research is to be conducted with respect and concern for the welfare and dignity of the participants.

The Participant at Minimal Risk.

One of the major consideration on whether the study should be conducted or not is the decision as to whether the participant will be a "subject at risk" or a "subject at minimal risk." If there is the possibility of serious risk for the participant, the possible outcome of the research should indeed be of considerable value before proceeding. Researchers in this category should consult with colleagues before continuing. At most universities, there is a special committee that reviews research proposals to determine if the value of the research is worth placing participants at risk. At all times, the researcher must take steps to prevent harm to the participant. Student research projects should be conducted with the minimum amount of risk on the participants.

Fairness, Responsibility and Informed Consent.

Prior to participation, the researcher and the participant should enter into an agreement that clarifies the obligation and responsibilities. With certain studies this involves informed consent. Here the participant agrees

to tolerate deception, discomfort, and boredom for the good of science. In return, the experimenter guarantees the safety and well being of the participant. Psychological research differs from medical research in this regard. Medical research ethics requires the researcher to inform the participant what will be done to them and for what purpose. Most behavioral and social science research is not this restrictive. The behavioral science researcher needs to tell only those aspects of the study that may influence the participant's willingness to participate. Informed consent is not required in minimal-risk research. Still, it is a good idea for the investigator in all research to establish a clear and fair agreement with research participants prior to their participation.

Deception

Demand characteristics exists with many behavioral science studies. Participants volunteer to participate with the knowledge that something not harmful will be done to them. Their expectations and their desire to "do what the researcher wants" could influence the outcome of the study. Hence validity of the study would be compromised. The famous Hawthorne study is a case in point. In the Hawthorne study, the factory workers were told ahead of time that some people will be coming to the factory to do a study on worker productivity. The workers knowing that they will be studied for their productivity behaved in ways that they would not normal behave, i.e. punctual, working harder, shorter breaks, etc. As a result the investigators in the study were unable to get a true measure of worker productivity. Hence, enters deception. Like a magic show, the participants' attentions are misdirected. If the investigators had entered the factory as "ordinary" workers, they could have obtained a clearer picture of worker productivity.

If the researcher can justify that deception is of value, and if alternative procedures are not available, then the participant must be provided with a sufficient explanation as soon as possible after the end of the experiment. This explanation is called ***debriefing***. Any deceptive procedure that presents the participants with a negative perception of themselves must be avoided.

Freedom from Coercion

Participants must always be made to feel that they can withdraw from the study at any time without penalty or repercussions. Participants need to be informed of this prior to the beginning of the experimental session. The researcher at a university that uses introductory psychology students as participants should make it clear that their participation is voluntary. At some universities, the introductory psychology course has a graded research component. This component cannot be based solely on participation in research studies. For those that wish so, the research component can be satisfied through other means such as a research paper. Giving extra credit points for participation can be conceived as coercion.

Debriefing

After collecting the data from the participant the nature of the research is carefully explained to the participant. Debriefing is an attempt to remove any misconceptions the participant may have about the study. This is an extremely important element in conducting a research study. Even the explanation of the study should not be aversive. It needs to be worded in such a way so that those that have just been deceived would not feel foolish or stupid or embarrassed. In the event of student researchers, it would be beneficial for both the researcher and participant to review the data. The debriefing session could be used as a learning experience so that the student participant can become more knowledgeable about behavioral science research. Showing a student around the laboratory and explaining the apparatus is also advised if time permits.

For those studies where an immediate debriefing would compromise the validity of the study, the researcher may delay debriefing. However, the researcher must make all possible attempts to contact the participant once the entire study (data collection) is completed.

Protection of Participants

The researcher must inform the participant of all risks and dangers inherent in the study. The researcher should realize that the participant is doing the researcher a favor by participating. Participation in any research may produce at least some degree of stress. Additionally, the researcher is obligated to remove any undesirable consequences of participation. This is relevant in the case where participant were placed in a "do nothing" or control group. It would be unethical in a study that examines pain-management programs to place people who are in chronic pain into a control group where they received no treatment.

Confidentiality

The issue of keeping the participant from harm includes confidentiality. The researcher must assure the participant that the data collected from them will be safeguarded. That is, the information collected from the

participant will not be disclosed to the public in a way that could identify the participant. With sensitive data, the researcher must inform the participant how the data will be handled. In one study dealing with sexual behavior and AIDs, participants were asked to fill out the questionnaire, place the questionnaire in an unmarked envelope and deposit the envelope in a sealed box. The researcher assured the participants that the questionnaire would only be seen by data entry people who don't know who they are and cannot guess who they are. Smith & Garner (1976) for example took extra precautions to assure the anonymity of participants in their study of homosexual behavior among male college athletes.

ETHICS OF ANIMAL RESEARCH

To some people, the use of animals in research is inhumane and not necessary. However, research studies using animals have provided a number of worthwhile advancements for both animals and human. Neal Miller (1985) has noted the major contributions that animal research has provided for society. Unlike human participants, animals do not volunteer. Contrary to the belief of animal right activists, very few studies of today involve inflicting animals with pain. Experiments using animals as participants are generally permissible so long as the animals are treated humanely. APA provides guidelines on the use of animals for behavioral research and also logistic guidelines for their housing and care. There are eleven major points covered in APA's guidelines. They will be listed here:

1. General.: This involved the code behind the acquisition, maintenance and disposal of animals. The emphasis is on the familiarity with the code.
2. Personnel: This point involves the people who will be caring for the animals. This includes the availability of a veterinarian and supervisor of the facility.
3. Facilities: The housing of the animals must conform to the standard set by National Institute of Health (NIH) for the care and use of laboratory animals.
4. Acquisition of Animals: This point deals with how the animals are acquired. Covered are the rules for breeding and/or purchasing of animals.
5. Care and Housing of Animals: This deals with the condition of the facilities where the animals are kept.
6. Justification of Research: The purpose of the research using animals must be stated clearly.
7. Experimental Design: The research design of the study should include humane considerations. This would include the type of animal and how many animals.
8. Experimental Procedure. All experimental procedures must take into consideration the animal's well being. Procedures must involve no inflicted pain. Any amount of induced pain must be justified by the value of the study. Any aversive stimuli should be set at the lowest level possible.
9. Field Research. Researchers doing field research should disturb the population as little as possible. There must be respect for property and privacy of inhabitants
10. Educational Use of Animals. Alternative non-animal studies should be considered. Classroom demonstrations using animals should be used when educational objectives can not be made through the use of media. Psychologists need to include a presentation on the ethics of using animals in research.
11. Disposition of Animals. This point deals with what to do with the animal once the study is finished.

These guidelines (available from the American Psychological Association) should be made known to all personnel involved in the research and conspicuously posted wherever animals are maintained and used.

In assessing the research, the possibility of increasing knowledge about behavior, including benefit for health or welfare of humans and animals, should be sufficient to outweigh any harm or distress to the animals. Humane consideration for the well being of the animals should thus always be kept uppermost in mind. If the animal is likely to be subjected to distress or pain, the experimental procedures specified in the guidelines of the American Psychological Association should be carefully followed, especially for surgical procedures. No animal should be discarded until its death is verified, and it should be disposed of in a manner that is legal and consistent with health, environmental, and aesthetic concerns.

A recent book by Shapiro (1998) presents the history and current status on the use of animals in scientific research. This book contains articles that deal with ethics and situations when animal research is necessary and when it is not.

Study Suggestions

1. Some feel that society have placed too many restrictions of 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, wanted to determine if the way a woman dresses influences men's behavior. She plans to attend two bars at night. In one bar she will be dressed provocatively and in the other she will be dressed in a business suit. Her dependent variable is the number of men who would approach and talk to her. Can you find some ethical problems with this study?
4. Go to your library and find other incidences of fraud and unethical behavior of 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 one of the following articles:

Braunwald, E. (1987). On analyzing scientific fraud. *Nature, 325*, 215-216.

Brody, R. G. & Bowman, L. (1998). Accounting and psychology students' perceptions of whistle blowing. *College Student Journal. 32*, 162-166. Does the college curriculum need to include ethics?

Broad, W. J. & Wade, N. (1982). *Betrayers of the truth*. New York: Touchstone.

Fontes, L. A. (1998). Ethics in family violence research: Cross-cultural issues. *Family Relations*: *Interdisciplinary Journal of Applied Family Studies, 47*, 53-61.

Herrmann, D. & Yoder, C. (1998). The potential effects of the implanted memory paradigm on child subjects. *Applied Cognitive Psychology, 12*, 198-206. (Discusses the danger of implanted memory.)

Knight, J. A. (1984). Exploring the compromise of ethical principles in science. *Perspectives in Biology and Medicine, 27*, 432-442. (Explores the reasons for fraud and dishonesty in science.)

Stark, C. (1998). Ethics in the research context: Misinterpretations and misplaced misgivings. *Canadian Psychology. 39*,.202-211. (A look at the ethic codes for the Canadian Psychological Association.)

tChapter Summary

1. The Tuskegee and Milgrim studies used a form of deception and are often cited as reasons why scientific research with humans and animals need to be regulated.
2. Fraud is also an issue of concern, since the work of individuals such as Burt and Breuning had a lot of influence on legislation and how people perceived themselves and others.
3. Organizations such as the American Psychological Association have set up guidelines on the ethics of doing research. Also they have set up review boards to evaluate and take action on claims of ethical misconduct.
4. Researchers are obligated to do no physical or psychological harm to research participants.
5. Researchers must do research in a way that will produce useful information.
6. The ten ethical standards set up by the American Psychological Association include provisions for planning the research, protection of participants, confidentiality, debriefing, deception, informed consent, and freedom from coercion
7. Guidelines are also provided for the use of animals in research. These provide guidelines on the care, feeding and housing of animals and what to do with animals after the end of the study.

Chapter 7

Research Design: Purpose and Principles

Research Design is the plan and structure of investigation so conceived 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 "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 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 helps them 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. 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 directions 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 variables attribute. We can then act to manipulate the active variables and to categorize and measure the attribute variables. A design tells us what 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 wish to test discrimination in admissions. The idea for this example came from the unusual and ingenious experiment cited earlier: Walster, Cleary, and Clifford (1971). 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 of men and half women. Other things equal, we expect approximately equal numbers of acceptances and rejections.

	Treatments	
A_1 (Male)		A_2 (Female)
	Acceptance Scores	
MA_1		MA_2

Figure 7.1

Acceptance, then, is the dependent variable. It is measured on a three-point scale: full acceptance, qualified acceptance, and rejection. Call male A_1 and female A_2. The paradigm of the design is given in Figure 7.1.

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, MA_1 and MA_2 will be tested for statistical significance with a t or F test. The substantive hypothesis is: $MA_1 > MA_2$, or more males than females will be accepted for admission. If there is no discrimination in admissions, then MA_1 is statistically equal to MA_2. Suppose that an F test indicates that the means are not significantly different. Can we then be sure that there is no discrimination practiced (on the average)? While the design of Figure 7.1 is satisfactory as far as it goes, perhaps it does not go far enough.

A Stronger Design

Walster and her colleagues used two other independent variables, race and ability, in a factorial design. We drop race-it was not 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 7.1; use three levels. That is, in addition to the applications being designated male and female, they are also designated as high, medium, and low ability. For example, three of the applications 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 7.1. 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 7.2.

The design is a 2 x 3 factorial. One independent variable, A, is gender, the same as in Figure 7.1. 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 student records sent to the colleges were systematically adjusted to fit the six cells of Figure 7.2. A case in the A_1B_2 cell, for instance, would be the record of a male of medium ability. It is this record that the college judges for admission.

Let's assume that we believe discrimination against women takes a more subtle form than simply across-the board exclusion: that it is the women of lower ability who are discriminated against (compared to men). This is an interaction hypothesis. At any rate, we use this problem and the paradigm of Figure 7.2 as a basis for discussing some elements of research design.

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

The paradigm of Figure 7.2 suggests a number of things. First and most obvious, a fairly large number of subjects is needed.

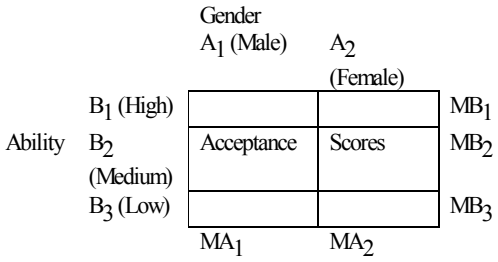


Figure 7.2

Specifically, 6n subjects are necessary (n = number of S's in each cell). If we decide that n should be 20, then we must have 120 S's for the experiment. Note the "wisdom" of the design here. If we were only testing the treatments and ignoring ability, only 2n S's would be needed. Please note that some such as Simon (1976, 1987), Simon & Roscoe (1984), and Daniel (1976) disagrees with this approach for all types of problems. They feel that many designs contains hidden replications and that one can do with a lot fewer subjects that 20 per cell. Such designs do require a lot more careful planning, but the researcher can come out with a lot more useful information.

There are ways to determine how many subjects are needed in a study. Such determination is part of the subject of "power," which refers to the ability of a test of statistical significance to detect differences in means (or other statistics) when such differences indeed exist. Chapter 8 discusses sample sizes and their relationship to research. Chapter 12, however, presents a method for estimating sample sizes to meet certain criteria. Power is a fractional value between 0 and 1.00 that is defined as $1 - \beta$, where β is the probability of committing a Type 2 error. The Type 2 error is failing to reject a false null hypothesis. If power is high (close to 1.00), this says that if the statistical tests was not significant, the research can conclude that the null hypothesis is true. Power also tells you how sensitive the statistical test is in picking up real differences. If the statistical test is not sensitive enough to detect a real difference, the test is said to have low power. A highly sensitive test that can pick up true differences is said to have high power. In Chapter 16, we discussed the difference between parametric and nonparametric statistical tests. Nonparametric tests are generally less sensitive than parametric tests. As a result nonparametric tests are said to have lower power than parametric tests. One of the most comprehensive books on the topic of power estimation is by Cohen (1988). Jaccard and Becker 1997) gives a easy to follow introduction to power analysis.

Second, the design indicates that the "subjects" (colleges, in this case) can be randomly assigned to both A and B because both are experimental variables. If ability were a nonexperimental attribute variable, however, then the subjects could be randomly assigned to A_1 and A_2, but not to B_1, B_2, and B_3. Third, according to the design the observations made on the "subjects" 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 of Figure 7.2 in effect prescribes the operations necessary for obtaining the measures that are appropriate for the statistical analysis. An F test depends on the assumption of the independence of the measures of the dependent variable. If ability here is an attribute variable and individuals are measured for intelligence, say, then the independence requirement is in greater jeopardy because of the possibility of one subject seeing another subject's paper, because teachers may unknowingly (or knowingly) "help" children with answers, and for other reasons. Researchers try to prevent such things-not on moral grounds but to satisfy the requirements of sound design and sound statistics.

A fourth point is quite obvious to us by now: Figure 7.2 suggests factorial analysis of variance, F tests, measures of association, and, perhaps, post hoc tests. If the research is well designed before the data are gathered-as it certainly was by Walster et al.-most statistical problems can be solved. In addition, certain troublesome problems can be avoided before they arise or can even be prevented from arising at all. With an inadequate design, however, problems of appropriate statistical tests may be very troublesome. One reason for the strong emphasis in this book on treating design and statistical problems concomitantly is to point out ways to avoid these problems. If design and statistical analysis are planned simultaneously, the analytical work is usually straightforward and uncluttered.

A highly useful dividend of design is this: A clear design, like that in Figure 7.2, suggests the statistical tests that can be made. A simple one-variable randomized design with two partitions, for example, two treatments, A1 and A_2, permits 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 7.2, however, three statistical tests are possible: (1) between A_1 and A_2; (2) among B_1, B_2, and B_3; and (3) the interaction of A and

B. In most investigations, all the statistical tests are not of equal importance. The important ones, naturally, are those directly related to the research problems and hypotheses.

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

It should be evident that research design is not static. A knowledge of design can help us to plan and do better research and can also suggest the testing of hypotheses. Probably more important, we may be led to realize that the design of a study is not in itself 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 of Figure 7.3. (To make matters simpler, we are only using two levels of B_1, B_2, and B_3, but changing B_3 to B_2. The design is thus reduced to 2 x 2.)

The important point to note is that *no adequate* test of the hypothesis is possible with this design. A, 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 $MA_1 > MA_2|B_2$ (MA_1 is greater than MA_2, under condition B_2), as hypothesized, the design cannot provide a clear possibility of confirming the hypothesized interaction since we cannot obtain information about the differences between A_1 and A_2 at the two levels of B, B_1 and B_2. Remember that an interaction hypothesis implies, in this case, that the difference between A, 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 connect. But presumptive evidence is not good enough, especially when we know that it is possible to obtain better evidence.

B_1, Condition		*B_2 Condition*	
Treatments		*Treatments*	
A_1	A_2	A_1	A_2
MA_1	MA_2	MA_1	MA_2

Figure 7.3

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

Figure 7.4

In Figure 7.3, suppose the means of the cells were, from left to right: 30, 30, 40, 30. This result would seem to support the interaction hypothesis, since there is a significant difference between A_1 and A_2 at level B_2, but not at level B_1. But we could not know this to be certainly so, even though the difference between A_1 and A_2 is statistically significant. Figure 7.4 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 vari*ance.

According to this principle, by constructing an efficient research design the investigator attempts (1) to

maximize the variance of the variable or variables of the substantive research hypothesis, (2) to control the variance of extraneous or "unwanted" variables that may have an effect on the experimental outcomes, and (3) to minimize the error or random variance, including so-called errors of measurement. Let us 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 1970's 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 and Cameron, 1996; Sharpley, 1988; McCullers, Fabes and Moran, 1987 and 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 and Falstrom, 1975; Ross and McBean, 1995; Senemoglu and Fogelman, 1995; Bergin, 1995). Although there appears to be some research supporting the efficacy of mastery learning. there is at least one study-and a fine study it is by Thompson (1980) whose results indicate that students taught through the mastery learning approach do no better than students taught with a conventional approach of lecture, discussion, and recitation. This is an exemplary study, done with careful controls over an extended time period. The example given below was inspired by the Thompson study. The design and controls in the example, however, are much simpler than Thompson's. Note, too, that Thompson had an enormous advantage: He did his experiment in a military establishment. This means, of course, that many control problems, usually recalcitrant in educational research, are easily solved.

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

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

Independent Variables

Methods	*Aptitude*
Mastery Learning, A_1	High Aptitude, B_1
Traditional, A_2	Low Aptitude, B_2

Dependent Variable

Science Achievement
Test scores in Science

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.

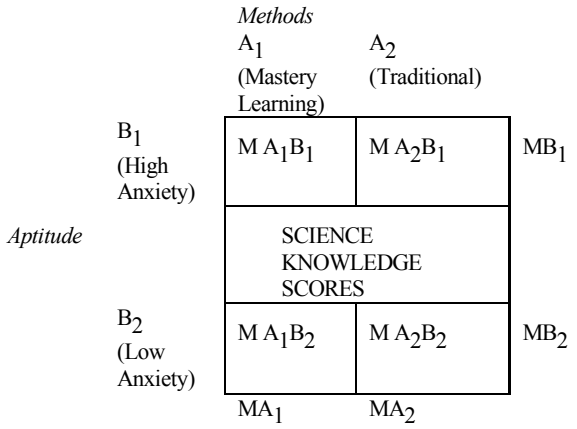


Figure 7.5

One, there are two independent variables. Two, 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 in Figure 7.5.

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. A*n attribute variable* is a measured variable or a variable that is a characteristic of people or groups, e.g., 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 subjects as high aptitude and low aptitude and assign them accordingly to B_1 and B_2. We can, however, randomly assign the students to A_1 and A_2, the methods groups. This is done in two stages: (1) the B_1 (high aptitude) students are randomly assigned to A_1 and A_2 and (2) the B_2 (low aptitude) students are randomly assigned to A_1 and A_2. By so randomizing the subjects we can assume that before the experiment begins, the students in A_1 are approximately equal to the students in A_2 in all possible characteristics.

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

Before tackling the application of the maxmincon principle in the present example, an important point should be discussed. Whenever we talk about variance, we must be sure to know what 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 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 influ*ence,* presumably, the variance of the dependent variable. Somewhat inaccurately put we "make" the measures of the dependent variable behave or vary as a presumed result of our manipulation and control of the independent variables. In an experiment, it is the dependent variable measures that are analyzed. Then, from the analysis we infer that the variances present in the total variance of the dependent variable measures are due to the manipulation and control of the independent variables and to error. Now, back to our principle.

MAXIMIZATION OF EXPERIMENTAL VARIANCE

The experimenter's most obvious, but not necessarily most important, concern is to maximize what we will call the *experimental variance.* This term is introduced to facilitate subsequent discussions and, in general, simply refers to the variance of the dependent variable influenced by the independent variable or variables of the substantive hypothesis. In this particular case, the experimental variance is the variance in the dependent

variable presumably due to methods, A_1 and A_2, and aptitude, B_1 and B_2. Although experimental variance can be taken to mean only the variance due to a manipulated *or active* variable, like methods, we shall also consider *attribute* variables, like intelligence, gender, and, in this case, aptitude, experimental variables. One of the main tasks of an experimenter is to maximize this variance. The methods must be "pulled" apart as much as possible to make A_1 and A_2 (and A_3, A_4, and so on, if they are in the design) as unlike as possible.

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

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

CONTROL OF EXTRANEOUS VARIABLES

The control of extraneous variables means that the influences of 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 subjects 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* subjects *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 of controlling all possible extraneous variables. Another way to phrase it is: if 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 many possibilities of inequality. If we match for intelligence, we may successfully achieve statistical equality in intelligence (at least in those aspects of intelligence measured), but we may suffer from inequality in other significantly influential independent variables like aptitude, motivation, and social class. A precept that springs from this equalizing power of randomization, then, is: *Whenever* it is *possible to do so, randomly assign subjects to experimental groups and conditions, and randomly assign conditions and other factors to experimental groups.*

The third means of controlling an extraneous variable is to build it right into the design as an independent variable. For example, assume that gender was to be controlled in the experiment discussed earlier and it was considered inexpedient or unwise to eliminate it. One could add a third independent variable, gender, to the design. Unless one were interested in the actual difference between the genders on the dependent variable or wanted to study the interaction between one or two of the other variables and gender, however, it is unlikely that this form of control would be used. One might want information of the kind just mentioned and also want to control gender, too. In such a case, adding it to the design as a variable might be desirable. The point is that building a variable into an experimental design "controls" the variable, since it then becomes possible to extract from the total variance of the dependent variable the variance due to the variable. (In the above case, this would be the "between-genders" variance.)

These considerations lead to another principle: An *extraneous variable can* be con*trolled 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* varia*bles.*

The fourth way to control extraneous variance is to match subjects. 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, say into high and low intelligence in a factorial design, and then randomize within each level as described above. Matching is a special case of this principle. Instead of splitting the subjects into two, three, or four parts, however, they are split into N/2 parts, N being the number of subjects 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 subjects 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 subjects. It is difficult to find matched subjects 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 the baby out with the bath, 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 subjects are used with different experimental treatments-called repeated measures or randomized blocks design-we have powerful control of variance. How match better on all possible variables than by matching a subject with oneself? Unfortunately, other negative considerations usually rule out this possibility. It should be forcefully emphasized that matching of any kind is no substitute for randomization. If subjects 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 subjects 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 later 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 subjects. 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 and lapses of memory, transient emotional states of subjects, 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 subjects 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, to this extent they are reliable. Imagine a completely unreliable measurement instrument, one that does not allow us to predict the future performance of individuals at all, one that gives one rank ordering of a sample of subjects at one time and a completely different rank ordering at another time. With such an instrument, it would not be possible to identify and extract systematic variances, since the scores yielded by the instrument would be like the numbers in a table of random numbers. This is the extreme case. Now imagine differing amounts of reliability and unreliability in the measures of the dependent variable. The more reliable the measures, the better we can identify and extract systematic variances and the smaller the error variance in relation to the total variance.

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

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

$$V_t = V_b + V_e$$

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

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

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

will inflate the within-groups or residual variance, the denominator of the above equation.

Study Suggestions

l. We have noted that research design has the purposes 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 subjects or repeated measures of the same subjects control variance? What is the relation between the research questions and hypotheses and a research design? In answering these questions, make up a research problem to illustrate what you mean (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, think of the maxmincon principle and F tests and t tests.

Chapter Summary

1. Research designs are plans and structures used to answer research questions.
2. Research designs have two basic purposes: (a) provide answers to research questions and (b) control variance.
3. Research designs work in conjunction with research hypotheses to yield a dependable and valid answer.
4. Research designs can also tell us what statistical test to use to analyze the data collected from that design.
5. When speaking of controlling variance, we can mean one or more of three things
   - maximize systematic variance
   - control extraneous variance
   - minimize error variance
6. To maximize systematic variance, one should have an independent variable where the levels are very distinct from one another.
7. To control extraneous variance the researcher need to eliminate the effects of a potential independent variable on the dependent variable. This can be done by
   - holding the independent variable constant. If one knows gender has a possible effect, gender can be held constant by doing the study with one type of gender, i.e. females.
   - randomization. This would mean randomly choosing subjects and then randomly assigning each group of subjects to treatment conditions (levels of the independent variable)
   - build the extraneous variable into the design by making it an independent variable.
   - matching subjects. This method of control might be difficult in certain situations. The 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.

Chapter 8

Inadequate Designs and Design Criteria

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

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

The principal focus of this chapter is on inadequate research designs. Such designs have been so common that they must be discussed. More important, the student should be able to recognize them and understand why they are inadequate. This negative approach has a virtue: the study of deficiencies forces one to ask why something is deficient, which in turn centers attention on the criteria used to judge both adequacies and inadequacies. So the study of inadequate designs leads us to the study of the criteria of research design. We take the opportunity, too, to describe the symbolic system to be used and to identify an important distinction between experimental and nonexperimental research.

EXPERIMENTAL AND NONEXPERIMENTAL APPROACHES

Discussion of design must be prefaced by an important distinction: that between experimental and nonexperimental approaches to research. Indeed, this distinction is so important that a separate chapter 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 which we discussed in an earlier chapter, they manipulated sex, race, and ability levels 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 subjects or treatments at random because the nature of the variables is such as to preclude manipulation. Subjects come to us with their differing characteristics intact, so to speak. They come to us with their sex, intelligence, occupational status, creativity, or aptitude "already there." Wilson (1996) used a nonexperimental design to study the readability, ethnic content and cultural sensitivity of patient education material used by nurses at local health department and community health centers. Here, the material pre-existed. 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 8 hospital units at two urban teaching hospitals. Edmondson did not choose these units or hospitals at random nor 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 & 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 and 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 means *an experimentally manipulated* independent variable (or variables). X_1 X_2, X_3, etc. mean independent variables 1, 2, 3, and so on, though we usually use X alone, even when it can mean more than one independent variable. (We also use X_1 X_2, etc. to mean partitions of an independent variable, but the difference will always be clear.) The symbol (X) indicates that the independent variable *is not manipu*lated-is not under the direct control of the investigator, but *is measured or imagined.* The dependent variable is Y: Y_b is the dependent variable *before* the manipulation of X, and Y_a the dependent variable *after* the manipulation of X. With ~X, we borrow the negation sign of set theory; ~X ("not-X") means that the experimental variable, the independent variable X, is not manipulated. (Note: (X) is a nonmanipulable variable and ~X is a manipulable variable that is not manipulated.) The symbol (R) will be used for the random assignment of subjects to experimental groups and the random assignment of experimental treatments to experimental groups.

The explanation of ~X, just given, is not quite accurate, because in some cases ~X can mean a different aspect of the treatment X rather than merely the absence of X. 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, ~X. For our purposes, however, ~X will do well enough, especially if we understand the generalized meaning of "control" discussed below. *An experimental group,* then, is a group of subjects receiving some aspect or treatment of X. In testing the frustration-aggression hypothesis, the experimental group is the group whose subjects are systematically frustrated. In contrast, the control group is one that is given "no" treatment.

In modern multivariate research, it is necessary to expand these notions. They are not changed basically; they are only expanded. It is quite possible to have more than one experimental group, as we have seen. Different degrees of manipulation of the independent variable are not only possible; they are often also desirable or even imperative. Furthermore, 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?-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 X_1 X_2, X_3, . . ., X_k, different forms or degrees of treatment.

If X is circled, (X), this means that the investigator “imagines" the manipulation of X, or assumes that X occurred and that it is the X of the hypothesis. It may also mean that X is measured and not manipulated.

Actually, we are saying the same thing here in different ways. The context of the discussion should make the distinction clear. Suppose a sociologist is studying delinquency and the frustration-aggression hypothesis. The sociologist observes delinquency, Y, and imagines that the delinquent subjects were frustrated in their earlier years, or (X). All nonexperimental designs will have (X). Generally, then, (X) means an independent variable not under the experimental control of the investigator.

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

FAULTY DESIGNS

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

Design 8.1: One-Group

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

Design 8.1 (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 8.1 (b) is the nonexperimental form of the one-group design. Y, the outcome, is studied, and X is assumed or imagined. An example would be to study delinquency by searching the past of a group of juvenile delinquents for factors that may have led to 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 8.1 is worthless. There is virtually no control of other possible influences on outcome. As Campbell (1957) long ago pointed out, the minimum of useful scientific information requires at least one formal comparison. The curriculum example requires, *at the least,* comparison of the group that experienced the new curriculum with a group that did not experience it. The presumed effect of the new curriculum, say such and-such achievement, might well have been about the same under any kind of curriculum. The point is not that the new curriculum did or did not have an effect. It was that without any formal, controlled comparison of the performance of the members of the "experimental" group with the performance of the members of some other group not experiencing the new curriculum, little can be said about its effect.

An important distinction should be made. It is not that the method is entirely worthless, but that it is *scientifically* worthless. In everyday life, of course, we depend on such scientifically questionable evidence; we have to. We act, we say, on the basis of our experience. We hope that we use our experience rationally. The everyday-thinking paradigm implied by Design 8.1 is not being criticized. Only when such a paradigm is used and said or believed to be scientific do difficulties arise. Even in high intellectual pursuits, the thinking implied by this design is used. Freud's careful observations and brilliant and creative analysis of neurotic behavior seem to fall into this category. The quarrel is not with Freud, then, but rather with assertions that his conclusions are "scientifically established. "

Design 8.2: One-Group, before-after (Pretest Posttest)

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

Design 8.2 is only a small improvement on Design 8.1. The essential characteristic of this mode of research is that a group is compared with itself. Theoretically, there is no better choice since all possible independent variables associated with the subjects' characteristics are controlled. The procedure dictated by such a design is as follows. A group is measured on the dependent variable, Y, before experimental manipulation. This is usually

called *a pretest.* Assume that the attitudes toward women of a group of subjects are measured. An experimental manipulation designed to change these attitudes is used. An experimenter might expose the group to expert opinion on women's rights, for example. After the interposition of this X, the attitudes of the subjects are again measured. The difference scores, or $Y_a - Y_b$, are examined for change in attitudes.

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

Measurement, History, Maturation

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

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

The Regression Effect

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

Research designs have to be constructed with the regression effect in mind. There is no way in Design 8.2 to control it. If there were 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 equal, are affected the same by regression and other influences. So if the groups differ in the posttest, it should be due to the experimental manipulation.

Design 8.2 is inadequate not so much because extraneous variables and the regression effect can operate (the extraneous variables operate whenever there is a time interval between pretest and posttest), but *because we do not know whether they have operated, whether they have affected the dependent-variable measures*. The design affords no opportunity to control or to test such possible influences.

Design 8.3: Simulated before-after

$$\frac{X \qquad Y_a}{Y_b}$$

The peculiar title of this design stems in part from its very nature. Like Design 8.2 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 8.4: Two Groups, No Control

(a) $\frac{X \qquad Y}{\sim X \qquad Y}$ (Experimental)

(b) $\frac{(X) \qquad Y}{(\sim X) \qquad \sim Y}$ (Nonexperimental

Design 8.4 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 subjects 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 ~Y in the bottom line). The first group is found to have experienced X, the second group not to have experienced X.

This design has a basic weakness. The two groups are *assumed to* be equal in independent variables other than X. It is sometimes possible to check the equality of the groups roughly by comparing them on different pertinent variables, for example, age, sex, income, intelligence, ability, and so on. This should be done if it is at all possible, but, as Stouffer (1950) says, "there is all too often a wide-open gate through which other uncontrolled variables can march," p. 522. Because randomization is not used-that is, the subjects 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 adequately answering the research questions. 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 subjects for reasons irrelevant to the research and then using an experimental group-control group type of design. Students often assume, because they match pupils on intelligence and sex, for instance, that their experimental groups are equal. They have heard that one should match subjects 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 subjects 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 means control of independent variables: the independent variables of the research study and extraneous independent variables. Extraneous independent variables, of course, are 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 subjects at random; assign subjects *to groups at random;* assign *experimental treatments to groups at random.*

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

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

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

As we have seen earlier, randomization theoretically satisfies this principle (see Chapter 8). When we test the empirical validity of an If p, then q proposition, we manipulate p and observe that q covaries with the

manipulation of p. But how confident can we be that our If p, then q statement is really "true"? Our confidence is directly related to the completeness and adequacy of the controls. If we use a design similar to Designs 8.1 through 8.4, we cannot have too much confidence in the empirical validity of the If p, then q statement, since our control of extraneous independent variables is weak or nonexistent. Because such control is not always possible in much psychological, sociological, and educational research, should we then give up research entirely? By no means. Nevertheless, we must be aware of the weaknesses of intrinsically poor design.

Generalizability

The third criterion, *generalizability,* is independent of other criteria because it is different in kind. This is an important point that will shortly become clear. It means simply: Can we *generalize the results of a study to other subjects, other groups, and other conditions?* Perhaps the question is better put: *How much* can we generalize the results of the study? This is probably the most complex and difficult question that can be asked of research data because it touches not only on technical matters like sampling and research design, but also on larger problems of basic and applied research. In basic research, for example, generalizability is not the first consideration, because the central interest is the relations among variables and why the variables are related as they are. This emphasizes the internal rather than the external aspects of the study. These studies are often designed to examine theoretical issues such as motivation or learning. The goal of basic research is to add information and knowledge to a field of study but usually without a specific practical purpose. Its results are generalizable, but not in the same realm as results found in applied research studies. In applied research, on the other hand, the central interest forces more concern for generalizability because one certainly wishes to apply the results to the other persons and to other situations. Applied research studies usually have their foundations in basic research studies. Using information found in a basic research study, applied research studies apply those findings to determine if it can solve a practical problem. Take the work of B.F. Skinner for example. His early research is generally considered as basic research. It was from his research that schedules of reinforcement were established. However, later, Skinner and others (Skinner, 1968; Garfinkle, Kline & Stancer, 1973) applied the schedules of reinforcement to military problems, educational problems and behavioral problems. Those who do research on the modification of behavior are applying many of the theories and ideas tested and established by B. F. Skinner. If the reader will ponder the following two examples of basic and applied research, one can get closer to this distinction.

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

Internal and External Validity

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

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

Earlier in this chapter we presented four possible threats to internal validity. Some textbook writers 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 & Stanley (1963) also list four other threats. They are instrumentation, selection, attrition and the interaction between one of more of these previous seven.

Instrumentation is a problem if the device used to measure the dependent variable changes over time. This is particularly true in studies using a human observer. Human observers or judges can be affected by previous events or fatigue. Observers may become more efficient over time and thus the later measurements are more accurate than earlier ones. On the other hand, with fatigue, the human observer would become less accurate in the later trials than the earlier ones. When this happens, the values of the dependent variable will change and that change was not 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 does 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 drop out of the study, the unbalance is a possible reason for the change in the dependent variable. Attrition also includes the drop out of subjects 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 of which one are volunteers (self-selected) and the other group are 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 subjects. 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* means *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 school children? All eighth-grade children? All eighth-grade children in this school system or the eighth-grade children of this school only? Or must the findings be limited to the eighth-grade children with whom we worked? This is a very important scientific question that should always be *asked-and-answered.*

Not only must sample generalizability be questioned. It is necessary to ask questions about the ecological and variable representativeness of studies. If the social setting in which the experiment was conducted is changed, will the relation of A and B still hold? Will A be related to B if the study is replicated in a lower-class school? In a western school? In a southern school? These are questions *of ecological representativeness.*

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

In the case of variables like achievement, aggression, aptitude, and anxiety, can the investigator assume that the "aggression" of the suburban subjects 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 will be 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 effect of testing, the interaction effects of selection biases and the independent variable, reactive effects of experimental arrangements and multiple-treatment interference.

In the reactive or interaction effect of testing, the reference is to the use of a pretest prior to administering treatment. Pre-testing 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 non-pretested 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 a person is participating in an experimental study may alter their 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 are 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 design-structural principles alone. This point should be emphasized because in the next chapter we will find that a perfectly good design structure can be poorly used. Thus it is necessary to learn and understand the two sources of research weakness: intrinsically poor designs and intrinsically good designs poorly used.

Study Suggestions

1. The faculty of a liberal arts college has decided to begin a new curriculum for all undergraduates. It asks a faculty 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 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 will not study any other school situation during the period so as not to bias the results, she says.

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? If so, why? If not why not? Which field is likely to have more basic

research: psychology or education'? Why? What implications does your conclusion have for generalizability?

4. What does replication of research have to do with generalizability? Explain. If it were possible, should all research be replicated? If so, why? What does replication have to, do with external and internal validity?

Chapter Summary

1. A study of designs that are faulty 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. Faculty 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. Faculty designs are discussed in terms of internal validity.
5. Internal validity is concerned with how strong of a statement the experimenter can state about the effect of the independent variable on the dependent variable. The more confident 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 8 basic classes of extraneous variables, which if not controlled may be confounded with the independent variable. These 8 basic classes are called threats to internal validity
8. Campbell’s threats to internal validity can be outlined as follows:
   a) History
   b) Maturation
   c) Testing or measurement
   d) Instrumentation
   e) Statistical Regression
   f) Selection
   g) Experimental Mortality or attrition
   h) Selection-Maturation interaction
9. External validity is concerned with how strong of 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
    a)Reactive or interaction effect of testing
    b) Interaction effects of selection biases and the independent variable
    c) Reactive effects of experimental arrangements
    d) Multiple-treatment interference

Chapter 9

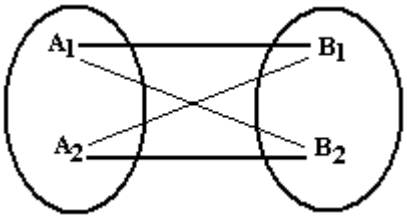
General Designs of Research

DESIGN is data discipline. The implicit purpose of all research design is to impose controlled restrictions on observations of natural phenomena. The research design tells the investigator, in effect: Do this and this; don't do that or that; be careful with this; ignore that; and so on. It is the blueprint of the research architect and engineer. If the design is poorly conceived structurally, the final product will be faulty. If it is at least well conceived structurally, the final product has a greater chance of being worthy of serious scientific attention. In this chapter, our main preoccupation is seven or eight "good" basic designs of research. In addition, however, we take up certain conceptual foundations of research and two or three problems related to design-for instance, the rationale of control groups and the pros and cons of matching.

CONCEPTUAL FOUNDATIONS OF RESEARCH DESIGN

The conceptual foundation for understanding research design was laid in Chapters 4 and 5, where sets and relations were defined and discussed. Recall that *a relation is* a set of ordered pairs. Recall, too, that *a Cartesian product is* all the possible ordered pairs of two sets. *A partition* breaks down a universal set U into subsets that are *disjoint* and *exhaustive. A cross partition* is a new partitioning that arises from successively partitioning *U* by forming all subsets of the form $A \cap B$. These definitions were elaborated in Chapters 5 and 6. We now apply them to design and analysis ideas.

Take two sets, A and B, partitioned into A_1 and A_2, B_1 and B_2. The Cartesian product of the two sets is:



The ordered pairs, then, are: A_1B_1, A_1B_2, A_2B_1, A_2B_2. Since we have a set of ordered pairs, this is a relation. It is also a cross partition. The reader should look back at Figures 4.7 and 4.8 of Chapter 4 to help clarify these ideas, and to see the application of the Cartesian product and relation ideas to research design. For instance, A_1 and A_2 can be two aspects of any independent variable: experimental-control, two methods, male and female, and so on.

A design is some subset of the Cartesian product of the independent variables and the dependent variable. It is possible to pair each dependent variable measure, which we call Y in this discussion, with some aspect or partition of an independent variable. The simplest possible cases occur with one independent variable and one dependent variable. In Chapter 10, an independent variable, A, and a dependent variable, B, were partitioned into $[A_1, A_2]$ and $[B_1, B_2]$ and then cross-partitioned to form the by-now familiar 2×2 crossbreak, with frequencies or percentages in the cells. We concentrate, however, on similar cross partitions of A and B, but with continuous measures in the cells.

Take A alone, using a one-way analysis of variance design. Suppose we have three experimental treatments, A_1, A_2 and A_3, and, for simplicity, two Y scores in each cell. This is shown on the left of Figure 9.1, labeled (a). Say that six participants have been assigned at random to the three treatments, and that the scores of the six individuals after the experimental treatments are those given in the figure.

The right side of Figure 9.1, labeled (b), shows the same idea in ordered-pair or relation form. The ordered pairs are A_1Y_1, A_1Y_2, A_2Y_3, . . . , A_3Y_6. This is, of course, not a Cartesian product, which would pair A_1 with all the Y's, A_2 with all the Y's, and A_3 with all the Y's, a total of $3 \times 6 = 18$ pairs. Rather, Figure 9.1(b) is a subset of the Cartesian product, $A \times B$. Research designs are subsets of $A \times B$, and the design and the research problem define or specify how the subsets are set up. The subsets of the design of Figure 9.1 are presumably dictated by the research problem.

When there is more than one independent variable, the situation is more complex. Take two independent variables, A and B, partitioned into $[A_1, A_2]$ and $[B_1, B_2]$. The reader should not confuse this with the earlier AB frequency paradigm, in which A was the

independent variable and B the dependent variable.

(a)

A_1	A_2	A_3
7	7	3
9	5	3

(b)

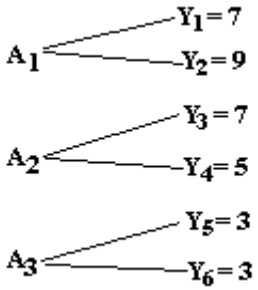


Figure 9.1

(a)

	B_1	B_2
A_1	$Y_1 = 8$	$Y_3 = 4$
	$Y_2 = 6$	$Y_4 = 2$
A_2	$Y_5 = 4$	$Y_7 = 8$
	$Y_6 = 2$	$Y_8 = 6$

(b)

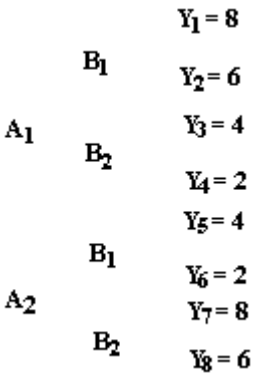


Figure 9.2

We must now have ordered triples (or two sets of ordered pairs): ABY. Study Figure 9.2. On the left side of the figure, labeled (a), the 2 × 2 factorial analysis of variance design and example used in Chapter 14 (see Figure 14.2 and Tables 14.3 and 14.4) is given, with the measures of the dependent variable, Y, inserted in the cells. That is, eight participants were assigned at random to the four cells. Their scores, after the experiment, are Y_1 Y_2, . . . Y_8. The right side of the figure, labeled (b), shows the ordered triples, ABY, as a tree. Obviously these are subsets of A × B × Y and are relations. The same reasoning can be extended to larger and more complex designs, like a 2 × 2 × 3 factorial (ABCY) or a 4 × 3 × 2 × 2 (ABCDY). (In these designations, Y is usually omitted because it is implied.) Other kinds of designs can be similarly conceptualized, though their depiction in trees can be laborious.

In sum, a research design is some subset of the Cartesian product of the independent and the dependent variables. With only one independent variable, the single variable is partitioned; with more than one independent variable, the independent variables are crosspartitioned. With three or more independent variables, the conceptualization is the same; only the dimensions differ, for example, A ¥ B ¥ C and A ¥ B x C ¥ D and the cross partitions thereof. Whenever possible, it is desirable to have "complete" designs-a complete design is a cross partition of the independent variables-and to observe the two basic conditions of disjointness and exhaustiveness. That is, the design must not have a case (a participant's score) in more than one cell of a partition or cross partition, and all the cases must be used up. Moreover, the basic minimum of any design is at least a partition of the independent variable into two subsets, for example, A into A_1 and A_2. There are also "incomplete" designs, but "complete" designs are emphasized more in this book. See Kirk (1995) for a more complete treatment of incomplete designs.

The term "general designs" means that the designs given in the chapter are symbolized or expressed in their most general and abstract form. Where a simple X, meaning independent variable, is given, it must be taken to mean more than one X-that is, X is partitioned into two or more experimental groups. For instance, Design 9.1, to be studied shortly, has X and ~X, meaning experimental and control groups, and thus is a partition of X. But X can be partitioned into a number of X's, perhaps changing the design from a simple one-variable design to, say, a factorial design. The basic symbolism

associated with Design 9.1, however, remains the same. These complexities will, we hope, be clarified in this and succeeding chapters.

A PRELIMINARY NOTE: EXPERIMENTAL DESIGNS AND ANALYSIS OF VARIANCE

Before taking up the designs of this chapter, we need to clarify one or two confusing and potentially controversial points not usually considered in the literature. Most of the designs we consider are experimental. As usually conceived, the rationale of research design is based on experimental ideas and conditions. They are also intimately linked to analysis of variance paradigms. This is of course no accident. Modern conceptions of design, especially factorial designs, were born when analysis of variance was invented. Although there is no hard law that says that analysis of variance is applicable only in experimental situations-indeed, it has been used many times in nonexperimental research-it is in general true that it is most appropriate for the data of experiments. This is especially so for factorial designs where there are equal numbers of cases in the design paradigm cells, and where the participants are assigned to the experimental conditions (or cells) at random.

When it is not possible to assign participants at random, and when, for one reason or another, there are unequal numbers of cases in the cells of a factorial design, the use of analysis of variance is questionable, even inappropriate. It can also be clumsy and inelegant. This is because the use of analysis of variance assumes that the correlations between or among the independent variables of a factorial design are zero. Random assignment makes this assumption tenable since such assignment presumably apportions sources of variance equally among the cells. But random assignment can only be accomplished in experiments. In nonexperimental research, the independent variables are more or less fixed characteristics of the participants, e.g., intelligence, sex, social class, and the like. They are usually systematically correlated. Take two independent manipulated variables, say reinforcement and anxiety. Because participants with varying amounts of characteristics correlated with these variables are randomly distributed in the cells, the correlations between aspects of reinforcement and anxiety are assumed to be zero. If, on the other hand, the two independent variables are intelligence and social class, both ordinarily nonmanipulable and correlated, the assumption of zero correlation between them necessary for analysis of variance cannot be made. Some method of analysis that takes account of the correlation between them should be used. We will see later in the book that such a method is readily available: multiple regression.

We have not yet reached a state of research maturity to appreciate the profound difference between the two situations. For now, however, let us accept the difference and the statement that analysis of variance is basically an experimental conception and form of analysis. Strictly speaking, if our independent variables are nonexperimental, then analysis of variance is not the appropriate mode of analysis. There are exceptions to this statement. For instance, if one independent variable is experimental and one nonexperimental, analysis of variance is appropriate. In one-way analysis of variance, moreover, since there is only one independent variable, analysis of variance can be used with a nonexperimental independent variable, though regression analysis would probably be more appropriate. In Study Suggestion 5 at the end of the chapter, an interesting use of analysis of variance with nonexperimental data is cited.

Similarly, if for some reason the numbers of cases in the cells are unequal (and disproportionate), then there will be correlation between the independent variables, and the assumption of zero correlation is not tenable. This rather abstract and abstruse digression from our main design theme may seem a bit confusing at this stage of our study. The problems involved should become clear after we have studied experimental and nonexperimental research and, later in the book, that fascinating and powerful approach known as multiple regression.

THE DESIGNS

In the remainder of this chapter we discuss four or five basic designs of research. Remember that a design is a plan, an outline for conceptualizing the structure of the relations among the variables of a research study. A design not only lays out the relations of the study; it also implies how the research situation is controlled and how the data are to be analyzed. A design, in the sense of this chapter, is the skeleton on which we put the variable-and-relation flesh of our research. The sketches given in

Designs 9.1 through 9.8, following, are designs, the bare and abstract structure of the research. Sometimes analytic tables, such as Figure 9.2 (on the left) and the figures of Chapter 8 (e.g., Figures 8.2, 8.3, and 8.5) and elsewhere are called designs. While calling them designs does no great harm, they are, strictly speaking, analytic paradigms. We will not be fussy, however. We will call both kinds of representations "designs."

Design 9.1: Experimental Group-Control Group: Randomized Participants

[R]	X	Y	(Experimental)
	~X	Y	(Control)

Design 9.1, with two groups as above, and its variants with more than two groups, are probably the "best" designs for many experimental purposes in behavioral research. Campbell and Stanley (1963) call this design the *posttest only control group design* while Isaac and Michael (1987) refers to it as the *randomized control group posttest only design*. The (R) before the paradigm indicates that participants are randomly assigned to the experimental group (top line) and the control group (bottom line). This randomization removes the objections to Design 8.4 mentioned in Chapter 8. Theoretically, all possible independent variables are controlled. Practically, of course, this may not be so. If enough participants are included in the experiment to give the randomization a chance to "operate," then we have strong control, and the claims of internal validity are rather well satisfied. This design controls for the effects of history, maturation and pre-testing but does not measure these effects.

If extended to more than two groups and if it is capable of answering the research questions asked, Design 9.1 has a number of advantages. The advantages are (1) it has the best built-in theoretical control system of any design, with one or two possible exceptions in special cases; (2) it is flexible, being theoretically capable of extension to any number of groups with any number of variables; (3) if extended to more than one variable, it can test several hypotheses at one time; and (4) it is statistically and structurally elegant.

Before taking up other designs, we need to examine the notion of the control group, one of the creative inventions of the last hundred years, and certain extensions of Design 9.1. The two topics go nicely together.

The Notion of the Control Group and Extensions of Design 9.1

Evidently the word "control" and the expression "control group" did not appear in the scientific literature before the late nineteenth century. This is documented by Boring (1954). The notion of controlled experimentation, however, is much older. Boring says that Pascal used it as early as 1648. Solomon (1949) searched the psychological literature and could not find a single case of the use of a control group before 1901. Perhaps the notion of the control group was used in other fields, though it is doubtful that the idea was well developed. Solomon (p.175) also says that the Peterson and Thurstone study of attitudes in 1933 was the first serious attempt to use control groups in the evaluation of the effects of educational procedures. One cannot find the expression "control group" in the famous eleventh edition (1911) of the Encyclopedia Britannica, even though experimental method is discussed. Solomon also says that control-group design apparently had to await statistical developments and the development of statistical sophistication among psychologists.

Perhaps the first use of control groups in psychology and education occurred in 1901 with the publication of Thorndike and Woodworth (1901). One of the two men who did this research, E. L. Thorndike, extended the basic and revolutionary ideas of this first research series to education (Thorndike, 1924). Thorndike's controls, in this gigantic study of 8564 pupils in many schools in a number of cities, were independent educational groups. Among other comparisons, he contrasted the gains in intelligence test scores presumably engendered by the study of English, history, geometry, and Latin with the gains presumably engendered by the study of English, history, geometry, and shopwork. He tried, in effect, to compare the influence of Latin and shopwork. He also made other comparisons of a similar nature. Despite the weaknesses of design and control, Thorndike's experiments and those he stimulated others to perform were remarkable for their insight. Thorndike even berated colleagues for not admitting students of stenography and typing who had not studied Latin, because he claimed to have shown that the influence of various participants on

intelligence was similar. It is interesting that he thought huge numbers of participants were necessary-he called for 18,000 more cases. He was also quite aware, in 1924, of the need for random samples.

The notion of the control group needs generalization. Assume that in an educational experiment we have four experimental groups as follows. A, is reinforcement of every response, As reinforcement at regular time intervals, A_3 reinforcement at random intervals, and A_4 no reinforcement. Technically, there are three experimental groups and one control group, in the traditional sense of the control group. However, A_4 might be another "experimental treatment"; it might be some kind of minimal reinforcement. Then, in the traditional sense, there would be no control group. The traditional sense of the term "control group" lacks generality. If the notion of control is generalized, the difficulty disappears. Whenever there is more than one experimental group and any two groups are given different treatments, control is present in the sense of comparison previously mentioned. As long as there is an attempt to make two groups systematically different on a dependent variable, a comparison is possible. Thus the traditional notion that an experimental group should receive the treatment not given to a control group is a special case of the more general rule that comparison groups are necessary for the internal validity of scientific research.

If this reasoning is correct, we can set up designs such as the following:

[R]	X_1	Y
	X_2	Y
	X_3	Y

or

[R]	X_{1a}	Y
	X_{1b}	Y
	X_{2a}	Y
	X_{2b}	Y

These designs will be more easily recognizable if they are set up in the manner of analysis of variance, as in Figure 9.3. The design on the left is a simple one-way analysis of variance design and the one on the right a 2 × 2 factorial design. In the right-hand design, X_{1a} might be experimental and X_{1b} control, with X_{2a} and X_{2b} either a manipulated variable or a dichotomous attribute variable. It is, of course, the same design as that shown in Figure 9.2(a).

X_1	X_2	X_3
Y Measures		

	X_1	
	X_{1a}	X_{1b}
X_2	Y	Measures
X_{2a}		
X_{2b}		

Figure 9.3

The structure of Design 9.2 is the same as that of Design 9.1. The only difference is that participants are matched on one or more attributes. For the design to take its place as an "adequate" design, however, randomization must enter the picture, as noted by the small r attached to the M (for "matched"). It is not enough that matched participants are used. The members of each pair must be assigned to the two groups at random. Ideally, too, whether a group is to be an experimental or a control group is also decided at random. In either case, each decision can be made by flipping a coin or by using a table of random numbers, letting odd numbers mean one group and even numbers the other group. If there are more than two groups, naturally, a random number system must be used.

Design 9.2 :Experimental Group-Control Group: Matched Participants

$[M_r]$	X	Y	(Experimental)
	~X	Y	(Control)

As in Design 9.1, it is possible, though often not easy, to use more than two groups. (The difficulty of matching more than two groups was discussed earlier.)

There are times, however, when a matching design is an inherent element of the research situation. When the same participants are used for two or more experimental treatments, or when participants are given more than one trial, matching is inherent in the situation. In educational research, when schools or classes are in effect variables-when, say, two or more schools or classes are used and the experimental treatments are administered in each school or class then Design 9.2 is the basis of the design logic. Study the paradigm of a schools design in Figure 9.4. It is seen that variance due to the differences between schools, and such variance can be substantial, can be readily estimated.

	X_{e1}	X_{e2}	X_c
Schools	Experimental 1	Experimental 2	Control
1			
2			
3		Y Measures	
4			
5			

Figure 9.4

Matching versus Randomization

Although randomization, which includes random selection and random assignment, is the preferred method for controlling extraneous variance, there is merit to the use of matching. In a number of situations outside academic circles, the behavioral scientist will not be able to use randomization in achieving constancy between groups prior to the administration of treatment. Usually in a university a participant pool is available to draw from. Researchers in this situation can afford to use randomization procedures. In business-marketing research, however, this may not be the case. Popular among market researchers is the controlled-store test. The controlled store test is an experiment that is done in the field. The second author has conducted such studies for a number of market research firms and a grocery chain in southern California. One of the goals of the controlled store test is to be very discreet. If a manufacturer of soap products wants to determine the effects of a cents-off coupon on consumer purchasing behavior, that manufacturer does not want the competing manufacturer of the same product to know about it. Why? Well if a competitor knew that a research study was going on in a store, they could go in and buy up their own products and hence contaminate the study.

To return to our discussion of randomization versus matching often times a grocery chain or a chain of department stores has a finite number of stores to use in a study. Location and clientele have a lot of influence on sales. Sales are usually the dependent variable in such studies. With a limited number of stores to choose from in order to perform the research, random assignment often does not work in equating groups of stores. One store in the chain might do 3 to 4 times the volume of business as another. If it is chosen at random, the group in falls into will create a great deal of unbalance especially if the other group does not have one to balance it. In short, the groups will no longer be equal. Hence, the solution here is to match the stores on an individual basis. One half of the matched pair is randomly assigned to one experimental condition and the other half gets the other condition. With more than 2 conditions, more stores would have to be matched and then assigned to treatment conditions.

In some human factors engineering studies using simulators, the use of randomization is sometimes not economically or practically feasible. Consider the testing of two test configurations for a simulator. Researcher may want to know which one leads to fewer perceptual errors. Processes of randomization would say that the experimenter should randomly assigned participants to conditions as they enter the study. However, when it requires 3 to 6 months to change the configuration of the simulator from one to the other, it is no longer feasible to do it the “usual” way.

However, an important point to remember is that randomization when it can be done correctly and appropriately is better than matching. It is perhaps the only method for controlling unknown sources of variances. One of the major shortcomings of matching is that one can never be sure that an exact match has been made. Without that exactness, the inexactness can be an alternative explanation on why the dependent variable is different between the treatment conditions after treatment.

Some ways of matching groups.

Matching by Equating Participants

The most common method of matching is to equate participants on one or more variables to be controlled. Christensen (1996) refers to this method as the *precision control method* and Matheson, Bruce and Beauchamp (1978) calls it the *matched-by-correlated criterion design*. To control for the influence of intelligence on the dependent variable, for example, the researcher must make sure that the participants in each of the treatment groups are of the same intelligence level. The goal here is to create equivalent groups of participants. Using our example of intelligence, if we had only two treatment conditions, we would select pairs of participants with identical or near identical intelligence test scores. Half of each pair would be assigned at random to one treatment condition and the other half is assigned to the other treatment condition. In a controlled stored test, where location is an important variable to be controlled for, we would find two stores of similar locale and call them a match. After we have built up, lets say 10 of such pairs, we can then take one half of each pair and assign them to one test environment and the other half to the other. If we had required matching for three conditions, we would then have to find three people with the same intelligence score or three stores in the same locale.

The major advantage in using this method is that it is able to detect small differences (increase in sensitivity) by ensuring that the participants in the various groups are equal on at least the paired variables. However an important requirement is that the variables on which participants are matched must be correlated significantly with the dependent variable. We had shown in an earlier chapter that matching is most useful when the variables on which participants are matched correlate greater than 0.5 or 0.6 with the dependent variable.

This method of matching has two major flaws or disadvantages. First, it is difficult to know which are the most important variables to match. In most instances, there are many potentially relevant variables. In a study the researcher might match on age, sex, race, marital status and intelligence. However, the researcher could have selected many other variables. The researcher should select those variables that show the lowest correlation with each other but the highest correlation with the dependent variable.

A second problem is the decrease in finding eligible matched participants as the number of variables used for matching increases. To choose 3 or 4 variables to match on and then find enough participants that meets the matching criteria requires a lot of participants to choose from. The researcher would need a large pool of available participants in order to obtain a just a few that are matched on all of the relevant variables. Matching affects the generalizability of the study. The researcher can only generalize the results to other individuals having the same characteristics.

The Frequency Distribution Matching Method

The individual-by-individual technique of matching presented above is a very good for developing equal groups, but many participants must be eliminated because they cannot be matched. Frequency distribution method attempts to overcome this disadvantage while retaining some of the advantages of matching. This technique, as the name implies, matches groups of participants in terms of overall distribution of the selected variable or variables rather than on an individual-by-individual basis. Let us say that we want to have two or more groups matched on intelligence. Let's further say we want to use the frequency distribution method of matching. First we would need an intelligence test score on each child. We then need to create the two or more groups in such a way that the groups of participants would have to have the same average intelligence test score, as well as the same standard deviation and skewness of the scores. Here, each group would be statistically equal. The mean, standard deviation and skewness between each group would be statistically equivalent. A statistical test of hypotheses could be utilized, but the researcher needs to be aware that both types of errors should be considered. If more than one variable was considered to be relevant on which to match participants, each group of participants would be required to have the same statistical measures on all of these variables. The number of participants lost using this technique would not be as great as the number lost using the individual-by-individual method, because each additional participant would merely have to contribute to producing the appropriate statistical measures rather than be identical to another participant on the relevant variables. Hence, this technique is more flexible in terms of being able to use a particular participant.

The major disadvantage of matching by the frequency distribution method occurs only when there is

matching on more than one variable. Here the combinations of variables may be mismatched in the various groups. If age and reaction time were to be matched, one group might include older participants with slower reaction times and younger participants with quicker reaction times, where the other group has the opposite combination. The mean and distribution of the two variables would be equivalent but the participants in each group would be completely different. This difference may affect the dependent variable.

Matching by Holding Variables Constant

Holding the adventitious variable constant for all experimental groups is another technique that can be used to create equal groups of participants. All participants in each experimental group will have the same degree or type of extraneous variable. If we need to control the variation caused by gender differences, we can hold sex constant by using only males or only females in the study. This has the effect of matching all participants in terms of the sex variable. This matching procedure creates a more homogeneous participant sample, because only participants with a certain type or amount of the fortuitous variable are used. A number of student research projects at universities use this method especially when the participant pool has a majority of male or female participants. This technique of holding variables constant has at least two problems that could affect the validity of the study. The severity of the problem increases if too many variables are held constant. The first disadvantage is that the technique restricts the size of the participant population. Consequently, in some cases, it may be difficult to find enough participants to participate in the study. The early split-brain research of Roger Sperry has often been criticized by the restriction of the participants used in the study. His early studies used only epileptic patients. So a study using this method could be criticized of a selection bias.

The second drawback is more critical in that the results of the study only generalizable to the type of participant used in the study. The results obtained from the epileptic patients study could only be generalized to other epileptic patients. If someone wanted to know whether non-epileptic patients would experience the same perceptual changes the researcher would have to conduct a similar study using non-epileptic patients. Conclusions from such a study might indeed be the same as those obtained from the epileptic patient study but separate studies have to be conducted. The only way we can find out if the results of one study can be generalized to the population is to replicate the study using the participants with different characteristics.

Matching by incorporating the Nuisance Variable into the Research Design

Another way of trying to develop equal groups is to use the nuisance or extraneous variable as an independent variable in the research design. Assume that we were conducting a learning experiment on rats and wanted to control for the effects of weight. The thought here is that the animal with the greater weight will need to consume more food after a period of deprivation and hence is more motivated. If we had used the method of holding weight constant, we would have a lot fewer participants. By using weight as an independent variable, we can use a lot more participants in the study. In statistical terms, the increase in the number of participants means an increase in power and sensitivity. By using an extraneous variable as an independent variable in the design we can isolate a source of systematic variance and also determine if the extraneous variable has an effect on the dependent variable.

However, building an extraneous variable into the design should not be done indiscriminately. Making the extraneous variable a part of the research design seems like an excellent control method. However, this method is best used when there is an interest in the differences produced by the extraneous variable or in the interaction between the extraneous variable and other independent variables. For a variable measured on a continuous scale, a researcher can still incorporate it into the design. The difference between a discrete and continuous extraneous variable would lie in the data analysis part of the research process. With a continuous variable regression or analysis of covariance would be preferable over analysis of variance.

Participant as Own Control

Since each individual is quite unique it is very difficult if not impossible to find another individual that would be a perfect match. However, a single person is a perfect match to oneself. One of the more powerful

techniques for achieving equality or constancy of experimental groups prior to the administration of treatment is to use that person in every condition of the experiment. Some refers to this as using the participants as their own control. Other than the reactivity of the experiment itself, the possibility of extraneous variation due to individual-to-individual differences is drastically minimized. This method of achieving constancy is common in some areas of the behavioral sciences. In psychology, the study of the interface of humans and machines (human factors or human engineering) utilizes this method. Dr. Charles W. Simon (1976) has presented a number of interesting experimental designs, which uses the participant over many treatment conditions. However, this method does not fit all applications. Some studies involved with learning are not suitable because a person cannot unlearn a problem so that he or she can now apply a different method. The use of this method also requires a lot more planning than others.

Additional Design Extensions: Design 9.3 using a Pretest

Design 9.3 has many advantages and is frequently used. Its structure is similar to that of Design 8.2, with two important differences: Design 8.2 lacks a control group and randomization. Design 9.3 is similar to Designs 9.1 and 9.2, except that the "before" or pretest feature has been added. It is used frequently to study change. Like Designs 9.1 and 9.2, it can be expanded to more than two groups.

Design 9.3: Before and after Control Group (Pretest-Posttest)

(a)	[R]	Y_b	X	Y_a	(Experimental)
		Y_b	~X	Y_a	(Control)
(b)	[M_r]	Y_b	X	Y_a	(Experimental)
		Y_b	~X	Y_a	(Control)

In Design 9.3(a), participants are assigned to the experimental group (top line) and the control group (bottom line) at random and are pretested on a measure of Y, the dependent variable. The investigator can then check the equality of the two groups on Y. The experimental manipulation X is performed, after which the groups are again measured on Y. The difference between the two groups is tested statistically. An interesting and difficult characteristic of this design is the nature of the scores usually analyzed: difference, or change, scores, $Y_a - Y_b = D$. Unless the effect of the experimental manipulation is strong, the analysis of difference scores is not advisable. Difference scores are considerably less reliable than the scores from which they are calculated. A clear explanation of why this is so is given by Friedenberg (1995) and Sax (1997). There are other problems. We discuss only the main strengths and weaknesses (see Campbell and Stanley, 1963 for a more complete discussion of this). At the end of the discussion the analytic difficulties of difference or change scores will be taken up.

Probably most important, Design 9.3 overcomes the great weakness of Design 8.2, because it supplies a comparison control group against which the difference, $Y_a - Y_b$, can be checked. With only one group, we can never know whether history, maturation (or both), or the experimental manipulation X produced the change in Y. When a control group is added, the situation is radically altered. After all, if the groups are equated (through randomization), the effects of history and maturation, if present, should be present in both groups. If the mental ages of the children of the experimental group increase, so should the mental ages of the children of the control group. Then, if there is still a difference between the Y measures of the two groups, it should not be due to history or maturation. That is, if something happens to affect the experimental participants between the pretest and the posttest, this something should also affect the participants of the control group. Similarly, the effect of testing-Campbell's reactive measures-should be controlled. For if the testing affects the members of the experimental group it should similarly affect the members of the control group. (There is, however, a concealed weakness here, which will be discussed later.) This is the main strength of the well-planned and well-executed before-after, experimental-control group design.

On the other hand, before-after designs have a troublesome aspect, which decreases both internal and external validity of the experiment. This source of difficulty is the pretest. A pretest can have a sensitizing effect on participants. On internal validity, for example,

the participants may possibly be alerted to certain events in their environment that they might not ordinarily notice. If the pretest is an attitude scale, it can sensitize participants to the issues or problems mentioned in the scale. Then, when the X treatment is administered to the experimental group, the participants of this group may be responding not so much to the attempted influence, the communication, or whatever method is used to change attitudes, as to a combination of their increased sensitivity to the issues and the experimental manipulation.

Since such interaction effects are not immediately obvious, and since they contain a threat to the external validity of experiments, it is worthwhile to consider them a bit further. One would think that, since both the experimental and the control groups are pretested, the effect of pretesting, if any, would ensure the validity of the experiment. Let us assume that no pretesting was done, that is, that Design 9.2 was used. Other things equal, a difference between the experimental and the control groups after experimental manipulation of X can be assumed to be due to X. There is no reason to suppose that one group is more sensitive or more alert than the other, since they both face the testing situation after X. But when a pretest is used, the situation changes. While the pretest sensitizes both groups, it can make the experimental participants respond to X, wholly or partially, because of the sensitivity.

What we also have is a lack of generalizability or external validity in that it may be possible to generalize to pretested groups but not to unpretested ones. Clearly such a situation is disturbing to the researcher, since who wants to generalize to pretested groups?

If this weakness is important, why is this a good design? While the possible interaction effect described above may be serious in some research, it is doubtful that it strongly affects much behavioral research, provided researchers are aware of its potential and take adequate precautions. Testing is an accepted and normal part of many situations, especially in education. It is doubtful, therefore, that research participants will be unduly sensitized in such situations. Still, there may be times when they can be affected. The rule Campbell and Stanley (1963) give is a good one: When unusual testing procedures are to be used, use designs with no pretests.

Difference Scores

Look at Design 9.3 again, particularly at changes between Y_b and Y_a. One of the most difficult problems that has plagued-and intrigued-researchers, measurement specialists, and statisticians is how to study and analyze such difference, or change scores. In a book of the scope of this one, it is impossible to go into the problems in detail. The interested reader can read one or both two excellent edited books written by Harris (1963) and Collins & Horn (1991). General precepts and cautions, however, can be outlined. One would think that the application of analysis of variance to difference scores yielded by Design 9.3 and similar designs would be effective. Such analysis can be done if the experimental effects are substantial. But difference scores, as mentioned earlier, are usually less reliable than the scores from which they are calculated. Real differences between experimental and control groups may be undetectable simply because of the unreliability of the difference scores. To detect differences between experimental and control groups, the scores analyzed must be reliable enough to reflect the differences and thus to be detectable by statistical tests. Because of this difficulty some researchers such as Cronbach and Furby (1970) even say that difference or change scores should not be used. So what can be done?

The generally recommended procedure is to use so-called residualized or regressed gain scores. These are scores calculated by predicting the posttest scores from the pretest scores on the basis of the correlation between pretest and posttest and then subtracting these predicted scores from the posttest scores to obtain the residual gain scores. (The reader should not be concerned if this procedure is not too clear at this stage. Later, after we study regression and analysis of covariance, it should become clearer.) The effect of the pretest scores is removed from the posttest scores; that is, the residual scores are posttest scores purged of the pretest influence. Then the significance of the difference between the means of these scores is tested. All this can be accomplished by using both the procedure just described and a regression equation or by analysis of covariance.

Even the use of residual gain scores and analysis of covariance is not perfect, however. If participants have not been assigned at random to the experimental and control groups, the procedure will not save the situation. Cronbach & Furby (1970) have pointed out that when groups differ systematically before experimental

treatment in other characteristics pertinent to the dependent variable, statistical manipulation does not correct such differences. If, however, a pretest is used, use random assignment and analysis of covariance, remembering that the results must always be treated with special care. Finally, multiple regression analysis may provide the best solution of the problem, as we will see later. It is unfortunate that the complexities of design and statistical analysis may discourage the student of research, sometimes even to the point of feeling hopeless. But that is the nature of behavioral research: it merely reflects the exceedingly complex character of psychological, sociological, and educational reality. This is at one and the same time frustrating and exciting. Like marriage, behavioral research is difficult and often unsuccessful-but not impossible. Moreover, it is one of the best ways to acquire reliable understanding of our behavioral world. The point of view of this book is that we should learn and understand as much as we can about what we are doing, use reasonable care with design and analysis, and then do the research without fussing too much about analytic matters. The main thing is always the research problem and our interest in it. This does not mean a cavalier disregard of analysis. It simply means reasonable understanding and care and healthy measures of both optimism and skepticism.

Design 9.4: Simulated before-after, Randomized

[R]		X	Y_a
	Y_b		

The value of Design 9.4 is doubtful, even though it is considered to among the adequate designs. The scientific demand for a comparison is satisfied: there is a comparison group (lower line). A major weakness of Design 8.3 (a pallid version of Design 9.4) is remedied by the randomization. Recall that with Design 8.3 we were unable to assume beforehand that the experimental and control groups were equivalent. Design 9.4 calls for participants to be assigned to the two groups at random. Thus, it can be assumed that they are statistically equal. Such a design might be used when one is worried about the reactive effect of pretesting, or when, due to the exigencies of practical situations, one has no other choice. Such a situation occurs when one has the opportunity to try a method or some innovation only once. To test the method's efficacy, one provides a base line for judging the effect of X on Y by pretesting a group similar to the experimental group. Then Y_a is tested against Y_b.

This design's validity breaks down if the two groups are not randomly selected from the same population or if the participants are not assigned to the two groups at random. Furthermore, even if randomization is used there is no real guarantee that it worked in equating the two groups prior to treatment. It has the weaknesses mentioned in connection with other similar designs, namely, other possible variables may be influential in the interval between Y_b and Y_a. In other words, Design 9.4 is superior to Design 8.3, but it should not be used if a better design is available.

Design 9.5: Three-Group, before-after

	Y_b	X	Y_a	(Experimental)
[R]	Y_b	~X	Y_a	(Control 1)
		X	Y_a	(Control 2)

Design 9.5 is better than Design 9.4. In addition to the assets of Design 9.3 it provides a way to possibly avoid confounding due to the interactive effects due to the pretest. This is achieved by the second control group (third line). (It seems a bit strange to have a control group with an X, but the group of the third line is really a control group.) With the Y_a measures of this group available, it is possible to check the interaction effect. Suppose the mean of the experimental group is significantly greater than the mean of the first control group (second line). We may doubt whether this difference was really due to X. It might have been produced by increased sensitization of the participants after the pretest and the interaction of their sensitization and X. We now look at the mean of Y_a of the second control group (third line). It, too, should be significantly greater than the mean of the first control group. If it is, we can assume that the pretest has not unduly sensitized the participants, or that X is sufficiently strong to override a sensitization-X interaction effect.

Design 9.6: Four-Group, before-after (Solomon)

	Y_b	X	Y_a	(Experimental)
[R]	Y_b	~X	Y_a	(Control 1)
		X	Y_a	(Control 2)
		~X	Y_a	(Control 3)

This design, proposed by Solomon (1949) is strong and aesthetically satisfying. It has potent controls. Actually, if we change the designation of Control 2 to Experimental 2, we have a combination of Designs 9.3 and 9.1, our two best designs, where the former design forms the first two lines and the latter the second two lines. The virtues of both are combined in one design. Although this design can have a matching form, it is not discussed here nor is it recommended. Campbell (1957) says that this design has become the new ideal for social scientists. While this is a strong statement, probably a bit too strong, it indicates the high esteem in which the design is held.

Among the reasons why it is a strong design is that the demand for comparison is well satisfied with the first two lines and the second two lines. The randomization enhances the probability of statistical equivalence of the groups, and history and maturation are controlled with the first two lines of the design. The interaction effect due to possible pretest participant sensitization is controlled by the first three lines. By adding the fourth line, temporary contemporaneous effects that may have occurred between Y_a and Y_b can be controlled. Because Design 9.1 and 9.3 are combined, we have the power of each test separately and the power of replication because, in effect, there are two experiments. If Y_a of Experimental is significantly greater than Control 1, and Control 2 is significantly greater than Control 3, together with a consistency of results between the two experiments, this is strong evidence, indeed, of the validity of our research hypothesis.

What is wrong with this paragon of designs? It certainly looks fine on paper. There seem to be only two sources of weakness. One is practicability-it is harder to run two simultaneous experiments than one and the researcher encounters the difficulty of locating more participants of the same kind.

The other difficulty is statistical. Note that there is a lack of balance of groups. There are four actual groups, but not four complete sets of measures. Using the first two lines, that is, with Design 9.3, one can subtract Y_b from Y_a or do an analysis of covariance. With the two lines, one can test the Y_a 's against each other with a t test or F test, but the problem is how to obtain one overall statistical approach. One solution is to test the Y_a 's of Controls 2 and 3 against the average of the two Y_b's (the first two lines), as well as to test the significance of the difference of the Y_a 's of the first two lines. In addition, Solomon originally suggested a 2×2 factorial analysis of variance, using the four Y_a sets of measures. Solomon's suggestion is outlined in Figure 9.5. A careful study will reveal that this is a fine example of research thinking, a nice blending of design and analysis. With this analysis we can study the main effects, X and ~X, and Pretested and Not Pretested. What is more interesting, we can test the interaction of pretesting and X and get a clear answer to the previous problem.

	X	~X
Pretested	Y_a, Experimental	Y_a, Control 1
Not Pre-tested	Y_a, Control 2	Y_a, Control 3

Figure 9.5

While this and other complex designs have decided strengths, it is doubtful that they can be used routinely. In fact, they should probably be saved for very important experiments in which, perhaps, hypotheses already tested with simpler designs are again tested with greater rigor and control. Indeed, it is recommended that designs like 9.5 and 9.6 and certain variants of Designs 9.6, to be discussed later, be reserved for definitive tests of research hypotheses after a certain amount of preliminary experimentation has been done.

Concluding Remarks

The designs of this chapter are general: they are stripped down to bare essentials to show underlying structure. Having the underlying structures well in mind-cognitive psychologists say that such structures are important in remembering and thinking-the student is in a position to use the more specific designs of analysis of variance and related paradigms. Knowing and understanding the general designs may enhance mental flexibility and the ability to cope conceptually and practically with

research problems and the design means of solving the problems.

Study Suggestions

1. The first sentence of this chapter is "Design is data discipline." What does this sentence mean? Justify it.

2. Suppose you are an educational psychologist and plan to test the hypothesis that feeding back psychological information to teachers effectively enhances the children's learning by increasing the teachers' understanding of the children. Outline an ideal research design to test this hypothesis, assuming that you have complete command of the situation and plenty of money and help. (These are important conditions, which are included to free the reader from the practical limitations that so often compromise good research designs.) Set up two designs, each with complete randomization, both following the paradigm of Design 9.1. In one of these use only one independent variable and one-way analysis of variance. In the second, use two independent variables and a simple factorial design. How do these two designs compare in their control powers and in the information they yield? Which one tests the hypothesis better? Why?

3. Design research to test the hypothesis of Study Suggestion 2, above, but this time compromise the design by not having randomization. Compare the relative efficacies of the two approaches. In which of them would you put greater faith? Why? Explain in detail.

4. Suppose that a team of sociologists, psychologists, and educators believed that competent and insightful counseling can change the predominantly negative attitudes of juvenile offenders for the better. They took 30 juvenile offenders-found to be so by the courts-who had been referred for counseling in the previous year and matched each of them to another nonoffender youngster on sex and intelligence. They compared the attitudes of the two groups at the beginning and the end of the year (the duration of the counseling), and found a significant difference at the beginning of the year but no significant difference at the end. They concluded that counseling had a salutary effect on the juvenile offenders' attitudes. Criticize the research. Bring out its strengths and weaknesses. Keep the following in mind: sampling, randomization, group comparability, matching, and control. Is the conclusion of the researchers empirically valid, do you think? If not, outline a study that will yield valid conclusions.

5. The advice in the text not to use analysis of variance in nonexperimental research does not apply so much to one-way analysis of variance as it does to factorial analysis. Nor does the problem of equal numbers of cases in the cells apply (within reason). In a number of nonexperimental studies, in fact, one-way analysis of variance has been profitably used. One such study is: Jones & Cook (1975). The independent variable was attitude toward African-Americans, obviously not manipulated. The dependent variable was preference for social policy affecting African-Americans: remedial action involving social change or action involving self-improvement of African-Americans. One-way analysis of variance was used with social policy preference scores of four groups differing in attitudes toward African-Americans. (Attitudes toward African-Americans were also measured with an attitude scale.)

It is suggested that students read and digest this excellent and provocative study. It will be time and effort well-spent. You may also want to do an analysis of variance of the data of the authors' Table 1, using the method outlined earlier of analysis of variance using n's, means, and standard deviations (see Addendum, Chapter 13).

Chapter Summary

1. The design of a study is its blueprint or plan for the investigation.
2. A design is a subset of a Cartesian cross-product of levels of the independent variable.
3. The experimental design is the one where at least one of the independent variables for the study is manipulated.
4. Non experimental designs are those where there is no randomization to equate the groups prior to administering treatment.
5. For experimental designs, usually the most appropriate statistical method is analysis of variance.
6. The assumptions of the analysis of variance are usually violated for non-experimental designs. Multiple regression may be a more appropriate method of analyzing data from non-experimental

designs.

7. The experimental group-control group design with randomized participants (Design 9.1) is the best design for many experimental behavioral research studies.
8. The Solomon four group design (Design 9.6) is a design that handles many of the concerns of behavioral research. However, it uses the resources of two studies and may not be economically efficient.
9. Design 9.2 is like Design 9.1 except it uses matched participants.
10. The use of matched participants is useful in some situations where randomization will not work properly.
11. There is several way of matching participants. The most popular is the individual-by-individual method.
12. Matching has problems in that the researcher is never sure that all the important variables have been used in the match. Additionally, if too many variables are used to match, it becomes more difficult to get participants that match.
13. Design 9.3 uses a pretest. Pretesting is one way of determining if the groups are equal or whether randomization has worked. However, pretesting also sensitizes the participants to the experiment.
14. Difference scores are often used in designs using a pretest. However, there are some problems with doing this. Namely, difference scores can be unreliable.
15. Design 9.4 is a simulated before after design using randomized participants. The second (control) group is only measured on a pretest. The experimental group receives treatment and the posttest.
16. Design 9.5 is a 3 group before-after design. It is just like design 9.3 except a third group receiving treatment and no pretest is used.

Chapter 10

Research Design Applications: Randomized Groups and Correlated Groups

IT is difficult to tell anyone how to do research. Perhaps the best thing to do is to make sure that the beginner has a grasp of principles and possibilities. In addition, approaches and tactics can be suggested. In tackling a research problem, the investigator should let one's mind roam, speculate about possibilities, even guess the pattern of results. Once the possibilities are known, intuitions can be followed and explored. Intuition and imagination, however, are not much help if we know little or nothing of technical resources. On the other hand, good research is not just methodology and technique. Intuitive thinking is essential because it helps researchers arrive at solutions that are not merely conventional and routine. It should never be forgotten, however, that analytic thinking and creative intuitive thinking both depend on knowledge, understanding, and experience.

The main purposes of this chapter are to enrich and illustrate our design and statistical discussion with actual research examples and to suggest basic possibilities for designing research so that the student can ultimately solve research problems. Our summary purpose, then, is to supplement and enrich earlier, more abstract design and statistical discussions.

SIMPLE RANDOMIZED SUBJECTS DESIGN

In Chapters 13 and 14 the statistics of simple one-way and factorial analysis of variance were discussed and illustrated. The design behind the earlier discussions is called *randomized* subjects *design.* The general design paradigm is Design 9.1:

[R]	X	Y	(Experimental)
	~X	Y	(Control)

Research Examples

The simplest form of Design 9.1 is a one-way analysis of variance paradigm in which k groups are given k experimental treatments and the *k* means are compared with analysis of variance or separate tests of significance. A glance at Figure 9.3, left side, shows this simple form of 9.1 with $k = 3$. Strange to say, it is not used too often. Researchers more often prefer the factorial form of Design 9.1. Two one-way examples are given below. Both used random assignment. Unfortunately, some researchers do not report how participants were assigned to groups or treatments. The need to report the method of participant selection and assignment to experimental groups should by now be obvious.

Dolinski & Nawrat: Fear-then-Relief and Compliance.

Studies on compliance have been of great interest to social psychologists. In Chapter 17 where we discussed ethics of doing behavioral science research we mentioned the influence of the Milgrim study on how we now do behavioral research. Milgrim, if you recall was interested in why the Nazis during World War 2 complied to commit unspeakable acts of brutality to other humans. In a study by Dolinski and Nawrat (1998), they explored another method that was used to induce compliance. This was a method used by Nazis and Stalinists to get Polish prisoners to testify against themselves, their friends and/or their families. Dolinski and Nawrat call this method as "fear-then-relief." This method involves putting a prisoner in a high state of anxiety. This is done by yelling, screaming and threatening the prisoner by jailers. After achieving this with the prisoner, the anxiety producing stimuli are abruptly removed. The prisoner is then treated kindly. The usual result of this procedure is the intensification of compliance behavior. Dolinski and Nawrat claim that compliance is due to the reduction of fear and not the fear itself. Although Dolinski and Nawrat use a very extreme example to illustrate their point, they also explain that the method is often used in some shape and form by dyads in everyday life. This can occur between parent and child, teacher and student, and employer and employee. Police often use similar tactics with their "good-cop, bad-cop" routine. This routine usually involves one police officer ("bad-cop") berating,

screaming and threatening a prisoner. When the prisoner reaches a level of high anxiety, another police officer ("good-cop") removes the "bad-cop" and talks kindly and sweetly to the prisoner. This method is also used by terrorists on hostages.

Dolinski and Nawrat designed and conducted four experiments to test the "fear-then-relief' method's ability to induce compliance. We will describe one of those experiments here. In this experiment, 120 volunteer high school students from Opole, Poland were randomly assigned to one of three experimental conditions. All participants were told that they were to partake in a study on the effects of punishment on learning. Group 1 experienced anxiety. They were told that they would be given mild, not painful electrical shock for every error they made. Group 2 participants experienced anxiety that was then reduced. They were initially given the same description as Group I, but then later were told that they would participate in a different study instead. This other study involved visual-motor coordination and no shock will be given. Group 3 was the control condition. These participants were told that they would be participating in a visual-motor coordination study. During the waiting period before the start of the experiment, each participant was asked to complete an anxiety questionnaire. After completing the questionnaire. A female student who was a confederate of the experimenter, but appeared to be totally unattached from the experiment introduced herself and asked each participant to join a charity action for an orphanage. For those that complied, they were asked how many hours they were willing to work for this action.

The manipulated independent variable in this study was the level of induced anxiety and relief. The dependent variables were compliance; amount of anxiety and the number of hours of donated time for a good cause. Using a one-way analysis of variance, Dolinski and Nawrat obtained a significant F-value. Group 2, the group that felt anxiety and then had it reduced had the highest rate of compliance and were willing to donate the greatest number of days. The level of anxiety for each group was in the expected direction. Group 1 experienced the highest degree of anxiety, followed by Group 2 and then Group 3. Table 10.1 presents the summary data for the study.

The results of the study upheld Dolinski and Nawrat's hypothesis that it was the "fear-then-relief" and not the emotion of anxiety itself that led to a higher degree of compliance. Simply creating a state of anxiety in people is not enough to create compliance. In fact, this study found that the participants in Group 1 (induced anxiety) who felt the greatest amount of anxiety complied less than the participants in Group 3 (control-low or no anxiety).

Hilliard, Nguyen and Domjan: One-trial Learning of Sexual Behavior

Can an animal be taught sexual behavior in one learning trial? This was the problem investigated by Hilliard, Nguyen and Domjan (1997). Using a classical conditioning paradigm, Hilliard, Nguyen and Domjan sought to show that sexual reinforcement can produce one-trial learning. The study is significant in that one-trial learning using classical conditioning is unusual outside the use of noxious or aversive stimuli. Hilliard, Nguyen and Domjan performed two studies, both of which are true experiments.

Table 10.1 Anxiety levels, compliance, number of days willing to volunteer by induced anxiety and F-values (Dolinski & Nawrat (1998) study).

Condition Group	Mean Anxiety Reported	Percentage Complying	Mean # days volunteered
Electrical shock study	53.25	37.5	0.625
Electrical shock study changed to visual-motor coordination study	43.05	75.0	1.150
Visual-motor coordination study	34.45	52.5	1.025
F- value	108.9 ($p < .001$)	6.13 ($p < .003$)	2.11 ($p > .05$)

In the first study, eighteen male quails were used as participants. They were randomly assigned to one of two groups: Group Paired or Group Unpaired. The learning trial for the Group Paired participants consisted of being placed in a completely dark experimental chamber for 4 minutes followed by a 35-second exposure to a test object (conditioned stimulus or CS).

After exposure to the test object, the male quail was allowed to interact with a sexually receptive female quail (unconditioned stimulus or US). The test object consisted of a mock figure of a quail made of stuffed terry cloth ovoid mounted with a taxidermically prepared female quail head. In the Group Unpaired, participants experienced exactly the same procedure as Group Paired except they were not exposed to the sexually receptive female quail (no US). The effect of conditioning was assessed the following day by placing the participant in the experimental chamber with the test object only. The dependent variables for the study were (1) the amount of time the male quail waited until grabbing the female quail, (2) the number of cloacal contact movements and (3) the amount of time in the test zone during the two minute test period. The test zone is the area around the test object. Hilliard, Nguyen and Domjan deemed the analysis of the third dependent variable the most important. A one-way analysis of variance was used to test the difference between the two conditions. Hilliard, Nguyen and Domjan found that participants from the Group Paired condition spent a longer amount of time in the test zone and they mounted the test object more frequently than the Group Unpaired participants. Table 10.2 gives the summary statistics Study 1.

In their second study, Hilliard, Nguyen and Domjan wanted to determine if sexual conditioning is directly related to the duration of context exposure prior to presentation of US. Context exposure is the process of exposing the participant to the test object. Twenty-six male quails were randomly assigned to three context exposure conditions. In Study 1, the context exposure was 35 seconds

Table 10.2 Means for Group Paired and Group Unpaired on Time Spent in Test Zone and Number of Mountings on Test Object (Hilliard, Nguyen and Domjan Data)

Condition	Time Spent in Test Zone	Number of Mounts on Test Object
Paired	60.70	3.00
Unpaired	3.37	0.00
F-value	7.41 ($p < .05$)	4.86 ($p < .05$)

. In this study, one group of male quails was allowed to interact with a sexually receptive female quail after 0 minutes of exposure to the test object. Another group was allowed 2 minutes of exposure and the third group was given 4 minutes. There was also a fourth group that received no pairing of context exposure and access to the female quail. During the test session where the participant was allowed to interact with the test object, participants in all three conditions showed strong approach to the test object. For the unpaired group, there was considerably less time spent in the test zone. A one-way analysis of variance was used to analyze the data and the F-value ($F = 5.92$, $p < .01$) was statistically significant. There was no statistical difference between the three conditions, however, there were significant differences when each condition was compared to the unpaired group. Hilliard, Nguyen and Domjan claims to have presented the first clear demonstration of one-trial classical conditioning using an appetitive unconditioned stimulus.

Table 10.3 Mean Time Spent in Test Zone for the 3 Context Exposure Rates

Condition	Zero Minutes	2 Minutes	4 minutes	Unpaired
Time spent in Test Zone	70.96	86.06	85.44	14.99

FACTORIAL DESIGNS

The basic general design is still Design 9.1, though the variation of the basic experimental group-control pattern is drastically altered by the addition of other experimental factors or independent variables. Following an earlier definition of factorial analysis of

variance, *factorial design is the structure of research in which two or more independent variables are juxtaposed in order to study their independent and interactive effects on a dependent variable.*

The reader may at first find it a bit difficult to fit the factorial framework into the general experimental group-control group paradigm of Design 9.1. The discussion of the generalization of the control-group idea in Chapter 9 however, should have clarified the relations between Design 9.1 and factorial designs. The discussion is now continued. We have the independent variables A and B and the dependent variable Y. The simplest factorial design, the 2 × 2, has three possibilities: both A and B active; A active, B attribute (or vice versa); and both A and B attribute. (The last possibility, both independent variables attributes, is the nonexperimental case. As indicated earlier, however, it is probably not appropriate to use analysis of variance with nonexperimental independent variables.) Returning to the experimental group-control notion, A can be divided into A_1 and A_2, experimental and control, as usual, with the additional independent variable B partitioned into B_1 and B_2. Since this structure is familiar to us by now, we need only discuss one or two procedural details.

The ideal participant assignment procedure is to assign participants to the four cells at random. If both A and B are active variables, this is possible and easy. Simply give the participants numbers arbitrarily from 1 through N, N being the total number of participants. Then, using a table of random numbers, write down numbers 1 through N as they turn up in the table. Place the numbers into four groups as they turn up and then assign the four groups of participants to the four cells. To be safe, assign the groups of participants to the experimental treatments (the four cells) at random, too. Label the groups l, 2, 3, and 4. Then draw these numbers from a table of random numbers. Assume that the table yielded the numbers in this order: 3, 4, 1, and 2. Assign Group 3 participants to the upper left cell, Group 4 participants to the upper right cell, and so on.

Often B will be an attribute variable, like gender, intelligence, achievement, anxiety, self perception, race, and so on. The participant assignment must be altered. First, since B is an attribute variable, there is no possibility of assigning participants to B_1 and B_2 at random. If B were the variable gender, the best we can do is to assign males first at random to the cells A_1B_1 and A_2B_1, and then females to the cells A_1B_2 and A_2B_2.

Factorial Designs with More than Two Variables

We can often improve the design and increase the information obtained from a study by adding groups. Instead of A_1 and A_2, and B_1 and B_2, an experiment may profit from A_1, A_2, A_3 and A_4, and B_1, B_2 and B_3. Practical and statistical problems increase and sometimes become quite difficult as variables are added. Suppose we have a 3 × 2 × 2 design that has 3 × 2 × 2 = 12 cells, each of which has to have at least two participants, and preferably many more. (It is possible, but not very sensible, to have only one participant per cell if one can have more than one. There are, of course, designs that have only one participant per cell.) If we decide that 10 participants per cell are necessary, 12 × 10 = 120 participants will have to be obtained and assigned at random. The problem is more acute with one more variable and the practical manipulation of the research situation is also more difficult. But the successful handling of such an experiment allows us to test a number of hypotheses and yields a great deal of information. The combinations of three-, four-, and five-variable designs give a wide variety of possible designs: 2 × 5 × 3, 4 × 4 × 2, 3 × 2 × 4 × 2, 4 × 3 × 2 × 2, and so on.'

Research Examples of Factorial Designs

Examples of two- and three-dimensional factorial designs were described in Chapter 14. (The restudy of these examples is recommended, because the reasoning behind the essential design can now be more easily grasped.) Since a number of examples of factorial designs were given in Chapter 14, we confine the examples given here to studies with unusual features or interesting results.

Flowers: Groupthink

In a highly provocative article, "Group Think," Janis (1971) discussed the possible deleterious consequences of the drive for concurrence (often called consensus) in cohesive groups. He said that consensus-seeking becomes so dominant in cohesive in groups that it

overrides realistic appraisal of alternative courses of action. In support of his thesis, Janis cited the Bay of Pigs, the Vietnam war, and other "fiascoes." The article and its arguments are impressive. Will it hold up under experimental testing? Flowers (1977) tested Janis' basic hypothesis in the laboratory. Flowers' hypothesis was that cohesiveness and leadership style in groups interacts to produce groupthink. That is, in high cohesive groups with closed leadership groupthink, as measured by number of problem solutions proposed and the use of information outside the groups, will develop. Cohesiveness was operationalized as follows: groups of acquaintances = high cohesive; groups of strangers = low cohesive. Style of leadership was open-leader encouraged divergent opinions and emphasized wise decisions-or closed-leader encouraged unanimity at all costs and focused on the preferred solution of the leader. The dependent variables were: number of suggested solutions to problems and the use of information outside the group. (There were, of course, many more details in operationalizing these variables.)

Part of the results obtained is given in Table 10.4. These data were obtained from 40 *groups,* with 10 groups in each cell. The unit of analysis was the group, therefore, an unusual feature of this research. The only effect that was significant was open and closed leadership, as shown by the means of (a), 6.45 and 5.15 and (b),16.35 and 11.75. The predicted interaction between leadership style and cohesiveness did not emerge in either set of data. Evidently style of leadership is the crucial variable. Part of Janis' thesis was supported.

Table 10.4 Mean Numbers of Solutions Proposed, (a), and Emergent Facts, (b), Flowers Study[a]

(a)

	Open	Closed	
High Coh.	6.7	4.94	5.82
Low Coh.	6.2	5.35	5.78
	6.45	5.15	

(b)

	Open	Closed	
High Coh.	16.8	11.8	14.3
Low Coh.	15.9	11.7	13.8
	16.35	11.75	

[a]N = 40 groups, 4 in each group. The layout of the data is mine, as are the calculations of the marginal means. F_a (open-closed) = 6.44 ($p < .05$); F_b (open-closed) =16.76 ($p < .01$).

The student should particularly note that group measures were analyzed. Also note the use of two dependent variables and two analyses of variance. That the main effect of open and closed leadership was significant with number of proposed solutions and with facts used is much more convincing than if only one of these had been used. An interesting and potentially important experiment! Indeed, Flowers' operationalization of Janis' ideas of groupthink and its consequences is a good example of experimental testing of complex social ideas. It is also a good example of replication and of Design 9.1 in its simplest factorial form.

Sigall and Ostrove: Attractiveness and Crime

It has often been said that attractive women are reacted to and treated differently than men and less attractive women. In most cases, perhaps, the reactions are "favorable": attractive women are perhaps more likely than less attractive women to receive the attention and favors of the world. Is it possible, however, that their attractiveness may in some situations be disadvantageous? Sigall and Ostrove (1975) asked the question: How is physical attractiveness a criminal defendant related to juridic sentences, and does the nature of the crime interact with attractiveness? They had their participants assign sentences, in years, to swindle and burglary offenses of attractive, unattractive, and control defendants. The factorial paradigm of the experiment, together with the results, is given in Table 10.5. (We forego describing many of the experimental details; they were well handled.)

In the burglary case, the defendant stole $2,200 in a high-rise building. In the swindle case, the defendant ingratiated herself with and swindled a middle-aged bachelor of $2,200. Note that the Unattractive and Control conditions did not differ much from each other. Both Attractive-Swindle (5.45) and Attractive-Burglary (2.80) differed from the other two conditions-but in opposite directions! Attractive-Swindle received the heaviest mean sentence: 5.45 years, whereas Attractive-Burglary received the lowest mean sentence: 2.80 years. The statistics support the preceding verbal summary:

the interaction was statistically significant: The Attractiveness-Offense F, at 2 and 106 degrees of freedom, was 4.55, $p < .025$. In words, attractive defendants have an advantage over unattractive defendants, except when their crimes are attractiveness-related (swindle).

Table 10.5 Mean Sentences in Years of Attractive, Unattractive, and Control Defendants for Swindle and Burglary, Sigall and Ostrove Study

	Defendant Condition		
	Attractive	Unattractive	Control
Swindle	5.45	4.35	4.35
Burglary	2.80	5.20	5.10

[a]N =120, 20 per cell. F (interaction) = 4.55 ($p < .025$).

Zakay, Hayduk and Tsal: Personal space and distance misperception.

How much of an invasion of our personal space is perceived when a person approaches us or when they are departing us? Are these two perceptions different? Zakay, Hayduk and Tsal (1992) say yes. The results of their study were contrary to current theories of personal space. They found that distance perception was more distorted in an invading (approaching) condition than in a departing condition. Thirty-two female and 32 male participants were used in this study. The participants gender and the experimenter's gender were two non-manipulated (measured) independent variables. The manipulated or active independent variable was the approaching-departing condition. Sixteen females and 16 males were randomly assigned to either the approaching condition or the departing condition. In the approaching condition, the experimenter was placed 210 centimeters (82.67 inches) away from the participant. The experimenter proceeded to walk slowly toward the participant. The participant was instructed to ask the experimenter to stop when he or she felt the experimenter was 140 centimeters (55.1 inches) away from the participant. In the departing condition, the experimenter was placed 70 centimeters (27.5 inches) sway from the participant. The experimenter proceeded to walk slowly away from the participant. The participant was instructed to ask the experimenter to stop when he or she felt the experimenter was 140 centimeters (55.1 inches) away from the participant. The dependent variable was the actual distance measurement between the experimenter and the participant when the participant asked the experimenter to stop.

The design for the study was a $2 \times 2 \times 2$ factorial. A three-way analysis of variance showed no sex-of participant effect, no sex of experimenter effect and no interaction effect. However there was an effect for the approaching-departing condition. Participants halted the experimenter at significantly larger distances in the approaching than in the departing condition. The F-value for 1 and 56 degrees of freedom was 28.01 and was significant at the $p < .01$ level. Table 10.6 gives the summary statistics for the study. Zakay, Hayduk and Tsal conclude that the "impending" or "anticipating" system seems to account for the stronger approaching condition.

Table 10.6 Means of Measured Distances between Experimenter and Participant (in centimeters) from Zakay, Hayduk and Tsal study.

	Female Participant	
Conditions	Male Experimenter	Female Experimenter
Approaching	177.12	187.12
Departing	151.75	167.50
Combined	164.43	177.31

	Male Participant		
Conditions	Male Experimenter	Female Experimenter	Combined
Approaching	172.11	174.42	177.62
Departing	156.66	156.00	158.00
Combined	164.38	165.21	

Quilici and Mayer: Examples, schema and learning.

Do examples help students learn statistics? This was the basic question posed by cognitive scientists Quilici and Mayer (1996). In their study on analytic problem solving Quilici and Mayer examined only one of three

processes that define analogical thinking. These researchers were concerned only with the recognition process that involves two techniques: (1} focus on the surface similarities between the example and actual problem to be solved or (2) focus on the structural similarities. Surface similarities deal with the shared attributes of objects in the problem cover story. With structural similarity, the concern is with the shared relations between objects in both example and problem. To study this phenomena, Quilici and Mayer used learning how to solve word problems in statistics. Quilici and Mayer felt that students who learn the structure of statistical word problems will be better able to solve other problems they encounter in the future by properly classifying them into the correct statistical method of analysis, e.g. t-test, correlation, etc. A few examples are given below to illustrate the difference between surface and structural similarities.

Example 1: A personnel expert wishes to determine whether experienced typists are able to type faster than inexperienced typists. Twenty experienced typists and 20 inexperienced typists are given a typing test. Each typist's average number of words typed per minute is recorded.

Example 2: A personnel expert wishes to determine whether typing experience goes with faster typing speeds. Forty typists are asked to report how many years they have worked as typists and are given a typing test to determine their average number of words typed per minute.

Example 3: After examining weather data for the last 50 years, a meteorologist claims that the annual precipitation varies with average temperature. For each of 50 years, she notes the annual rainfall and average temperature.

Example 4: A college dean claims that good readers earn better grades than poor readers. The grade point average is recorded for 50 first-year students who scored high on the reading comprehension test and for 50 first year students who scored low on a reading comprehension test.

In examining these four problems taken from Quilici and Mayer (1996, p.146). Example 1 and Example 2 would have the same surface features. Both deal with typists and typing. To solve Example 1 a t-test would be used to compare experienced with inexperienced typists. However, to solve Example 2, one would use a correlation since the question asks for a relation between typing experience and average number of words typed per minute. Hence Example 1 and Example two would be different structurally. Example 3 also looks at the relation between two variables: amount of rainfall and temperature. It would have the same structure as Example 2 but a different surface. They have the same structure because both require the use of correlation to solve the problem. Example 4 and Example 1 have the same structure but a different surface.

Quilici and Mayer (1996) designed a study to determine if experience with examples foster structural schema construction. One of their hypotheses stated that students who are exposed to statistical word problem examples are more likely to sort future problems on the basis of structure and less likely to sort on the basis of surface features. Students who are not exposed to statistical word problem examples will not exhibit such behavior. They also hypothesized that those exposed to three examples will be able to exhibit the behavior to a higher degree than those exposed to only one example. These researchers used a 3×4 factorial design. The first independent variable was structural characteristics (t-test, chi-square and correlation). The second independent variable was surface characteristics (typing, weather, mental fatigue and reading). There were two dependent variables: a structure usage score and a surface usage score. Participants were randomly assigned to treatment conditions. A two way analysis of variance confirmed their hypothesis that those exposed to example would use a structure based schema while those not exposed to examples would not. However there was no statistical difference between those that were exposed to three examples and those that got one example. Table 10.7 gives the summary statistics for the study.

Table 10.7 Mean Structure and Surface Scores by Number of Examples. Quilici & Mayer Study.

	3-examples	1-example	No example	
Structure Score	.327	.323	.049	F = 35.93, p < .001
Surface Score	.441	.488	.873	F = 17.82, p < .001

S

Hoyt: Teacher Knowledge and Pupil Achievement

We now outline an educational study done many years ago because it was planned to answer an important theoretical and practical question and because it clearly illustrates a complex factorial design. The research question was: What are the effects on the achievement and attitudes of pupils if teachers are given knowledge of the characteristics of their pupils? Hoyt's (1955) study explored several aspects of the basic question and used factorial design to enhance the internal and external validity of the investigation. The first design was used three times for each of three school participants and the second and third was used twice, once in each of two school systems.

The paradigm for the first design is shown in Figure 10.1. The independent variables were treatments, ability, gender, and schools. The three treatments were no information (N), test scores (T), and test scores plus other information (TO). These are self-explanatory. Ability levels were high, medium, low IQ. The variables gender and schools are obvious. Eighth-grade students were assigned at random within gender and ability levels. It will help us understand the design if we examine what a final analysis of variance table of the design looks like. Before doing so, however, it should be noted that the achievement results were mostly indeterminate (or negative). The F ratios, with one exception, were not significant. Pupils' attitudes toward teachers, on the other hand, seemed to improve with increased teacher knowledge of pupils, an interesting and potentially important finding. The analysis of variance table is given in Table 10.8. One experiment yields 14 tests! Naturally, a number of these tests are not important and can be ignored. The tests of greatest importance (marked with asterisks in the table) are those involving the treatment variable. The most important test is between treatments, the first of the main effects. Perhaps equally important are the interactions involving treatments. Take the interaction treatments × sex. If this was significant, it would mean that the amount of information a teacher possesses about students has an influence on student achievement, but boys are influenced differently than girls. Boys with teachers who possess information about their pupils may do better than boys whose teachers do not have such information, whereas it may be the opposite with girls, or it may make no difference one way or the other.

Second-order or triple interactions are harder to interpret. They seem to be rarely significant. If they are significant, however, they require special study. Crosstabulation tables of the means are perhaps the best way, but graphic methods, as discussed earlier, are often enlightening. The student will find guidance in Edwards' (1984) book or Simon's (1976) manuscript.

		N		T		TO	
		M	F	M	F	M	F
School A	High IQ						
	Medium IQ						
	Low IQ						
School B	High IQ						
	Medium IQ						
	Low IQ						

Dependent Variable Measures

Figure 10.1

Table 10.8 Sources of Variance and Degrees of Freedom for a 3 × 3 × 2 × 2 Factorial Design with Variables Treatments, Ability, Sex, and School (Total and Within Degrees of Freedom are Omitted)

Source	*df*
Main Effects	
*Between Treatments	2
Between Ability Levels	2
Between Gender	1
Between Schools	1
First-Order Interactions	
*Interaction: Treatments × Ability	4
*Interaction: Treatments × Gender	2

*Interaction: Treatments × School 2
Interaction: Ability × Gender 2
Interaction: Ability × School 2
Interaction: Gender × School 1

Second-Order Interactions:

*Interaction: Treatments × Ability × Gender 4
*Interaction: Treatments × Ability × School 4
Interaction: Ability × Gender × School 2

Third-Order Interactions:
Interaction: Treatments × Ability ×
Gender × School 4

Within or Residual

Total

EVALUATION OF RANDOMIZED SUBJECTS DESIGNS

Randomized subjects designs are all variants or extensions of Design 9.1, the basic experimental group-control group design in which participants are assigned to the experimental and control groups at random. As such they have the strengths of the basic design, the most important of which is the randomization feature and the consequent ability to assume the preexperimental approximate equality of the experimental groups in all possible independent variables. History and maturation are controlled because very little time elapses between the manipulation of X and the observation and measurement of Y. There is no possible contamination due to pretesting.

Two other strengths of these designs, springing from the many variations possible, are flexibility and applicability. They can be used to help solve many behavioral research problems, since they seem to be peculiarly well suited to the types of design problems that arise from social scientific and educational problems and hypotheses. The one-way designs, for example, can incorporate any number of methods, and the testing of methods is a major educational need. The variables that constantly need control in behavioral research-gender intelligence, aptitude, social class, schools, and many others-can be incorporated into factorial designs and thus controlled. With factorial designs, too, it is possible to have mixtures of active and attribute variables another important need.

There are also weaknesses. One criticism has been that randomized subjects designs do not permit tests of the equality of groups, as do before-after designs. Actually, this is not a valid criticism for two reasons: with enough participants and randomization, it can be assumed that the groups are equal, as we have seen; and it is possible to check the groups for equality on variables other than Y, the dependent variable. For educational research, data on intelligence, aptitude, and achievement, for example, are available in school records. Pertinent data for sociology and political science studies can often be found in county and election district records.

Another difficulty is statistical. One should have equal numbers of cases in the cells of factorial designs. (It is possible to work with unequal *n*'s, but it is both clumsy and a threat to interpretation. Dropping out cases at random or the use of missing data methods (see Dear, 1959; Gleason & Staelin, 1975 for two excellent references on estimating missing data.) can cure small discrepancies.) This imposes a limitation on the use of such designs; because it is often not possible to have equal numbers in each cell. One-way randomized designs are not so delicate: unequal numbers are not a difficult problem. How to adjust and analyze data for unequal n's is a complex, thorny, and much-argue problem. For a discussion in the context mostly of analysis of variance, see Snedecor and Cochran (1989). Discussion in the context of multiple regression, which is actually a better solution of the problem, can be found in Kerlinger and Pedhazur (1973) and Pedhazur (1996). Pedhazur's discussions are detailed and authoritative. He reviews the issues and suggests solutions.

Compared to matched groups designs, randomized subjects designs are usually less precise, that is, the error term is ordinarily larger, other things equal. It is doubtful, however, whether this is cause for concern. In some cases it certainly is-for example, where a very sensitive test of a hypothesis is needed. In much behavioral research, though, it is probably desirable to consider as nonsignificant any effect that is insufficiently powerful to make itself felt over and above the random noise of a randomized subjects design.

All in all, then, these are powerful, flexible, useful and widely applicable designs. In the opinion of the

authors they are the best all-round designs, perhaps the first to be considered when planning the design of a research study.

CORRELATED GROUPS

A basic principle is behind all correlated-groups' designs: there is systematic variance in the dependent variable measures due to the correlation between the groups *on some* variable *related to the dependent variable.* This correlation and its concomitant variance can be introduced into the measures-and the design-in three ways:

(1) use the same units, for example, participants, in each of the experimental groups,

(2) match units on one or more independent variables that are related to the dependent variable, and

(3) use more than one group of units, like classes or schools, in the design.

Despite the seeming differences among these three ways of introducing correlation into the dependent variable measures, they are basically the same. We now examine the design implications of this basic principle and discuss ways of implementing the principle.

THE GENERAL PARADIGM

With the exception of correlated factorial designs and so-called nested designs, all analysis of variance paradigms of correlated-groups designs can be easily outlined. The word "group" should be taken to indicate set of scores. Then there is no confusion when repeated trials experiment is classified as a multigroup design. The general paradigm is given in Figure 10.2. To emphasize the sources of variance, means of columns and rows have been indicated. The individual dependent variable measures (Y's) have also been inserted.

It is useful to know the system of subscripts to symbols used in mathematics and statistics. A rectangular table of numbers is called a matrix. The entries of a matrix are letters and/or numbers. When letters are used, it is common to identify any particular matrix entry with two (sometimes more) subscripts. The first of these indicates the number of the row, the second the number of the column. Y_{32}, for instance, indicates the Y measure in the third row and the second column. Y_{52} indicates the Y measure of the fifth row and the second column. It is also customary to generalize this system by adding the letter subscripts. In this book, ***i*** symbolizes any row number and ***j*** any column number. *Any* number of the matrix is represented by Y_{ij}. *Any* number of the third row is Y_{3j} and any number of the second column is Y_{i2}

It can be seen that there are two sources of systematic variance: that due to columns, or treatments, and that due to rows-individual or unit differences. Analysis of variance must be the two-way variety.

	Treatments							
Units	X_1	X_2	X_3	.	.	.	X_k	Row
1	Y_{11}	Y_{12}	Y_{13}	.	.	.	Y_{1k}	M_1
2	Y_{21}	Y_{22}	Y_{23}	.	.	.	Y_{2k}	M_2
	Y_{31}	Y_{32}	Y_{33}	.	.	.	Y_{3k}	M_3
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
n	Y_{n1}	Y_{n2}	Y_{n3}	.	.	.	Y_{nk}	M_n
	M_{x1}	M_{x2}	M_{x3}	.	.	.	M_{xk}	(M_t)

Figure 10.2

The reader who has studied the correlation-variance argument of Chapter 15, where the statistics and some of the problems of correlated-groups designs were presented, will have no difficulty with the variance reasoning of Figure 10.2. The intent of the design is to maximize the between-treatments variance, identify the between-units variance, and the error (residual) variance. The maxmincon principle applies here as elsewhere. The only difference, really, between designs of correlated groups and randomized subjects is the rows or units variance.

Units

The units used do not alter the variance principle. The word "unit" is deliberately used to emphasize that units can be persons or participants, classes, schools, districts, cities, even nations. In other words, "unit" is a generalized rubric that can stand for many kinds of entities. The important consideration is whether the units, whatever they are, differ from each other. If they do, *variance between* units is introduced. In this sense,

talking about correlated groups or participants is the same as talking about variance between groups or participants. The notion of individual differences is extended to unit *differences.*

The real value of correlated-groups design beyond allowing the investigator to isolate and estimate the variance due correlation is in guiding the investigator to design research to capitalize on the differences that frequently exist between units. If a research study involves different classes in the same school, these classes are a possible source of variance. Thus it may be wise to use "classes" as units in the design. The well-known differences between schools are very important sources of variance in behavioral research. They may be handled as a factorial design, or they may be handled in the manner of the designs in this chapter. Indeed, if one looks carefully at a factorial design with two independent variables, one of them *schools,* and at a correlated groups design with units *schools,* one finds, in essence, the same design. Study Figure 10.3. On the left is a factorial design and on the right a correlated-groups design. But they look the same! They are the same, in variance principle. (The only differences might be numbers of scores in the cells and statistical treatment.)

		Treatments	
Schools		A_1	A_2
1	B_1		
2	B_2		
3	B_3		

Factorial Designs

		Treatments	
Schools		A_1	A_2
1	B_1		
2	B_2		
3	B_3		

Figure 10.3

One group Repeated Trials Design

In the one-group repeated trials design, as the name indicates, one group is given different treatments at different times. In a learning experiment, the same group of participants may be given several tasks of different complexity, or the experimental manipulation may be to present learning principles in different orders, say from simple to complex, from complex to simple, from whole to part, from part to whole.

It was said earlier that the best possible matching of participants is to match participant with oneself, so to speak. The difficulties in using this solution of the control problem have also be n mentioned. One of these difficulties resembles pretest sensitization, which may produce an interaction between the pretest and the experimentally manipulated variable. Another is that participants mature and learn over time. A participant who has experienced one or two trials of an experimental manipulation and is facing a third trial is now a different person from the one who faced trial one. Experimental situations differ a great deal, of course. In some situations, repeated trials may not unduly affect the performances of participants on later trials; in other situations, they may. The problem of how individuals learn or become unduly sensitized during an experiment is a difficult one to solve. In short, *history,* maturation, and *sensitization* are possible weaknesses of repeated trials. The regression effect can also be a weakness because, as we saw in an earlier chapter, low scorers tend to get higher scores and high scorers lower scores on retesting simply due to the imperfect correlation between the groups. A control group is, of course, needed.

Despite the basic time difficulties, there may be occasions when a one-group repeated trials design is useful. Certainly in analyses of "time" data this is the implicit design. If we have a series of growth measurements of children, for instance, the different times at which the measurements were made correspond to treatments. The paradigm of the design is the same as that of Figure 102. Simply substitute "participants" for "units" and label $X_1, X_2, \ldots$ as "trials "

From this general paradigm special cases can be derived. The simplest case is the one-group, before-after design, Design 8.2 (a), where one group of participants was given an experimental treatment preceded by a pretest and followed by a posttest. Since the weaknesses of this design have already been mentioned, further discussion is not necessary. It should be noted, though, that this design, especially in its nonexperimental form,

closely approximates much commonsense observation and thinking. A person may observe educational practices today and decide that they are not good. In order to make this judgment, one implicitly or explicitly compares today's educational practices with educational practices of the past. From a number of possible causes, depending on the particular bias, the researcher will select one or more reasons for what one believes to be the sorry state of educational affairs: "progressive education," "educationists," "moral degeneration," "lack of firm religious principles," and so on.

Two group, Experimental Group-Control Group Designs

This design has two forms, the better of which (repeated here) was described in Chapter 9 as Design 9.2:

$$[M_r] \quad \begin{array}{lll} X & Y & \text{(Experimental)} \\ \hline \sim X & Y & \text{(Control)} \end{array}$$

In this design, participants are first matched and then assigned to experimental and control groups at random. In the other form, participants are matched but not assigned to experimental and control groups at random. The latter design can be indicated by simply dropping the subscript r from M_r (described in Chapter 8 as Design 8.4, one of the less adequate designs).

The design-statistical paradigm of this war-horse of designs is shown in Figure 10.4. The insertion of the symbols for the means shows the two sources of systematic variance: *treatments* and *pairs, columns* and rows. This is in clear contrast to the randomized designs in an earlier section of this chapter, where the only systematic variance was *treatments* or columns.

The most common variant of the two-group, experimental group-control group design is the before-after, two-group design. [See Design 9.3 (b).] The design-statistical paradigm and its rationale are discussed later.

	Treatments		
Pairs	X_e	X_c	
1	Y_{1e}	Y_{1c}	M_1
2	Y_{2e}	Y_{2c}	M_2
3	Y_{3e}	Y_{3c}	M_3
.	.	.	.
.	.	.	.
n	Y_{ne}	Y_{nc}	M_n
	M_e	M_c	

Figure 10.4

RESEARCH EXAMPLES OF CORRELATED-GROUPS DESIGNS

Hundreds of studies of the correlated-groups kind have been published. The most frequent designs have used matched participants or the same participants with pre- and posttests. Correlated-groups designs, however, are not limited to two groups; the same participants, for example, can be given more than two experimental treatments. The studies described below have been chosen not only because they illustrate correlated-groups design, matching, and control problems, but also because they are historically, psychologically, or educationally important.

Thorndike's Transfer of Training Study

In 1924, E. L. Thorndike published a remarkable study of the presumed effect on intelligence of certain school participants. Students were matched according to scores on Form A of the measure of the dependent variable, intelligence. This test also served as a pretest. The independent variable was One Year's Study of Participants, such as history, mathematics, and Latin. A posttest, Form B of the intelligence test, was given at the end of the year. Thorndike (1924) used an ingenious device to separate the differential effect of each school subject. He did this by matching on Form A of the intelligence test those pupils who studied, for instance, English, history, geometry, and *Latin* with those pupils who studied English, history, geometry, and *shopwork.* Thus, for these two groups, he was comparing the differential effects *of Latin* and *shopwork.* Gains in final intelligence scores were considered a joint effect of growth plus the academic subjects studied.

Despite its weaknesses, this was a colossal study. Thorndike was aware of the lack of adequate controls, as revealed in the following passage on the effects of selection:

> The chief reason why good thinkers seem superficially to have been made such by having taken certain school studies, is that good thinkers have taken such studies. . . . When the good thinkers studied Greek and Latin, these studies seemed to make good thinkers. Now that the good thinkers study Physics and Trigonometry, these seem to make good thinkers. If the abler pupils should all study Physical Education and Dramatic Art these subjects would seem to make good thinkers. (p. 98)

Thorndike pointed the way to controlled educational research, which has led to the decrease of metaphysical and dogmatic explanations in education. His work struck a blow against the razor-strop theory of mental training, the theory that likened the mind to a razor that could be sharpened by stropping it on "hard" subjects.

It is not easy to evaluate a study such as this, the scope and ingenuity of which is impressive. One wonders, however, about the adequacy of the dependent variable, "intelligence" or "intellectual ability." Can school subjects studied for one year have much effect on intelligence? Moreover, the study was not experimental. Thorndike measured the intelligence of students and let the independent variables, school subjects, operate. No randomization, of course, was possible. As mentioned above, he was aware of this control weakness in his study, which is still a classic that deserves respect and careful study despite its weaknesses in history and selection (maturation was controlled).

Miller and DiCara: Learning of Autonomic Functions

In a previous chapter we presented data from one of the set of remarkable studies of the learning of autonomic functioning done by Miller and his colleagues.(Miller, 1971; Miller & DiCara, 1968). Experts and non-experts believe that it is not possible to learn and control responses of the autonomic nervous system. That is, glandular and visceral responses-heart beat, urine secretion, arid blood pressure, for example-were supposed to be beyond the "control" of the individual. Miller believed otherwise. He demonstrated experimentally that such responses are subject to instrumental learning. The crucial part of his method consisted of rewarding visceral responses when they occurred. In the study whose data were cited in an earlier chapter of this book for example, rats were rewarded when they increased or decreased the secretion of urine. Fourteen rats were assigned at random to two groups called "Increase Rats" and "Decrease Rats." The rats of the former group were rewarded with brain stimulation (which was shown to be effective for *increases in* urine secretion), while the rats of the latter group were rewarded for *decreases in* urine secretion during a "training" period of 220 trials in approximately three hours.

To show part of the experimental and analytic paradigms of this experiment, the data before and after the training periods for the Increase Rats and the Decrease Rats are given in Table 10.9 (extracted from the Miller & DiCara's Table 1). The measures in the table are the milliliters of urine secretion per minute per 100 grams of body weight. Note that they are very small quantities. The research design is a variant of **Design 9.3 (a)**:

$$[R] \quad \begin{array}{llll} Y_b & X & Y_a & \text{(Experimental)} \\ \hline Y_b & \sim X & Y_a & \text{(Control)} \end{array}$$

The difference is that ~X, which in the design means absence of experimental treatment for the control group, now means reward for decrease of urine secretion. The usual analysis of the after-training measures of the two groups is therefore altered.

We can better understand the analysis if we analyze the data of Table 10.9 somewhat differently than Miller and DiCara did. (They used *t* tests.) We did a two-way (repeated measures) analysis of variance of the Increase Rats data, Before and After, and the Decrease Rats data, Before and After. The Increase Before and After means were .017 and .028, and the Decrease means were .020 and .006. The Increase F ratio was 43.875 (df =1,6); the Decrease Rats F was 46.624. Both were highly significant. The two Before means of .017 and .020 were not significantly different, however, In this case, comparison of the means of the two After groups, the usual comparison with this design, is probably not appropriate because one was for increase and the other

for decrease in urine secretion.

This whole study, with its highly controlled experimental manipulations and its "control" analyses, is an example of imaginative conceptualization and disciplined competent analysis. The above analysis is one example. But the authors did much more. For example, to be more sure that the reinforcement affected only urine secretion, they compared the Before and After heart rates (beats per minute) of both the Increase and the Decrease rats. The means were 367 and 412 for Increase rats, and 373 and 390 for the Decrease rats. Neither difference was statistically significant. Similar comparisons of blood pressure and other bodily functions were not significant.

Table 10.9 Secretion of Urine Data, Miller and DiCara Study: Increase Rats and Decrease Rats, Before and After Training

Increase Rats[a]

Rats	Before	After	Σ
1	.023	.030	.053
2	.014	.019	.033
3	.016	.029	.045
4	.018	.030	.048
5	.007	.016	.023
6	.026	.044	.070
7	.012	.026	.038
Means	.017	.028	

Decrease Rats[b]

Rats	Before	After	Σ
1	.018	.007	.025
2	.015	.003	.018
3	.012	.005	.017
4	.015	.006	.021
5	.030	.009	.039
6	.027	.008	.035
7	.020	.003	.023
	.020	.006	.023

[a] Increase Before-After: $F = 43.875$ ($p < .001$); $\omega^2 = .357$. The measures in the table are milliliters per minute per 100 grams of weight.

[b] Decrease, Before-After: $F = 46.624$ ($p < .001$); $\omega^2 = .663$.

Students will do well to study this fine example of laboratory research until they clearly understand what was done and why. It will help students learn more about controlled experiments, research design, and statistical analysis than most textbook exercises. It is a splendid achievement!

Tipper, Eissenberg and Weaver: Effects of Practice on Selective Attention.

When speaking about selective attention, one may recall the classic study by Stroop (1935). Stroop had demonstrated the role of interference on selective attention. Irrelevant stimulus can compete with the target stimulus for control of perceptual action. For those unfamiliar with this study, one memorable part of the Stroop study was presenting participants with words such as "green" and "blue" that was printed in red or yellow. Participants were asked to name the colors in which the word was written but would instead read the words. People find it very difficult to suppress the habit of reading words even when one is asked not to. In order to do the task correctly, the participant must slow down and consciously suppress reading the words. This interference was called the Stroop effect. A large number of studies have been performed on selective attention since Stroop's famous study. Tipper, Eissenberg and Weaver (1992) is one of them. This study is different in that they take issue with a number of studies that have been performed on selective attention. For one, Tipper, Eissenberg and Weaver hypothesizes that in any selective attention experiment that uses a participant for one hour or so may be tapping into a different perceptual mechanism than those used in every day life. Laboratory experiments usually require the participant to be present for about an hour. Within an hour the entire experimental experience is still novel. It may be that attentional selectivity is achieved by a different mechanism as the familiarity with the stimuli increases.

Tipper, Eissenberg and Weaver (1992) designed a study to test their hypotheses concerning selective attention using a completely within subjects design. All of the participants experienced all of the treatment conditions. They looked at the effect of interference on reaction time and errors. They had each participant experience both levels of interference: negative priming

and response inhibition across 11 blocks or trials taken over 4 days (practice effect). Their results showed that there was an interference effect ($F = 35.15$, $p < .001$) when using reaction time as the dependent variable. The reaction times were longer when the distraction was present. They also found a practice effect (blocks) {$F = 9.62$, $p < .0001$) and no interaction effect. The practice effect indicated that the reaction time of participants became faster with the increase in practice. The fact that the interaction effect was not significant indicates that the interfering effects of the irrelevant stimulus remained constant even after extended practice. The findings of Tipper, Eissenberg and Weaver (1992) does suggest that other mechanisms for selection attention exist and operate with different levels of experience.

MULTIGROUP CORRELATED-GROUPS DESIGNS

Units Variance

While it is difficult to match three and four sets of participants, and while it is ordinarily not feasible or desirable in behavioral research to use the same participants in each of the groups, there are natural situations in which correlated groups exist. These situations are particularly important in educational research. Until recently, the variances due to differences between classes, schools, school systems, and other "natural" units have not been well controlled or often used in the analysis of data. Perhaps the first indication of the importance of this kind of variance was given in Lindquist's (1940) fine book on statistical analysis in educational research. In this book, Lindquist placed considerable emphasis on schools variance. Schools, classes, and other educational units tend to differ significantly in achievement, intelligence, aptitudes, and other variables. The educational investigator has to be alert to these unit *differences,* as well as to individual differences.

Consider an obvious example. Suppose an investigator chooses a sample of five schools for their variety and homogeneity. The goal of course is external validity: representativeness. The investigator uses pupils from all five schools and combines the measures from the five schools to test the mean differences in some dependent variable. In so doing, the investigator is ignoring the variance due to the differences among schools. It is understandable that the means do not differ significantly; the schools variance is mixed in with the error variance.

Gross errors can arise from ignoring the variance of units such as schools and classes. One such error is to select a number of schools and to designate certain schools as experimental schools and others as control schools. Here the between-schools variance gets entangled with the variance of the experimental variable. Similarly, classes, school districts, and other educational units differ and thus engender variance. The variances must be identified and controlled, whether it be by experimental or statistical control, or both.

FACTORIAL CORRELATED GROUPS

Factorial models can be combined with the units notion to yield a valuable design: *factorial correlated groups* design. Such a design is appropriate when units are a natural part of a research situation. For instance, the research may require the comparison of a variable before and after an experimental intervention, or before and after an important event. Obviously there will be correlation between the before and after dependent variable measures. Another useful example is shown in Figure 10.5. This is a 3×2 factorial design with five units (classes, school, and so forth) in each level, B_1 and B_2.

The strengths and weaknesses of the factorial correlated-groups design are similar to those of the more complex factorial designs. The main strengths are the ability to isolate and measure variances and to test interactions. Note that the two main sources of variance, *treatments* (A) and *Levels (B),* and the *units* variance can be evaluated; that is, the differences between the A, B, and *units* means can be tested for significance. In addition three interactions can be tested: *treatments by levels, treatments by units,* and *levels by units.* If individual scores are used in the cells instead of means, the triple interaction, too, can be tested. Note how important such interaction can be, both theoretically and practically. For example, questions like the following can be answered: Do treatments work differently in different units? Do certain methods work differently at different intelligence levels or with different sexes or with children of different socioeconomic levels? The advanced student will want to know how to handle units (schools, classes, etc.) and units variance in factorial designs. Detailed guidance is given in Edwards (1984) and in Kirk (1995). The subject is difficult. Even the names of the designs become complex: randomized

blocks, nested treatments, split-plot designs. Such designs are powerful however: they combine virtues of factorial designs and correlated groups designs. When needed Edwards and Kirk are good guides. It is suggested, in addition, that help be solicited from someone who understands both statistics and behavioral research. It is unwise to use computer programs because their names seem appropriate. It is also unwise to seek analytic help from computer personnel. One cannot expect such people to know and understand, say, factorial analysis of variance. That is not their job. More will be said about computer analysis in later chapters.

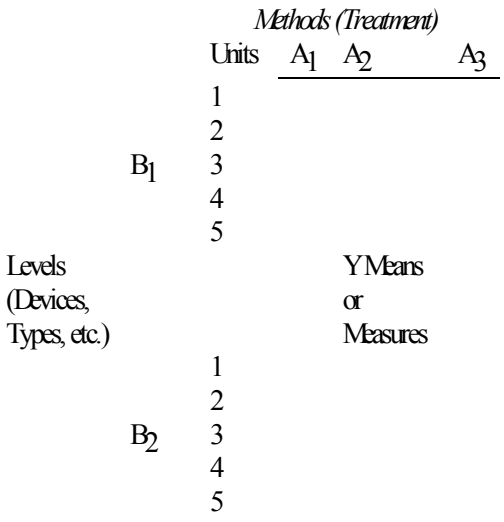


Figure 10.5

Suedfeld and Rank: Revolutionary Leaders and Conceptual Complexity

Suedfeld and Rank (1976), in a study mentioned earlier in another context, tested the intriguing notion that successful revolutionary leaders-Lenin, Cromwell, Jefferson, for example are conceptually simple in their public communications *before* revolution and conceptually complex *after* revolution. Unsuccessful revolutionary leaders, on the other hand, do not differ in conceptual complexity before and after revolution. The problem lends itself to a factorial design and to repeated measures analysis. The design and the data on conceptual complexity are shown in Table 10.10. It can be seen that the successful leaders became conceptually more complex-from 1.67 to 3.65-but unsuccessful leaders did not change very much–2.37 and 2.21. The interaction F ratio was 12.37, significant at the .005 level. The hypothesis was supported.

A few points should be picked up. One, note the effective combining of factorial design and repeated measures. When appropriate, as in this case, the combination is highly useful mainly because it sets aside, so to speak, the variance in the dependent variable measures due to individual (or group or block) differences. The error term is thus smaller and better able to assess the statistical significance of mean differences. Two, this study was nonexperimental: no experimental variable was manipulated.

Table 10.10 Factorial Design with Repeated Measures: Suedfeld & Rank Study of Revolutionary leaders[a]

	Pre takeover	Post takeover	
Success	1.67	3.65	2.66
Failure	2.37	2,22	2.30
	1.96	3.05	

[a]Tabled measures are means of conceptual complexity measures. Interaction F =12.37 ($p < .005$).

Three and most important, the intrinsic interest and significance of the research problem and its theory and the ingenuity of measuring and using conceptual complexity as a variable to "explain" the success of revolutionary leaders overshadow possible questionable methodological points. The above sentence, for instance, may be incongruent with the use of variables in this study. Suedfeld and Rank analyzed measures of the independent variable, conceptual complexity. But the hypothesis under study was actually: If conceptual complexity (after revolution), then successful leadership. But with a research problem of such compelling interest and a variable of such importance (conceptual complexity) imaginatively and competently measured, who wants to quibble?

Perrine, Lisle and Tucker: Offer of Help and Willingness to Seek Support

Teachers at all levels of education use a course syllabus to introduce the course to students. How much and what features in the syllabus have the greatest impact on

students even before classroom instruction starts? Perrine, Lisle and Tucker (1995) developed a study to see if the offer of help on an instructor's syllabus encourages college students of different ages to seek help from their instructors. According to Perrine, Lisle and Tucker, to their best knowledge this is the first study to explore the use of social support by college and university instructors to benefit students. Perrine, Lisle and Tucker also studied the effect of class size on the student's willingness to seek help. The study used 104 undergraduate students of whom 82 were female and 22 were males. Each participant was asked to read a description of two psychology classes. The descriptions include statements made by the instructor of each class on the course syllabus. In the description, class size was manipulated. It was set either to 15, 45 or 150 students. The course was described as demanding with a lot of work, but was enjoyable. It also encouraged the student not to fall behind in the readings and assignments. The two separate statements from the instructors consisted of one that was supportive and one that was neutral. In the supportive statement, the student was encouraged to approach the instructor for help if the student ever encountered problems in the class. The neutral one did not include such a statement. Each participant read both descriptions. After reading the descriptions, the participant responded to questions about their willingness to seek help from the instructor for six possible academic problems encountered in the class. The six problems were (1) trouble understanding textbook, (2) low grade on first exam, (3) hard to hear instructor's lectures, (4) study skills ineffective for course, (5) thinking of dropping the course and (6) trouble understanding major topic. The participant used a 6 point rating scale: 0 = definitely no to 6 = definitely yes.

The design was a 3 × 2 × 2 (class size × syllabus statement × student age) factorial design. The design contained one manipulated (active) independent variable, one measured (attribute) independent variable and one within-subjects (correlated) independent variable. Class size was the randomized and manipulated dependent variable. Student age was the measured independent variable and syllabus statement was the correlated independent variable. Using the appropriate analysis of variance (usually referred to as mixed anova when at least one independent variable is between subjects and at least one other is within-subjects) participants expressed significantly more willingness to seek help from the instructor when the supportive statement appeared on the course syllabus than when only the neutral statement appeared.

Table 10.1l. Means and F-Values for Syllabus Statement Differences and Age Differences. Perrine, Lisle and Tucker Study.

Academic Problem	Syllabus		
	Supportive	Neutral	F
Trouble understanding textbook	4.7	3.7	76.08**
Low grade on first exam	4.8	4.0	49.89**
Hard to hear instructor's lectures	4.4	3.8	36.05**
Study skills ineffective for course	4.7	3.6	79.57**
Thinking about dropping the course	4.9	3.8	61.80**
Trouble understanding major topic	5.3	4.2	82.97**

Academic Problem	Age		
	Older	Younger	F
Trouble understanding textbook	4.8	4.1	5.48*
Low grade on first exam	5.2	4.3	7.64*
Hard to hear instructor's lectures	4.4	4.0	1.01
Study skills ineffective for course	4.8	4.0	6.32*
Thinking about dropping the course	4.8	4.3	2.18
Trouble understanding major topic	5.3	4.6	7.69*

*p<.05, **p<.01

Younger students under the age of 25) expressed less willingness than older students. There was also an age × syllabus interaction ($F = 4.85$, $p < .05$) that was significant. The response to the offer of help was different between age groups. The statements affected younger students less than older students. Class size did not appear to be a significant factor on whether students were willing to seek help or not. Table 10.11 presents the summary statistics for the study.

ANALYSIS OF COVARIANCE

The invention of the analysis of covariance by Ronald Fisher was an important event in behavioral research methodology. Here is a creative use of the variance principles common to experimental design and to correlation and regression theory-which we study later in the book-to help solve a long-standing control problem.

Analysis of covariance is a form of analysis of variance that tests the significance of the differences among means of experimental groups after taking into account initial differences among the groups and the correlation of the initial measures and the dependent variable measures. That is, analysis of covariance analyzes the differences between experimental groups on Y, the dependent variable, after taking into account either initial differences between the groups on Y (pretest), or differences between the groups in some potential independent variable or variables, X, substantially correlated with Y, the dependent variable. The measure used as a control variable-the pretest or pertinent variable is called a covariate.

The reader should be cautious when using the analysis of covariance. It is particularly sensitive to violations of its assumptions. The potential misuse of this method was of such concern that the journal of *Biometrics* in 1957 devoted an entire issue to it. Elashoff (1969) wrote an important article for educational researchers on the use of this method. The consensus is that it is generally not a good idea to use this method for nonexperimental research designs.

Clark and Walberg: *Massive Reinforcement and Reading Achievement*

There is little point to describing the statistical procedures and calculations of analysis of covariance. First, in their conventional form, they are complex and hard to follow. Second, we wish here only to convey the meaning and purpose of the approach. Third and most important, there is a much easier way to do what analysis of covariance does. Later in the book we will see that analysis of covariance is a special case of multiple regression and is much easier to do with multiple regression. To give the reader a feeling for what analysis of covariance accomplishes, let us look at an effective use of the procedure in educational and psychological studies

Clark and Walberg (1968) thought that their participants, potential school dropouts doing poorly in school, needed far more reinforcement (encouragement, reward, etc.) than participants doing well in school. So they used massive reinforcement with their experimental group participants and moderate reinforcement with their control group participants. Since their dependent variable, reading achievement, is substantially correlated with intelligence, they also needed to control intelligence. A one-way analysis of variance of the reading achievement means of the experimental and control groups yielded an F of 9.52, significant at the .01 level, supporting their belief. It is conceivable, however, that the difference between the experimental and control groups was due to intelligence rather than to reinforcement. That is, even though the S's were assigned at random to the experimental groups, an initial difference in intelligence in favor of the experimental group may have been enough to make the experimental group reading mean significantly greater than the control group reading mean, since intelligence is substantially correlated with reading. With random assignment, it is unlikely to happen, but it can happen. To control this possibility, Clark and Walberg used analysis of covariance.

Table 10.12 Analysis of Covariance Paradigm, Clark and Walberg Study

	Experimental (Massive Reinforcement)	
	X (Intelligence)	Y (Reading)
Means	92.05	31.62

	Control (Moderate Reinforcement)	
	X (Intelligence)	Y (Reading)
Means	90.73	26.86

Study Table 10.4, which shows in outline the design and analysis. The means of the X and Y scores, as

reported by Clark and Walberg, are given at the bottom of the table. The Y means are the main concern. They were significantly different. Although it is doubtful that the analysis of covariance will change this result, it is possible that the difference between the X means, 92.05 and 90.73, may have tipped the statistical scales, in the test of the difference between the Y means, in favor of the experimental group. The analysis of covariance F test, which uses Y sums of squares and mean squares purged of the influence of X, was significant at the .01 level: F = 7.90. Thus the mean reading scores of the experimental and control groups differed significantly, after adjusting or controlling for intelligence.

Taris: Locus of Control, Situations and Driving Behavior.

This study by Taris (1997) uses a rather unusual design. This is a 2 × 2 × 2 × 2 factorial with two covariates. All four of the independent variables in this study were within-subject factors. Taris' study examines the extent situational and personality factors influences driving behavior of young adults. The study's four within factors are (1) desirability {2) verifiability, (3} scenario and (4) controllability. The covariates are locus of control and annual mileage. Annual mileage was used as a measure of driving experience. Locus of control is a personality measure. Each participant received descriptions of four situations in which Controllability and Verifiability were systematically manipulated. High controllability was generally in the form of a statement where the participant was at ease, "with plenty of time." Low controllability meant a situational statement where the participant is in a hurry to get somewhere. High verifiability generally meant someone was watching, i.e. "an area heavily patrolled by police," or "passenger in the car." In each situation there are two scenarios. One scenario asked for a decision on the speed at which one was driving and the other involved a decision to either stop or not stop for a red traffic light. In each scenario there is a desirable and an undesirable course of action. Taris presented the situations, scenarios and actions in random order.

An example of a low controllability, high verifiability condition created by Taris is

> "You are driving home after work. The traffic is rather dense. You are aware that the police often patrol the area. The meeting at your job took more time than you expected. Now you will be late for the home match of your volleyball team in which you are drafted. Your team counts on you." (p. 990).

An example of high controllability and low verifiability would read:

> "You are driving home after work. The traffic is rather dense. You are aware that the police seldom patrol the area. You left your job at the usual time. Tonight you will play a home match with your volleyball team in which you are drafted. Your team counts on you " (p. 990}.

As mentioned previously, there are two scenarios. One scenario deals with obeying speed laws and the other deals with stopping at a red light. For each scenario there is a desirable or undesirable action. From Taris' example, this would be a statement such as "You drive on. At times you exceed the speed limit," versus " You drive on. Nowhere do you exceed the speed limit."

The dependent variable is a probability judgment regarding the likelihood that one would engage in that particular action. Results showed a Desirability Effect ($F = 532.21$, $p < .001$). Desirable actions were considered more likely than undesirable ones. This shows that youthful drivers are not likely to act in an undesirable way. The significant interaction between Desirability and Verifiability ($F = 4.22$, $p < .05$) states that undesirable actions were more likely if verifiability was low. The interaction between Controllability and Desirability was not significant. So Taris was unable to state that undesirable behavior would be more likely to occur when controllability was low. The interaction between Verifiability and Locus of Control ($F = 8.42$, $p < .01$) was significant. This tells us that people with an internal locus of control found verifiability less important while people that are externally controlled found verifiability very important. Taris also found a significant three-way interaction effect ($F = 4.37$, $p < .05$) between Desirability, Verifiability and Locus of Control. Upon examining the data, Taris found no differences for desirable actions. However there was a effect for undesirable actions. Those who drove a large number of miles were more likely to engage in undesirable acts. Taris concludes from these results that both locus of control and situational factors are

important determinants of the choice to act desirably or undesirably. Although most people know that the greater presence of police deter undesirable behavior, this study has an important finding in that not all youthful drivers act in the same way. Taris has found that internally oriented youths are less likely to have their behavior influenced by situational factors. However, externally oriented youth are more affected by verifiability. Table 10.13 gives the summary statistics for this study.

Table 10.13 Mean Ratings for the Likelihood of Action as a Function of Desirability, Verifiability and Locus of Control. Taris Study.

	External		Internal	
Verifiability	Desirable	Undesirable	Desirable	Undesirable
Low	7.61	3.76	7.56	3.37
High	7.62	3.20	7.58	2.99

RESEARCH DESIGN AND ANALYSIS: CONCLUDING REMARKS

Four major objectives have dominated the organization and preparation of Part Six. The first was to acquaint the student with the principal designs of research. By so doing, it was hoped that narrowly circumscribed notions of doing research with, say, only one experimental group and one control group, or with matched participants or with one group, before and after, may be widened. The second objective was to convey a sense of the balanced structure of good research designs, to develop sensitive feeling for the architecture of design. Design must be formally as well as functional (y fitted to the research problems we seek to solve. The third objective was to help the reader understand the logic of experimental inquiry and the logic of the various designs. Research designs are alternative routes to the same destination: reliable and valid statements of the relations among variables. Some designs, if practicable, yield stronger relational statements then other designs.

In a certain sense, the fourth objective of Part Six: to help the student understand the relation between the research design and statistics has been the most difficult to achieve. Statistics is, in one sense, the technical discipline of handling variance. And, as we have seen, one of the basic purposes of design is to provide control of systematic and error variances. This is the reason for treating statistics in such detail in Parts Four and Five before considering design in Part Six. Fisher (1951) expresses this idea succinctly when he says, "Statistical procedure and experimental design are only two different aspects of the same whole, and that whole comprises all the logical requirements of the complete process of adding to natural knowledge by experimentation." p.3

A well-conceived design is no guarantee of the validity of research findings. Elegant designs nicely tailored to research problems can still result in wrong or distorted conclusions. Nevertheless, the chances of arriving at accurate and valid conclusions are better with sound designs than with unsound ones. This is relatively sure: if design is faulty, one can come to no clear conclusions. If, for instance, one uses a two-group, matched-subjects design when the research problem logically demands a factorial design, or if one uses a factorial design when the nature of the research situation calls for a correlated-groups design, no amount of interpretative or statistical manipulation can increase confidence in the conclusions of such research.

It is fitting that. Fisher (1951) should have the last word on this subject. In the first chapter of his book, *The Design of Experiments,* he said:

> If the design of an experiment is faulty, any method of interpretation that makes it out to be decisive must be faulty too. It is true that there are a great many experimental procedures which are well designed that may lead to decisive conclusions? However, on other occasions they may fail to do so. In such cases, if decisive conclusions are in fact drawn when they are unjustified, we may say that the fault is wholly in the interpretation, not in the design. But the fault of interpretation. . . lies in overlooking the characteristic features of the design which lead to the result being sometimes inconclusive, or conclusive on some questions but not on all. To understand correctly the one aspect of the problem is to understand the other. (p. 3)

Study Suggestions

Randomized Groups

1. In studying research design, it is useful to do analyses of variance-as many as possible: simple one-way analyses and two-variable factorial analyses. Try even a three-variable analysis. By means of this statistical work you can get a better understanding of the designs. You may well attach variable names to your "data," rather than work with numbers alone. Some useful suggestions for projects with random numbers follow.

(a} Draw three groups of random numbers 0 through 9. Name the independent and dependent variables. Express a hypothesis and translate it into design-statistical language. Do a one-way analysis of variance. Interpret.

(b) Repeat 1 (a) with five groups of numbers.

(c) Now increase the numbers of one of your groups by 2, and decrease those of another group by 2. Repeat the statistical analysis.

(d) Draw four groups of random numbers, 10 in each group. Set them up, at random, in a 2×2 factorial design. Do a factorial analysis of variance.

(e) Bias the numbers of the two right-hand cells by adding 3 to each number. Repeat the analysis. Compare with the results of 1(d).

(f) Bias the numbers of the data of 1 (d}, as follows: add 2 to each of the numbers in the upper left and lower right cells. Repeat the analysis. Interpret.

2. Look up Study Suggestion 2 and 3, Chapter 14. Work through both examples again. (Are they easier for you now?)

3. Suppose that you are the principal of an elementary school. Some of the fourth- and fifth *grade* teachers want to dispense with workbooks. The superintendent does not like the idea, but is willing to let you test the notion that workbooks do not make much difference. (One of the teachers even suggests that workbooks may have bad effects on both teachers and pupils.) Set up two research plans and designs to test the efficacy of the workbooks: a one-way design and a factorial design. Consider the variables achievement, intelligence, and gender. You might also consider the possibility of teacher attitude toward workbooks as an independent variable.

4. Suppose an investigation using methods and gender as the independent variables and achievement as the dependent variable has been done with the results reported in Table 10.14. The numbers in the cells are fictitious means. The F ratios of methods and gender are not significant. The interaction F ratio is significant at the .01 level. Interpret these results statistically and substantively. To do the latter, give names to each of the three methods.

Table 10.14 Hypothetical Data Means} of a Fictitious Factorial Experiment

	Methods			
	A_1	A_2	A_3	
Male	45	45	36	42
Female	35	39	40	38
	40	42	38	

5. Although difficult and sometimes frustrating, there is no substitute for reading and studying original research studies. A number of studies using factorial design and analysis of variance have been cited and summarized in this chapter and in earlier chapters. Select and read two of these studies. Try summarizing one of them. Criticize both studies for adequacy of design and execution of the research (to the best of your present knowledge and ability). Focus particularly on the adequacy of the design to answer the research question or questions.

Correlated Groups

6. Can memory be improved by training? William James, the great American psychologist and philosopher, did a memory experiment on himself over 100 years ago. (See James, 1890) He first learned 158 lines of a Victor Hugo poem, which took him 131 5/6 minutes. This was his baseline. Then he worked for 20-odd minutes daily, for 38 days, learning the entire first book *of Paradise Lost.* (Book 1 is 22 tightly printed pages of rather difficult verse!) This was training of his memory. He returned to the Hugo poem and learned 158 additional lines in 151 1/2 minutes. Thus he took longer after the training than before. Not satisfied, he had others do similar tasks-with similar results. On the basis of this work, what conclusions could James come to? Comment on his research design. What design among

those in this book does his design approximate?

7. In the Miller and DiCara study outlined in this chapter, the authors did parallel analyses. In addition to their analyses of urine secretion, for example, they analyzed heart beat rate and blood pressure. Why did they do this?

8. In her classic study of "natural categories," Rosch (1973) replicated the original study of colors with forms (square, circle, etc.). What advantage is there in such replication?

9. We did a two-way (repeated measure) analysis of variance of the Miller and DiCara Increase Rats data of Table 10.9, with some of the results reported in the table. ω^2 (Hays omega-squared) was .357 ω^2 for the Decrease Rats data was .663. What do these coefficients mean? Why calculate them?

10. Kolb (1965), basing his work on the outstanding work of McClelland on achievement motivation, did a fascinating experiment with underachieving high school boys of high intelligence. Of 57 boys, he assigned 20 at random to a training program in which, through various means; the boys were "taught" achievement motivation (an attempt to build a need to achieve into the boys). The boys were given a pretest of achievement motivation in the summer, and given the test again six months later. The mean change *scores* were, for experimental and control groups, 6.72 and -. 34, respectively. These were significant at the .005 level.

(a) Comment on the use of change scores. Does their use lessen our faith in the statistical significance of the results?

(b) Might factors other than the experimental training have induced the change?

11. Lest the student believe that only continuous measures are analyzed and that analysis of variance alone is used in psychological and educational experiments, read the study by Freedman, Wallington and Bless (1967) on guilt and compliance. There was an experimental group (S's induced to lie) and a control group. The dependent variable was measured by whether a participant did or did not comply with a request for help. The results were reported in crosstabulation frequency tables. Read the study, and, after studying the authors' design and results, design one of the three experiments another way. Bring in another independent variable, for instance. Suppose that it was known that there were wide individual differences in compliance. How can this be controlled? Name and describe two kinds of design to do it.

12. One useful means of control by matching is to use pairs of identical twins. Why is this method a useful means of control? If you were setting up an experiment to test the effect of environment on measured intelligence and you had 20 pairs of identical twins and complete experimental freedom, how would you set up the experiment?

13. In a study in which training on the complexities of art stimuli affected attitude toward music, among other things, Renner (1970 used analysis of covariance, with the covariate being measures from a scale designed to measure attitude toward music. This was a pretest. There were three experimental groups. Sketch the design from this brief description. Why did Renner use the music attitude scale as a pretest? Why did she use analysis of covariance? (Note: The original report is well worth reading. The study, in part a study of creativity, is itself creative.)

14. In a significant study of the effect of liberal arts education on complex concept formation, inter and McClelland (1978) found the difference between senior and freshmen of a liberal arts college on measure of complex concept formation to be statistically significant ($M_S = 2.00$, $M_f = 1.22$; $t = 3.76$ ($p < .001$). Realizing that a comparison was needed, they also tested similar mean differences in a teachers college and in a community college. Neither of these differences was statistically significant. Why did Winter and McClelland test the relation in the teachers college and in the community college? It is suggested that students look up the original report-it is well worth study-and do analysis of variance from the reported n's, means, and standard deviations, using the method outlined in Chapter 13 (Addendum).

15. One virtue of analysis of covariance seldom mentioned in texts is that three estimates of the correlation between X and Y can be calculated. The three are the total r over all the scores, the between-groups r, which is the r between the X and Y means, and the within-groups r, the r calculated from an average of the r's between X and Y within the k groups. The within-groups r is the "best" estimate of

the "true" r between X and Y. Why is this so?

[Hint: Can a total r, the one usually calculated in practice, be inflated or deflated by between-groups variance?]

16. The 2 × 2 × 2 factorial design is used a good deal by social psychologists. Here are two unusual, excellent, even creative studies in which it was used:

Aronson, E. & Gerard E. (1966). Beyond Parkinson's law: The effect of excess time on subsequent performance. *Journal of Personality* and *Social Psychology, 3*, 336- 339.

Carlsmith, J. & Gross, A (1969). Some effects of guilt on compliance. *Journal of Personality and Social Psychology, 11*, 232-239.

The following article of Martin's uses a 2 × 3 × 5 and a 2 × 2 × 5 designs.

Martin, R. {1998). Majority and minority influence using the afterimage paradigm: A series of attempted replications. *Journal of Experimental Social Psychology, 34*, 1-26.

Read one of these studies.

Chapter Summary

1. Randomized subjects designs are the preferred designs of behavioral research.
2. Randomized subjects designs are true experiments with active, manipulated independent variables.
3. The usual statistical method to analyze data from randomized subjects designs is analysis of variance.
4. Randomized subject designs usually require a large number of participants to achieve the desired precision.
5. Correlated subjects designs usually involve
   a)using the same participants in each treatment condition
   b)matching participants on one or more independent variables related to the dependent variable
   c)using more than one group of participants, i.e., classrooms.
6. Units can be different kinds of entities. In psychological research, units are usually people or animals
7. Correlated subjects designs include the one group repeated trials (measures) design.
8. Design 9.2 is the better design to use when participants are matched and randomly assigned to treatment groups.
9. A covariate is a potential independent variable used to adjust for the individual differences between groups that are not due to the treatment. Pretests are the most common covariates.
10. Analysis of covariance is a correlated subjects method of statistical analysis. A covariate adjusts the dependent variable, then the adjusted values are used in an analysis of variance. Multiple regression is another statistical method one can use for this purpose.

Chapter 11

Quasi Experimental and N = 1 Designs of Research

VARIANTS OF BASIC DESIGNS

Designs 9.1 through 9.6 are the basic experimental designs. Some variants of these designs have already been indicated. Additional experimental and control groups can be added as needed, but the core ideas remain the same. It is always wise to consider the possibility of adding experimental and control groups. Within reason, the addition of such groups provides more validating evidence for the study's hypotheses. This design was a combination of two other basic designs. It combined the strengths of both and adding replication power, as well as further controls. Such advantages lead to the principle that, whenever we consider a research design, we should consider the possibility of adding experimental groups as replications or variants of experimental and control groups. One of the major goals of science is to find causal relations. The true experiment ("true here is taken to mean a manipulated independent variable where a causal statement is possible) if arranged and executed correctly can provide the researcher with a causal statement concerning the relation between X and Y. This is generally considered the highest form of experimentation. The weakening of the components of the true experiment is what we will discuss in this chapter.

Compromise Designs a.k.a. Quasi Experimental Designs

It is possible, indeed necessary, to use designs that are compromises with true experimentation. Recall that true experimentation requires at least two groups, one receiving an experimental treatment and one not receiving the treatment or receiving it in different form. The true experiment requires the manipulation of at least one independent variable, the random assignment of participants to groups, and the random assignment of treatments to groups. When one or more of these prerequisites is missing for one reason or another, we have a compromise design. Compromise designs are popularly known as quasi-experimental designs. They are called quasi because quasi means "almost" or "sort of." Cook and Campbell (1979) presented two major classifications of quasi-experimental design. The first is called the "non-equivalent control group designs" and the second is the "interrupted time series designs." A number of research studies performed outside of the laboratory could fall into this category. Many marketing research studies are in the form of quasi-experimental designs. Often a researcher is asked to "design" and analyze the data from a study that was unplanned. For example, a grocery buyer decides to stock a different brand of baby food. Her superiors may later ask if such a move was profitable. This buyer would then consult with a market researcher to determine what can be done to show whether her decision was a profitable one or not. Such analyses would not have the niceties of random selection and assignment and it would consist of data taken over time. Additionally, other ads or the season of the year could influence the baby food sales. The only component resembling a true experiment is the fact that the independent variable was manipulated. Not all stores received the different baby food product. With such problems, the researcher would turn to the use of quasi-experimental or compromise research designs.

Nonequivalent Control Group Design

Perhaps the most commonly used quasi-experimental design is the experimental group-control pattern in which one has no clear assurance that the experimental and control groups are equivalent. Some such as Cook and Campbell (1979), Christensen (1996), Ray (1997) and Graziano & Raulin (1993) refer to it as the nonequivalent control group design. Cook and Campbell present eight variations of this design that they state are "interpretable." The eight are

- no-treatment control group designs
- nonequivalent dependent variables designs
- removed treatment group designs
- repeated treatment designs
- reversed treatment nonequivalent control group design
- cohort designs

- posttest only designs
- regression continuity designs

In this book we will discuss in detail only one of these. This one is the most likely one to occur in the research literature in some shape and form. For a thorough discussion of these eight types of non-equivalent control group designs one should read Cook and Campbell (1979).

No-treatment Control Group Design

The structure of the no-treatment control group design has already been considered in Design 9.3. Cook and Campbell (1979) refers to this design as the untreated control group design with pretest and posttest. The compromise form is as follows:

Design 11.1: No-treatment Control Group Design

Y_b	X	Y_a	(Experimental)
Y_b	~X	Y_a	(Control)

The difference between Designs 9.3 and 11.1 is sharp. In Design 11.1, there is no randomized assignment of participants to groups, as in 9.3(a), and no matching of participants and then random assignment, as in 9.3(b). Design 11.1, therefore, is subject to the weaknesses due to the possible lack of equivalence between the groups in variables other than X. Researchers commonly take pains to establish equivalence by other means, and to the extent they are successful in doing so, to this extent the design is valid. This is done in ways discussed below.

It is often difficult or impossible to equate groups by random selection or random assignment, or by matching. Should one then give up doing the research? By no means. Every effort should be made, first, to select and to assign at random. If both of these are not possible, perhaps matching and random assignment can be accomplished. If they are not, an effort should be made at least to use samples from the same population or to use samples as alike as possible. The experimental treatments should be assigned at random. Then the similarity of the groups should be checked using any information available sex, age, social class, and so on. The equivalence of the groups could be verified using the means and standard deviations of the pretests: t tests and F tests will do. The distributions should also be checked. Although one cannot have the assurance that randomization gives, if these items all check out satisfactorily, one can go ahead with the study knowing at least that there is no known evidence against the equivalence assumption.

These precautions increase the possibilities of attaining internal validity. There are still difficulties, all of which are subordinate to one main difficulty, called selection. (These other difficulties will not be discussed here. For detailed discussion, see the Campbell and Stanley, 1963 or Cook & Campbell, 1979).

Selection is one of the difficult and troublesome problems of behavioral research. Since its aspects will be discussed in detail in Chapter 23 on nonexperimental research, only a brief description will be given here. One of the important reasons for the emphasis on random selection and assignment is to avoid the difficulties of selection. When participants are selected into groups on bases extraneous to the research purposes, we call this "selection," or alternatively, "self-selection." Take a common example: let us assume that volunteers are used in the experimental group and other participants are used as controls. If the volunteers differ in a characteristic related to Y, the dependent variable, the ultimate difference between the experimental and control groups may be due to this characteristic rather than to X, the independent variable. Volunteers, for instance, may be more intelligent (or less intelligent) than nonvolunteers. If we were doing an experiment with some kind of learning as the dependent variable, obviously the volunteers might perform better on Y because of superior intelligence, despite the initial likeness of the two groups on the pretest. Note that, if we had used only volunteers and had assigned them to experimental and control groups at random, the selection difficulty is lessened. External validity or representativeness, however, would be decreased.

Cook and Campbell (1979) claim that even in very extreme cases, it is still possible to draw strong conclusions if all the threats to validity are considered and accounted for. Without the benefit of random assignment, attempts should be made through other means to eliminate rival hypotheses. We consider only the design that uses the pretest because the pretest could provide useful information concerning the effectiveness

of the independent variable on the dependent variable. The pretest could provide data on how equal the groups are to each other prior to administering treatment to the experimental group.

Another more frequent example in educational research is to take some school classes for the experimental group and others for the control group. If a fairly large number of classes are selected and assigned at random to experimental and control groups, there is no great problem. But if they are not assigned at random, certain ones may select themselves into the experimental groups, and these classes may have characteristics that predispose them to have higher mean Y scores than the other classes. For example, their teachers may be more alert, more intelligent, and more aggressive. These characteristics interact with the selection to produce, irrespective of X, higher experimental group than control group Y scores. In other words, something that influences the selection processes, as do the volunteer participants, also influences the dependent variable measures. This happens even though the pretest may show the groups to be the same on the dependent variable. The X manipulation is "effective," but it is not effective in and of itself. It is effective because of selection, or self-selection. Additionally, an educational researcher may have to receive the school district's approval for research. At times, the district will assign the school and the classroom that a research may use.

A classic study by Sanford and Hemphill (1952) reported in Campbell and Stanley (1963) used this design. This study was conducted at the United States Naval Academy at Annapolis. This study was done to see if a psychology course in the curriculum increased the students' (midshipmen) confidence in social situations. The second-year class took the psychology course. The second-year midshipmen were the first group of students to take the psychology course. The comparison or control group was the third-year class. The third year students did not take the course in their second year. A social situation questionnaire was administered to both classes at the beginning of the academic year and at the end of the year. The results showed an increase in confidence scores for the second year class from 43.26 to 51.42. The third-year class also showed an increase. However, their increase was considerably smaller changing from 55.80 to 56.78. One might conclude from these data that taking the psychology course did have an effect of increasing confidence in social situations. However, other explanations are also possible. One could explain that the greater gains made by the second-year class were the result of some maturational development that has its largest growth in the second year with a smaller growth in the third year. If such a process exists, the larger score increase for the second-year class would have occurred even if the midshipmen had not taken the psychology class. The fact that the second year class started with a lower score than the third year class might indicate that these students had not yet reached an equivalent level to the third year class. Plus, the end-of-year scores of the second year class were not equivalent to the beginning scores for the third year class. A better and stronger design would be to create two equivalent groups from the second-year class through random selection and give at random the psychology class to one of them.

Possible outcomes from this design are given in Figure 11.1. There is the possibility of a different interpretation on causality depending on which outcome the researcher obtains. In almost all of the cases the most likely threat to internal validity would be the selection-maturation interaction. You might recall that this interaction occurs when the two groups are different to begin with as measured by the posttest. Then one of the groups experience greater differential changes such as getting more experienced, more accurate, more tired, etc. than the other group. The difference after treatment as observed in the posttest can not exactly be attributed to the treatment itself.

In Figure 11.1(a), there are three possible threats to internal validity. As mentioned above, the most prevalent one is the selection-maturation interaction. With the outcome in Figure 11.1a, Cook and Campbell (1979) states that there are four alternative explanations. The first is selection-maturation interaction. Let's say the study involves comparing two strategies or methods of problem solving. Group A has higher intelligence than Group B. Group A scores higher on the pretest than group B. Group A sees an increase in the posttest scores after treatment. Group B sees little or no change. One might feel that the treatment that Group A receives is superior to the one received by Group B. However, with selection-maturation, interaction, Group A's increase may be due to their higher level of intelligence. With a higher level of intelligence, these participants

can process more or grow faster than Group B. A second explanation is one of instrumentation.

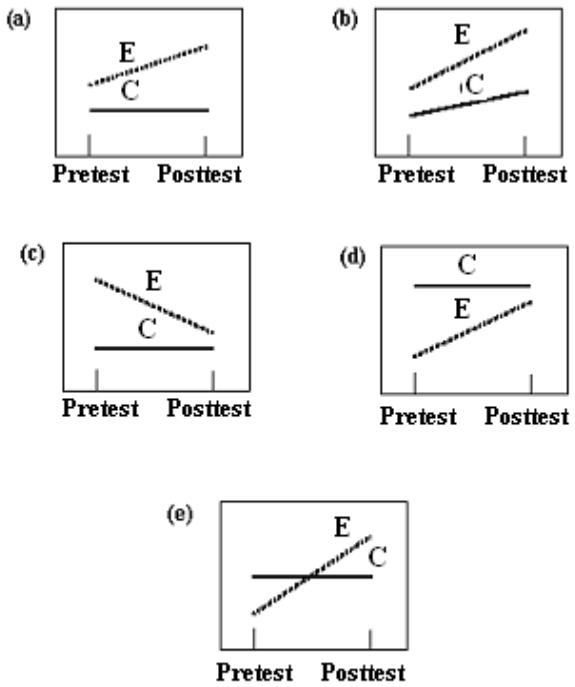


Figure 11.1 Five possible outcomes for the non-equivalent control group design

The scale used to measure the dependent variable may be more sensitive at certain levels than others. Take percentiles for example. Percentiles have an advantage over raw scores in that they convey direct meaning without other pieces of information. However, percentiles are nonlinear transformations of the raw scores. As such changes near the center of the distribution are more sensitive than at the tails. While a change of only 2 or 3 points on the raw score scale can reflect a 10 percentile point change in the center of the distribution. A change of 15 raw score points might be necessary to see a 10 percentile point increase at the tail. Hence, Group B may not change much because the measurements are not sensitive enough to detect the changes. However, Group A will show a change because they happen to be in the more sensitive part of the measurement scale.

The third explanation is statistical regression. Let's say that the two groups, A and B actually comes from different populations and Group B is the group of interest. The researcher wants to introduce an educational plan to help increase the intellectual functioning of these participants. These participants are selected because they generally score low on intelligence tests. The researcher creates a comparison or control group from normal scoring students. This group is depicted as Group A in Figure 11.1a. These students would be at the low end of the test score scale, but not as low as Group B. If this is the setup, then statistical regression is a viable alternative explanation. The increase in scores by Group A would be due to their selection on the basis of extreme scores. On the posttest, their scores would go up because they would be approaching the population baseline. The fourth explanation centers on the interaction between history and selection. Cook and Campbell (1979) refer to this as the local history effect. In this situation, something other than the independent variable will affect one of the groups (Group A) and not the other (Group B). Let's say a market researcher wanted to determine the effectiveness of an ad for soup starters. Sales data are gathered before and after introducing the ad. If two groups are used where they are from different regions of the country, the growth in sales seen by one of the groups (A) may not be necessarily due to the ad. Let us say one group is from southern California and the other is in the Midwestern United States. Both groups may have similar purchasing behavior during the spring and summer, i.e., not a great deal of need for soup starters. However, as the fall season approaches, the sale of soup starters may increase for the group in the Midwest. In southern California, where the temperatures are considerably warmer all year around, the demand for soup starters would remain fairly constant. So here the explanation would be the season of the year and not the ad.

All of the threats mentioned for Figure 11.1a is also true for Figure 11.1b. While in Figure 11.1a one of the groups (Group B) remains constant, in Figure 11.1b, both groups experience an increase from pretest to posttest. Selection-maturation is still a possibility since by definition the groups are growing (or declining) at a different rate where the lower scoring group (Group C) progresses at a lower rate than the high scoring group (Group T). To determine if selection-maturation is playing a main role for the results, Cook and Campbell (1979) recommends two methods. The first involves looking at only the data for the experimental group (Group T). If the within group variance for the posttest is considerably greater than the within-group variance of the pre-test, then there is evidence of a selection-maturation interaction. The second method is to

develop two plots and the regression line associated with each plot. One plot is for the experimental group (Group T). The pretest scores are plotted against the maturational variable. The maturation variable can be age or experience. The second plot would be the same except it would be for the control group (Group C). If the regression line slopes for each plot differ from each other, then there is evidence of a differential average growth rate, meaning that there is the likelihood of a selection-maturation interaction (see Figure 11.2)

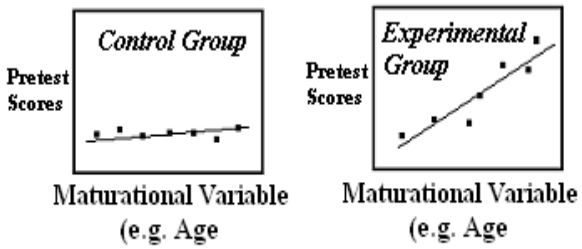


Figure 11.2 Comparison of Experimental & Control Groups

The outcome shown in Figure 11.1c is more commonly found in clinical psychology studies. The treatment is intended to lead to a decline of an undesired behavior. Like the previous two outcomes, this one is also susceptible to selection-maturation interaction, statistical regression, instrumentation and local history effects. In this outcome, the difference between the experimental and control groups are very dramatic on the pretest but after the treatment they are closer to one another. An example where this might happen is in a study where the researcher tests the effectiveness of two different diets on weight loss. The initial weight of Group E is considerably higher than Group C. After 90 days on the diet, Group E shows a greater loss of weight than Group C. This may be attributed to the diet plan used by Group E, if it were not for the fact that Group E was considerably heavier to begin with and may through local history (hearing the dangers of over-weightedness, or eating certain foods on television) lose weight.

The fourth outcome is shown in Figure 11.1d. This is different from the previous three in that the control group (Group C) starts out higher than the experimental group (Group E) and remains higher even at posttest. However for Group E, they showed a great gain from pretest to posttest. Statistical regression would be a threat if the participants in Group E were selected on the basis of their extremely low score. Cook and Campbell (1979) states that the selection-maturation threat can be ruled out since this effect usually results in a slower growth rate for low scores and a faster growth rate for high scorers. Here, the low scorers show the greater growth in scores than the high scorers. This evidence lends support to the effectiveness of the treatment condition received by Group E. What cannot be easily ruled out are the threats from instrumentation and local history that we saw in the previous three outcomes of non-equivalent control group designs.

With the final outcome shown in Figure 11.1e, the means of the experimental (Group E) and control (Group C) groups are significantly different from one another at both pretest and posttest. However, the differences are in the reverse direction in the posttest than in the pretest. The trend lines cross over one another. Group E initially starts low but then overtakes Group C who initially scored high. Cook and Campbell (1979) found this outcome to be more interpretable than the previous four. Instrumentation or scaling is ruled out because no transformation of the scores could remove or reduce this cross-over or interaction effect. Statistical regression becomes untenable because it is extremely rare that a low score can regress enough to overtake an initially high score. Other than a very complicated selection-maturation interaction effect, this pattern is not akin to selection-maturation threats. Maturation for example does not generally start off different, meet and then grows apart in the opposite direction. Hence, outcome 11.1e seems to be the strongest one that would enable the researcher to make a causal statement concerning treatment. Cook and Campbell, however, warn that researchers should not plan on developing quasi-experimental research in the hopes of obtaining this outcome. Definitely, the designing of a nonequivalent control group study should be done with care and caution.

Research Examples

Nelson, Hall and Walsh-Bowers: Non-Equivalent Control Group Design.

The research study by Nelson, Hall and Walsh-Bowers (1997) specifically state that they used a non equivalent control group design to compare the long term effects of supportive apartments (SA), group homes (GH) and board-and-care homes (BCH) for psychiatric residents.

Supportive apartments and group homes are run by non-profit organizations. Board-and-care homes are run for profit. The main goal was to compare the two intervention groups: supportive apartments and group homes. They were unable to randomly assign participants to different housing settings. Nelson, Hall and Walsh-Bowers tried their best to match the residents but there were some significant differences in the composition of the groups that led them to use the non-equivalent control group design. With this design they decided to use BCH residents as the comparison group. They could not correct through matching the following variables that could have an effect on the dependent variables. The SA and GH groups tended to be younger than the BCH group (33 years versus 45) and have spent less time in residence (2.5 years versus 39 years). The SA and GH residents had a higher level of education than those in the BCH group did. Nelson, Hall and Walsh-Bowers found a significant difference between these groups on these variables. Even though gender was not significant there were more men than women in the SA and GH groups. In the BCH group there were more women than men residents.

Nelson, Hall and Walsh-Bowers (1997) state that the difference they found between these three groups on posttest measures could have been due to the selection problem and not the type of care facility.

Chapman and McCauley: Quasi-experiment

In this study, Chapman and McCauley (1993) examined the career growth of graduate students who applied for a National Science Foundation Graduate Fellowship Award. Although one can perhaps think of this study as a non-experimental one, Chapman and McCauley felt that it falls under the classification of quasi-experimental. We shall see why. In comparing the award winners and non-winners, the choice of winners was not exactly done at random. The study did not look at the Quality Group 1 applicants. The Group 1 applicants were in the top 5% and all received awards. The Quality Group 2 NSF applicants made up of the next 10% and are considered a highly homogeneous group. Awards are given to approximately half of a homogeneous group of applicants in a procedure that Chapman and McCauley say approximates random assignment to either fellowship or honorable mention. The students were assigned with regard to academic promise. Chapman and McCauley assumed that differences in performance between Quality Group 2 applicants who are and are not awarded an NSF fellowship could reveal the effect of positive expectations associated with this prestigious award.

The results showed that those getting an NSF award were more likely to finish the Ph.D. However, Chapman and McCauley found no reliable fellowship effect on achieving faculty status, achieving top faculty status, or submitting or receiving an NSF or a National Institutes of Health research grant. It seems that the positive expectancies associated with this prestigious award have some influence in graduate school and no effect on accomplishments after graduate school.

Time Designs

Important variants of the basic quasi-experimental design are time designs. The form of Design 9.6 can be altered to include a span of time:

Y_b	X	Y_a	
Y_b	~X	Y_a	
	X		Y_a
	~X		Y_a

The Y_a's of the third and fourth lines are observations of the dependent variable at any specified later date. Such an alteration, of course, changes the purpose of the design and may cause some of the virtues of Design 9.6 to be lost. We might, if we had the time, the patience, and the resources, retain all the former benefits and still extend in time by adding two more groups to Design 9.6 itself.

A common research problem, especially in studies of the development and growth of children, involves the study of individuals and groups using time as a variable. Such studies are longitudinal studies of participants, often children, at different points in time. One such design among many might be:

Design 11.2: A Longitudinal Time Design (a.k.a. Interrupted Time Series Design)

Y_1	Y_2	Y_3	Y_4	X	Y_5	Y_6	Y_7	Y_8

Note the similarity to Design 8.2, where a group is compared to itself. The use of Design 11.2 allows us to avoid one of the difficulties of Design 8.2. Its use makes it possible to separate reactive measurement effects from the effect of X. It also enables us to see, if the measurements have a reactive effect, whether X has an effect over and above that effect. The reactive effect should show itself at Y4; this can be contrasted with Y5. If there is an increase at Y5 over and above the increase at Y4, it can be attributed to X. A similar argument applies for maturation and history.

One difficulty with longitudinal or time studies, especially with children, is the growth or learning that occurs over time. Children do not stop growing and learning for research convenience. The longer the time period, the greater the problem. In other words, time itself is a variable in a sense. With a design like Design 8.2, Yb X Ya, the time variable can confound X, the experimental independent variable. If there is a significant difference between Yb and Ya, one cannot tell whether X or a time "variable" caused the change. But with Design 11.2, one has other measures of Y and thus a base line against which to compare the change in Y presumably due to X.

One method of determining whether the experimental treatment had an effect is to look at a plot of the data over time. Caporaso (1973) has presented a number of additional possible patterns of behavior that could be obtained from time-series data. Whether or not a significant change in behavior followed the introduction of the treatment condition is determined by tests of significance. The most widely used statistical test is ARIMA (Autoregressive, integrated moving average) developed by Box and Jenkins, (1970). (Also see Gottman, 1981). This method consists of determining whether the pattern of postresponse measures differs from the pattern of preresponse measures. The use of such a statistical analysis requires the availability of many data points. If enough data points cannot be collected to achieve the desired level of sensitivity Cook and Campbell advocate plotting the data on graph paper and visually determine whether a discontinuity exists between the pre- and post measures. Naturally this approach should be used only when one cannot use an appropriate statistical test, and one should remember that the number of preresponse data points obtained must be large enough to identify all the plausible patterns that may exist.

The statistical analysis of time measures is a special and troublesome problem: the usual tests of significance applied to time measures can yield spurious results. One reason is that such data tend to be highly variable, and it is as easy to interpret changes not due to X as due to X. That is, in time data, individual and mean scores tend to move around a good bit. It is easy to fall into the trap of seeing one of these shifts as "significant," especially if it accords with our hypothesis. If we can legitimately assume that influences other than X, both random and systematic, are uniform over the whole series of Y's, the statistical problem can be solved. But such an assumption may be, and probably often is, unwarranted.

The researcher who does time studies should make a special study of the statistical problems and should consult a statistician. For the practitioner, this statistical complexity is unfortunate in that it may discourage needed practical studies. Since longitudinal single-group designs are particularly well-suited to individual class research, it is recommended that in longitudinal studies of methods or studies of children in educational situations analysis be confined to drawing graphs of results and interpreting them qualitatively. Crucial tests, especially those for published studies, however, must be buttressed with statistical tests.

Multiple Time-Series Design

The multiple time-series design is an extension of the interrupted time-series design. With the interrupted time series design, only one group of participants was used. As a result, alternative explanations can come from a history effect. The multiple time-series design has the advantage of eliminating the history effect by including a control group comprised of an equivalent-or at least comparable-group of participants that does not receive the treatment condition. This is shown in Design 11.3. In this design one experimental group receives the treatment condition and the control group does not. Consequently, the design offers a greater degree of control over sources of alternative explanations or rival hypotheses. The history effects, for example, are controlled because they would influence the experimental and control groups equally.

Design 11.3: A Multiple Time-Series Design

Y_1	Y_2	Y_3	Y_4	X	Y_5	Y_6	Y_7	Y_8	Experimental
Y_1	Y_2	Y_3	Y_4		Y_5	Y_6	Y_7	Y_8	Control

Naturally, there are other possible variations of Design 11.2 besides Design 11.3. One important variation is to add one or more control groups; another is to add more time observations. Still another is to add more X's, more experimental interventions. (see Gottman, 1981; Gottman, McFall & Barnett, 1969; Campbell & Stanley, 1963).

Single Subject Experimental Designs

The majority of today's behavioral research involves using groups of participants. However, there are other approaches. In this section we deal with strategies for achieving control in experiments using one or a few participants. These single subject designs are sometimes referred to as the N = 1 design. Single subject designs are an extension of the interrupted time-series design. Where the interrupted time series generally looks at a group of individuals over time, e.g. children, the single subject study uses only one participant or at most a few participants. Even when a few participants are used, each is studied individually and extensively. These will also be called single subject designs or studies. . Although they have different names, they all share the following characteristics:

Only one or a few participants are used in the study

Each subject participates in a number of trials (repeated measures). This is similar to the within-participants designs described in Chapter 10.

Randomization (i.e., random assignment and/or random selection) procedures are hardly ever used. The repeated measurements or time intervals are instead assigned at random to the different treatment conditions.

These designs observe the organism's behavior before the experimental treatment and used as a baseline measure. The observations after the treatment are then compared with these baseline observations. The participant serves as his or her own control. These designs are usually applied in school, clinical, and counseling research. They are used to evaluate the effects of behavioral interventions. This mode of research is popular among those that do operant learning experiments or behavior modification.

Research using single participants is not new. There were some such as Gustav Fechner, who developed the discipline of psychophysics in the 1860s using only two participants: himself and his brother-in-law. Fechner is credited with inventing the basic psychophysical methods that are still used today to measure sensory thresholds. Fechner heavily influenced Hermann Ebbinghaus, who is known for his experimental work on memory. He also used himself as his own subject. Wilhelm Wundt, who is credited with founding the first psychological laboratory in 1879, conducted experiments measuring various psychological and behavioral responses in individual participants. Finally, I. P. Pavlov did his pioneering work on instrumental conditioning using individual dogs. The list of psychologists using single participants is extensive with most of them occurring before 1930 and the advent of R.A. Fisher and William Sealy Gossett's work in modern statistics

Behavioral scientists doing research before the development of modern statistics attempted to solve the problem of reliability and validity by making extensive observations and frequent replication of results. This is a traditional procedure used by researchers doing single-subject experiments. The assumption is that individual participants are essentially equivalent and that one should study additional participants only to make certain that the original subject was within the norm.

The popularity of Fisher's work on analysis of variance and Gossett's work on the Student's t-test led the way for group oriented research methodology. Some claim that these works were so popular that the single-subject tradition nearly became extinct. In fact, even in today's world, there are hiring practices at major universities that depend on whether the candidate is a group-oriented research scientist or a single-participants design oriented researcher. Despite the popularity of Fisher's methods and group oriented research, certain psychologists continued to work in the single subject tradition The most notable of these was Burrus Frederick Skinner. Skinner refrained from using inferential statistics. He does not advocate the use of complex of inferential statistics. Skinner feels that one

can adequately demonstrate the effectiveness of the treatment by plotting the actions of the organism's behavior. Skinner called this the cumulative record. Some such as E. L. Thorndike called it a "learning curve." Skinner felt that it was more useful to study one animal for 1000 hours than to study 1000 animals for one hour each. Murray Sidman (1960) in his classic book describes Skinner's philosophy of research. Sidman makes a clear distinction between the single-subject approach and the group approach to research. The single subject approach assumes that the variance in the subject's behavior is dictated by the situation. As a result, this variance can be removed through careful experimental control. The group-difference-research attitude assumes that the bulk of the variability is inherent and it can be controlled and analyzed statistically.

Some Advantages of Doing Single Subject Studies

Group-oriented research usually involves the computation of the mean or some other measure of average or central tendency. Averages can be misleading. Take at look at the two figures in Figure 11.3. Both have exactly the same values. If we were to compute the mean for the data in each group, we would find that they are exactly equal. Even if we computed the standard deviation or variance, we would find that the two measures of variability are exactly the same. However, visual inspection for the data shows that the graph, Figure 11.3a exhibits a trend while Figure 11.3b does not. In fact, Figure 11.3b shows what appears to be a random pattern. The single subject approach does not have this problem. A participant is studied extensively over time. The cumulative record for that participant show the actual performance of the participant.

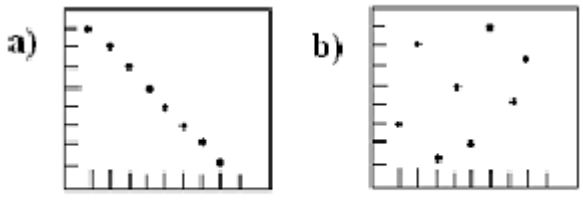


Figure 11.3 Systematic vs. Random

One of the major problems in using large samples is that statistical significance can be achieved for differences that are very small. With inferential statistics a large sample will tend to reduce the amount of error variance. Take the t-test as an example. Even if the mean difference remains the same, the increase in sample size will tend to lower the standard error. With a reduction of the standard error, the t-value gets larger, hence increasing its chance of statistical significance. However, statistical significance and practical significance are two different things. The experiment may have little practical significance even if it had plenty of statistical significance. Simon (1987) has criticized the indiscriminant use of large groups of participants. He finds them wasteful and unable to produce useful information. Simon advocates the use of screening experiments to find the independent variables that has the greatest effect on the dependent variable. These would be the powerful variables that produce large effects. Simon doesn't exactly endorse single-subject designs, but advocates for well constructed designs that uses only the number of participants necessary to find the strongest effects. He refers to these as "Economical Multifactor Designs. (1976). Single-subject researchers on the other hand, favor increasing the size of the effect instead of trying to lower error variance. They feel that this can be done through tighter control over the experiment.

In this same vein, single subject designs have the advantage over group-oriented designs in that with only a few participants they can test different treatments. In other words, they can determine the effectiveness or the ineffectiveness of a treatment intervention without employing a larger number of participants.

With single-subject studies, the researcher can avoid some of the ethical problems that face group-oriented researchers. One such ethical problems concerns the control group. In some situations, the control group does not receive any real treatment. Although in most of the studies done today the participants in the control group are not harmed in any way, there is still some ethical questions. Take for example the study by Gould and Clum (1995) to determine if self-help with minimal therapist contact is effective in the treatment of panic disorder. All participants in this study were sufferers of panic attacks. The participants were randomly assigned to either an experimental or control group. The experimental group received self-help material. The control "did not receive treatment during the course of the experiment."

(p. 536). Instead the control group was told that they were on the waiting list for treatment.

In the study of certain type of individuals, the size of the population is small and hence it would be difficult to do adequate sampling and obtain enough participants for the study. In fact, the study by Strube (1991) shows that even random sampling tends to fail when using small samples. If there are not enough participants of a certain characteristic available for study, the researcher can consider single-subject designs instead of abandoning the study. Simon (1987) cites the attempted study by Adelson & Williams in 1954 concerning the important training parameters in pilot training. The study was abandoned because there were too many variables to consider and not enough participants. Simon pointed out that the study could have been done, but not using the traditional group-oriented methodology.

Some Disadvantages of using Single-Subject Designs

Single-subject studies are not without their problems and limitations. Some of these will become more apparent when we actually discuss the types of single subject designs. Some of the more general problems with the single subject paradigm are external validity. Some find it difficult to believe that the findings from one study using one subject (Or maybe three of four) can be generalized to an entire population.

With repeated trial on one participant, one can question whether the treatment would be equally effective for a participant who has not experienced previous treatments. If we are talking about a therapeutic treatment, it may be the accumulation of sessions that is effective. The person going through the n-th trial can be a very different person from the one in the first trial. It is here that group-oriented research can eliminate this problem. Each person is given the treatment once.

Single-subject studies are perhaps even more sensitive to aberrations on the part of the experimenter and participant. These studies are effective only if the researcher can avoid biases and the participant is motivated and cooperative. The researcher can be prone to look only for certain effects and ignore others. We discussed earlier in this book about Blondlot. He was the only scientist able to see "N-Rays." It wasn't so much that he was a fraud, but that he was biased toward seeing something that was not there. A researcher doing single-subject research could be affected more so than the group-oriented researcher and need to develop a system of checks and balances to avoid this pitfall.

A number of research studies are by nature required to follow group-oriented methods and as such would be ill suited for single subject designs. For example, to study the behavior of jury members would require the use of groups and the influence of group dynamics. In a previous chapter, we discussed the research surrounding Janis' Groupthink. The study of this important phenomenon was best done with groups, since it was the group as a whole that displayed this phenomenon.

SOME SINGLE-SUBJECT RESEARCH PARADIGMS

The Stable Baseline: An Important Goal.

In a group-oriented design one group of participants is compared to another different group of participants. Or a group of participants receiving one condition is compared to the same set of participants receiving a different condition. We assume that the groups are equal prior to giving treatment so that if the dependent variable differs after treatment, we can associate that difference to the treatment. The determination of an effective treatment is done by statistically comparing the difference between the two groups on some outcome variable. When we use only one subject, however, a different tactic must be employed. In this one-subject situation we need to compare the behavior that occurs before and after the introduction of the experimental intervention. The behavior before the treatment intervention must be measured over a long enough time period so that we can obtain a stable baseline. This baseline, or operant level is important because it is compared to later behavior. If the baseline varies considerably, it could be more difficult to assess any reliable change in behavior following intervention. The baseline problem with single subject designs is an important one. For a complete description of the problems and possible solutions one should consult Barlow and Hersen (1976). Another excellent reference is Kazdin (1982).

An example where baseline measures are very important is in the use of a polygraph (lie detector). Here, the operator gets physiological measurements of

the person (defendant). The person is asked a number of questions where the participant is asked to give factual information which are known to be true (name, eye color, place of birth, etc.). The responses emitted are recorded and taken as the baseline measure for answering truthfully. Then another baseline is taken for responses to untrue statements. These are statements where the participant is told to lie to the question asked. After establishing these two baselines, the question of importance (i.e., did you commit the crime?) is asked and compared to the two baselines. If the physiological response resembles the lie baseline, the participant is told that they have lied.

Designs that use the Withdrawal of Treatment

The ABA Design.

The ABA design involves three major steps. The first is to establish a stable baseline (A). The experimental intervention is applied to the participant in the second step (B). If the treatment is effective, there will be a response difference from the baseline. In order to determine if the treatment intervention caused the change in behavior, the researcher exercises step three: a return to baseline (A). The third is required because we don't know what the response rate would have been if the participant received no treatment. We also would like to know whether the response change was due to the treatment intervention or something else.

A major problem with the ABA design is that the effect of the intervention may not be fully reversible. If the treatment involved surgery, where the hypothalamus is removed or the corpus callosum is severed, it would be impossible to reverse these procedures. A learning method that causes some permanent change in a participant's behavior would not be reversible.

There are also some ethical concerns about reverting the organism back to the original state if that state was an undesirable behavior (Tingstrom, 1996). Experiments in behavior modification seldom return the participant back to baseline. This return to baseline is called the withdrawal condition. To benefit the participant, the treatment is re-introduced. The design that does this is the ABAB design.

Repeating Treatments (ABAB Designs)

There are two versions of the ABAB design. The first was briefly described in the last section. It is the ABA design except that treatment is re-introduced to the participant and the participant leaves the study having achieved some beneficial level. Repeating the treatment also provides the experimenter with additional information about the strength of the treatment intervention. By demonstrating that the treatment intervention can bring the participant back to the beneficial level after taking that person back to baseline lends strength to the statement that treatment caused the change in behavior, i.e., evidence of internal validity. The ABAB design essentially produces the experimental effect twice.

The second variation of the ABAB design is called the alternating treatments design. In this variation there is no baseline taken. The A and B in this design are two different treatments that are alternated at random. The goal of this design is to evaluate the relative effectiveness of the two treatment interventions. The A and B may be two different methods of controlling overeating. The participant is given each treatment at different times. Over a period of time, one method might emerge as being more effective than the other. The advantage this design has over the first ABAB design is that there is no baseline to be taken and the participant is not subjected to withdraw procedures. Since this method involves comparing two sets of series of data, some have called it the between-series design.

There are some other interesting variations of the ABAB design where withdrawal of the treatment is not done. McGuigan (1996) calls it the ABCB design where in the third phase, the organism is given a "placebo" condition. This placebo condition is essentially a different method

Single subject designs are unlike group designs in that they only permit the researcher to vary one variable at a time. The researcher would not be able to determine which variable or which combination of variables caused the response changes if two or more variables are altered at the same time. The best that anyone can do is to make a statement that the combination of variables led to the change. However, the researcher won't be able to tell which one of how much of each. If there are two variables, called B and C, and the baseline is A, then a possible presentation sequence of the conditions would be A-B-A-B-BC-B-BC. In this sequence every condition was preceded and proceeded by the same

condition at least once with only one variable changing at a time.

The A-B-A-B-BC-B-BC design is often called an interaction design. All possible combinations of B and C, however, are not presented. Condition C is never presented alone (A represents the absence of B and C). The interaction here is different than the interaction discussed in the chapter on factorial designs. What is tested by this procedure is whether C adds to the effect of B.

In a learning experiment using this design, we could examine the effect of praising a student for giving the correct answer (C) to a question on geography along with a merit point (B). If we find that praise plus merit point has a greater effect than a merit point alone, we have information that is useful in designing a learning situation for this and other students. However, we will not know the singular effect of praise. Praise used by itself may have been just as effective as the merit point plus praise. Yet, praise by itself may have little or no effect. We can however assess praise by lengthening the single subject design. The sequence would be A-B-A-B-BC-B-BC-C-BC. However, lengthening a single subject experiment of this kind comes with other problems. A subject may become fatigued or disinterested. As a result, too long of a session may not produce useful information even though the design looks sound.

Some Research Examples.

Powell and Nelson: Example of an ABAB design.

This study by Powell and Nelson (1997) involved one participant, Evan, a 7-year-old boy who had been diagnosed with attention deficit hyperactivity disorder (ADHD). Evan was receiving 15 mg of Ritalin® per day. Most of Evan's behavior in the classroom was described as undesirable. Evan also had poor peer relations and did not understand his schoolwork. The undesirable behaviors included noncompliance, being away from his desk, disturbing others, staring *off,* and not doing work. Data were collected on the occurrence of interactions between Evan and his teacher.

The treatment intervention was letting Evan choose the class assignments he wanted to work on. There was choice and no-choice conditions. Baseline data were collected during the no-choice phase. Evan was given the same assignment as the rest of the class. During the choice phases, the teacher presented Evan with three different assignments and he chose one to complete. The assignment choices were identical in length and difficulty and varied only in content. Evan was not given the same choice of assignments twice.

Powell and Nelson used an ABAB design to evaluate the effects of choice making on Evan's undesirable behavior. During the no-choice condition, Evan was not given a choice of academic assignments. During the choice condition, he was allowed to choose his assignments.

The results showed that during the choice condition, the number of undesirable behaviors decreased. This study supported the efficacy of choice making as an antecedent control technique. These results suggest that educators attempting to manage the behaviors of students in classrooms may use choice procedures.

Rosenquist, Bodfish, & Thompson: Treating a Mentally Retarded Person with Tourette Syndrome.

The treatment of people with Tourette syndrome with a drug called haloperidol is quite common. However, not much was known about this drug's effectiveness on people who are mentally retarded and with Tourette syndrome. The identification of a person with both afflictions is difficult. Rosenquist, Bodfish, & Thompson (1997) wanted to determine the effectiveness of haloperidol. In this article, the person under study suffered from both conditions. Tourette syndrome is a neuropsychiatric condition where sufferers display simple and complex motor and vocal tics.

The single individual used in this study had a severe case of Tourette syndrome, which included catapults out of her chair; tic related choking, compulsive behavior and hyperactivity. Rosenquist, Bodfish, & Thompson used an ABABA design where A = baseline and B = treatment with haloperidol. The study was done over a 22-week period where the first 2 weeks served as the initial baseline. This was followed by 8 weeks of haloperidol treatment, a second 2 weeks of baseline, a second 8 weeks of haloperidol and then a final 2 weeks of baseline. The haloperidol was administered in a capsule that contained different dosages at different times. During the 8-week haloperidol treatments, dosages were altered. The dosages were increased every two weeks of the 8-week period except the last

two weeks where it was the medication washout period. The dosage changes were done without the administrator's knowledge. The entire study used a randomized videotape scoring procedure. This procedure was used to control experimenter bias. Videotapes were made of the participant, but the scoring of the tapes did not take place until the end of the study. Those viewing the tape were previously trained in using a tic checklist. Table 11.1 shows the results of this study. The dosage levels of haloperidol treatments are labeled numerically. The numbers under each condition are the mean number of tics exhibited by the participant. It appears from the data that the most effective dosage was 10 mg per day. Even though the study ended with the participant returning to baseline. The patient's guardian and the treatment team agreed to continue the haloperidol treatments at 10 mg per day.

Table 11.1 Mean Number of Tics Exhibited during Meal Time for Baseline and Haloperidol Treatment

Condition	Baseline	Haloperidol 1	Haloperidol 2	Haloperidol 5	Haloperidol 10
Simple Motor	34.8	11.0	12.5	21.4	6.3
Complex Motor	13.6	5.3	8.3	11.4	3.0
Simple Vocal	35.4	2.0	8.2	16.6	1.0
Complex Vocal	1.3	0.0	2.0	0.0	1.0

Using Multiple Baselines

There is a form of single subject research that uses more than one baseline. Several different baselines are established before treatment is given to the participant. These types of studies are called multiple baseline studies. There are three classes of multiple baseline research designs.

These are multiple baselines:
across behaviors
across participants
across environments

The use of multiple baselines is another approach to demonstrate the effectiveness of a treatment on behavior change. There is a common pattern for implementing all three classes of this design. That pattern is given in Figure 11.4.

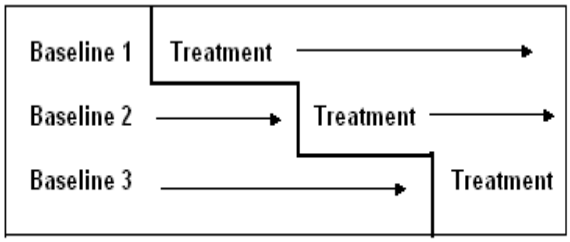


Figure 11.4 General Format of Multiple Baseline Design

With the multiple baselines across behaviors, the treatment intervention for each different behavior is introduced at different times. So looking in Figure 11.4, each baseline would be a baseline of a different behavior. In the case of an autistic child, Baseline 1 might be banging one's head against the wall. Baseline 2 would be talking constantly in different tones and noises and Baseline 3 would be hitting others. This is done to see if the change in behavior coincides with the treatment intervention. If one of the behaviors changes while the other behaviors remain constant or stable at the baseline, the researcher could state that the treatment was effective for that specific behavior. After a certain period of time has passed, the same treatment is applied to the second undesirable behavior. Every following behavior is subjected to the treatment in the same step-wise procedure. If the treatment intervention is effective in changing the response rate of each behavior, one can state that the treatment is effective.

An important consideration with this particular class of multiple baseline design is that one assumes the responses for each behavior are independent of the responses for the other behaviors. The intervention can be considered effective if this independence exists. If the responses are in some way correlated, then the interpretation of the results becomes more difficult.

In the multiple baseline design across participants, the same treatment is applied in series to the same behavior of different individuals in the same environment. When looking at Figure 11.4, each baseline is for a different participant. Each participant will receive the same treatment for the same behavior in the same environment. The study by Tingstrom, Marlow, Edwards, Kelshaw and Olmi (1997) is an example of a multiple baseline study across participants.

Their compliance-training package is the treatment intervention. This intervention uses time-in (physical touch and verbal praise) and time-out (a coercive procedure) to increase the rate of student compliance to teachers' instructions. The behavior of interest here is compliance to teachers' instructions. The environment is the classroom. The participants of this study were three students: A, B and C, who have demonstrated noncompliance behavior. All three students have articulation and language disorders. The design of the study followed the following intervention phases: baseline, time-in only, time-in/time-out combined and follow-up. Students B and C remained in the baseline phase while the time-in only phase was implemented for student A. When Student A showed a change in compliance, the time-in only phase was implemented for student B while student C remained in baseline. When student B showed a change in compliance, time-in only was implemented for student C. Tingstrom, Marlow, Edwards, Kelshaw and Olmi was able to demonstrate the effectiveness of the combined time-in and time-out intervention in increasing compliance.

In the multiple baseline design across environments, the same treatment is given to different participants who are in different environments. In Figure 11.4, each baseline would be for a different participant in a different environment. The treatment and behavior under study would be the same. Here we may have three different patients where each is a resident in a different type of psychiatric care facility such as those studied by Nelson, Hall and Walsh-Bowers (1997) discussed earlier in this chapter. In this study Nelson, Hall and Walsh-Bowers (1997) compared the long-term effects of supportive apartments (SA), group homes (GH) and board-and-care homes (BCH).

Study Suggestions

Look up each of the following studies and determine which ones are quasi-experimental, non-equivalent control group and single subject designs.

Adkins, V. K. & Matthews, R. M. (1997). Prompted voiding to reduce incontinence in community-dwelling older adults. *Journal of Applied Behavior Analysis, 30*, 153-156.

Streufert, S., Satish, U., Pogash, R., Roache, J. & Severs, W. (1997). Excess coffee consumption in simulated complex work settings: Detriment or facilitation of performance*?* *Journal of Applied Psychology, 82,* 774-782.

Lee, M. J. & Tingstrom, D. H. (1994). A group math intervention: The modification of cover, copy, and compare for group application. *Psychology in the Schools, 31*, 133-145.

Why is a baseline measure necessary in single subject designs?

Should the data from single subject designs be analyzed statistically? Why?

Give an example of where a single subject design should be used. Also cite a research situation where group designs is more appropriate.

A university student wants to do a time-series study on the effects of the full moon on psychiatric patients. What dependent variable should this student use? Where should this person look to locate the data for such a study?

Are single-subject studies applicable to medical research? Should medical school students be taught single subject designs? Read the following article:

Bryson-Brockmann, W. and Roll, D. (1996). Single-case experimental designs in medical education: An innovative research method. *Academic Medicine, 71*, 78-85.

Chapter Summary

True experiments are those where the experimenter can randomly select the participants, randomly assign the participants to treatment conditions and control the manipulation of the independent variable. The quasi-experimental design lacks one or more of these features.

Cook and Campbell (1979) covers 8 variations of the non-equivalent control group design. The one covered here is the no-treatment control group design. Five different results are discussed in terms of internal validity.

Time –series designs are longitudinal designs. It involves repeated measurements of the same dependent variables at different fixed intervals of time. Usually at some point, treatment intervention is introduced.

Selection and selection-maturation interactions are two alternative explanations that plague the results obtained from quasi-experimental designs.

Experiments using single participants are not new. The first people in experimental psychology used single subject designs

In single-subject researchers feel that with proper experimental control variability of the situation can be removed.

Group-oriented research feels variability be statistically analyzed.

Single-subject research has several advantages over group research in terms of flexibility and ethics. However it suffers from external validity credibility.

Small but statistically significant effects found in group research may have little clinical or practical significance and may have been artificially induced by large sample sizes. In this case, the effect size would be small. Single subject research concentrates on effect size and not sample size.

The establishment of a stable baseline is one of the most important tasks in single subject research.

The establishment of a baseline followed by administration of treatment followed by a withdrawal of the treatment is called the ABA design.

A major problem with the design is that the treatment may be irreversible or leaving the participant in the improved state rather than return the person to the original undesirable state.

A variation of the ABA design is the ABAB design where the participant is restored to the improved state.

In a single-subject study only one variable can be varied at a time.

The so-called interaction design does not permit the testing for an interaction as defined earlier in factorial designs. It merely examines two variables jointly.

There are three types of multiple-baseline designs. In each case, the intervention is introduced at different times for different behaviors, participants or environments. If behavior changes coincide with the introduction of treatment, this gives evidence that the treatment is effective.

References

Chapter 1

Barber, T. X. (1976). *Pitfalls in human research: Ten pivotal points*. New York: Pergmon

Blaser, M. J. (1996). The bacteria behind ulcers. *Scientific American, 274,* 104-107.

Braithwaite, R. (1996). *Scientific explanation*. Herndon, VA: Books International. (Original work published in 1953).

Bruno, J. E. (1972) (Ed.). *Emerging issues in education: Policy implications for the schools*. Lexington, MA: Heath

Buchler, J. (1955). *Philosophical writings of Peirce.* New York: Dover.

Conant, J. B. (1951). *Science and common sense*. New Haven: Yale University Press.

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth*. New York: The Free Press.

Dewey, J. (1991). *How we think*. Amherst: Prometheus. (Original work published 1933).

Graziano, A. M. & Raulin, M. L. (1993). *Research methods: A process of inquiry.* (2nd Ed.) New York: Harper Collins.

Hergenhahn, B. R. (1996). *Introduction to theories of learning.* (5th Ed.) Paramus, NJ: Prentice-Hall

Hoch, S. J. (1986). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, Cognition. 11*, 719-731.

Hurlock, E. (1925). An evaluation of certain incentives used in schoolwork. *Journal of Educational Psychology, 16*, 145-159.

Kagan, J. & Zentner, M. (1996). Early childhood predictors of adult psychopathology. *Harvard Review of Psychiatry, 3,* 341-350.

Kerlinger, F. (1977). The influence of research on educational practice. *Educational Researcher, 16*, 5-12.

Kerlinger, F. (1979). *Behavioral research: A conceptual approach.* New York: Holt, Rinehart and Winston,

Klayman, J. & Ha, Y.-W. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review, 94*, 21--28.

Klonoff, E. A. & Landrine, H. (1994). Culture and gender diversity in commonsense beliefs about the causes of six illnesses. *Journal of Behavioral Medicine, 17,* 407-418.

Lynch, M. P., Short, L. B., & Chua, R. (1995). Contributions of experience to the development of musical processing in infancy. *Developmental Psychobiology, 28,* 377-398.

Malone, J. C. (1991). *Theories of learning: A historical approach*. Pacific Grove: Brooks/Cole.

Marlow, A. G., Tingstrom, D. H., Olmi, D. J., & Edwards, R. P. (1997). The effects of classroom-based time-in/time-out on compliance rates in children with speech/language disabilities. *Child and Family Behavior Therapy, 19*, 1-15.

Michel, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, NJ: Erlbaum.

Nisbett, R. & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood-Cliffs, NJ: Prentice Hall

Polanyi, M. (1974). *Personal knowledge: Toward a post critical philosophy.* Chicago: University of Chicago Press. (Original work published in 1958).

Sampson, E. E. (1991). *Social worlds. Personal lives: An introduction to social psychology.* Orlando: Harcourt Brace Jovanovich.

Schunk, D. H. (1996). *Learning theories: An educational perspective*. Westerville, OH: Merrill.

Simon, C. W. (1976). *Economical multifactor designs for human factors engineering experiments*. Culver City, CA: Hughes Aircraft Company, 172 pp.

Simon, C. W. (1987). Will egg-sucking ever become a science? *Human Factors Society Bulletin, 30*, 1-4.

Simon, C. W. & Roscoe, S. N. (1984). Application of a multifactor approach to transfer of learning research. *Human Factors, 26*, 591-612.

Tang, J. (1993). Whites, Asians and Blacks in science and engineering: Reconsideration of their economic prospects. *Research in Social Stratification and Mobility, 12,* 289-291.

Tingstrom, D. H., Marlow, L. L. G., Edwards, R. P., Kelshaw, K. L., & Olmi, J. D. (1997, April). Evaluation of a compliance training package for children. In S. G. Little (Chair), *Behavior school psychology: Time-out revisited*. Symposium conducted at the 29th Annual Convention, National Association of School Psychologists, Anaheim, California

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content*. Cambridge, MA: Harvard University Press.

Weber, R. L. (1973). *A random walk in science: An anthology.* New York: Crane Rusak.

Wegner, D. M. (1989). *White bears and other unwanted thoughts: Suppression, obsession and the psychology of mental control.* New York: Penguin Books.

Wegner, D. M., Schneider, D. J., Carter, S. R., & White, T. L. (1987). Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology, 53(1*), 5-13.

Whitehead, A. N. (1992). *An introduction to mathematics.* New York: Oxford University Press. (Original work published in 1911).

Wood, R. W. (1973). N rays. In R. L. Weber's (Ed.) *A random walk in science*: *An anthology.* London: Institute of Physics.

Chapter 2

Anderson, C. D., Warner, J. L., & Spencer, C. C. (1984). Inflation bias in self-assessment examinations: Implications for valid employee selection. *Journal of Applied Psychology, 69*, 574-580.

Ayres, T. & Hughes, P. (1986). Visual acuity with noise and music at 107 dbA. *Journal of Auditory Research, 26*, 65-74.

Bahrick, H. P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *Journal of Experimental Psychology: General, 113*, 1-26.

Bahrick, H. P. (1992). Stabilized memory of unrehearsed knowledge. *Journal of Experimental Psychology: General, 121*, 112-113.

Berkowitz, L (1983).Aversively stimulated aggression: Some parallels and differences in research with humans and animals. *American Psychologist, 38,*1135-1144.

Bollen, K. (1980). Issues in the comparative measurement of political democracy. *American Sociological Review, 45*, 370-390.

Braithwaite, R. (1996). *Scientific explanation.* Herndon, VA: Books International. (Original work published in 1953).

Chamber, B. & Abrami, P. C. (1991). The relationship between student team learning outcomes and achievement, causal attributions and affect. *Journal of Educational Psychology, 83*, 140-146.

Cochran, S. D. & Mays, V. M. (1994). Depressive distress among homosexually active African-American men and women. *American Journal of Psychiatry, 15*, 524-529.

Cohen, M. R. (1997). *A Preface to logic.* New York: Meridian. (Original work published in 1956).

Cutler, W. B., Preti, G., Krieger, A. & Huggins, G. R. (1986). Human axilliary secretions influence women's menstrual cycles: The role of donor extract from men. *Hormones and Behavior, 20*, 463-473.

Dewey, J. (1982). *Logic: The theory of inquiry.*. New York: Irvington Publishers. (Original work published in 1938).

Dill, J.C. & Anderson, C.A. (1995). Effects of frustration justification in hostile aggression. *Aggressive-Behavior, 21(5),* 359-369.

Dion, K. L. & Cota, A. A. (1991). The Ms. stereotype: Its domain and its role of explicitness in title preference. *Psychology of Women Quarterly, 15*, 403-410.

Doctor, R. S., Cutris, D. & Isaacs, G. (1994). Psychiatric morbidity in policemen and the effect of brief psychotherapeutic intervention: A pilot study. *Stress Medicine, 10*, 151-157.

Dollard, J., Doob, L. Miller, N., Mowrer, O. & Sears, R. (1939). *Frustration and aggression*. New Haven: Yale University Press.

Elbert, J. C. (1993). Occurrence and pattern of impaired reading and written language in children with ADDS. *Annals of Dyslexia, 43*, 26-43.

Fallon, A. & Rozin, P. (1985). Sex differences in perception of desirable body shape. *Journal of Abnormal Psychology, 94*, 102-105.

Frentz, C., Gresham, F. M. & Elliot, S. N. (1991). Popular, controversial, neglected and rejected adolescents: Contrasts of social competence and achievement differences. *Journal of School Psychology, 29*, 109-120.

Glick, P., DeMorest, J. A., & Hotze, C.A. (1988). Keeping your distance. Group membership, personal space and request for small favors. *Journal of Applied Social Psychology, 18*, 315-330.

Guida, F. V. & Ludlow, L. H. (1989). A cross-cultural study of test anxiety. *Journal of Cross-Cultural Psychology, 20*, 178-190.

Hall, J. , Kaplan, D. & Lee, H. B. (October, 1994*). Counselor-Client matching on ethnic, gender and language: Implications for school psychology.* Paper presented at the Annual convention for the Association for the Advancement of Behavioral Therapy, San Diego, CA,

Hom, H. L. , Berger, M., Duncan, M. K., Miller, A. & Belvin, A. (1994). The effects of cooperative and individualistic reward on intrinsic motivation. *Journal of Genetic Psychology, 155*, 87-97.

Hurlock, E. (1925). An evaluation of certain incentives used in schoolwork, *Journal of Educational Psycholo*gy. *16*, 145-149.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E. & Nizam, A. (1997). *Applied regression analysis and other multivariable methods*. (3rd ed.) Belmont, CA: Duxbury.

Kumpfer, K. L., Turner, C., Hopkins, R. & Librett, J. (1993). Leadership and team effectiveness in community coalitions for the prevention of alcohol and other drug abuse. *Health Education Research, 8*, 359-374.

Langer, E. & Imber, L. (1980). When practice makes imperfect: Debilitating effects of overlearning, *Jour*nal *of Personality and Social Psychology, 37*, 2014-2024.

Lariviere, N. A. & Spear, N. E. (1996). Early Pavlovian conditioning impairs later Pavlovian conditioning. *Developmental Psychobiology, 29*, 613-635.

Little, S. G., Sterling, R. C., Tingstrom, D. H. (1996). The influence of geographic and racial cues on evaluation of blame. *Journal of Social Psychology, 136*, 373-379.

MacDonald, T. K., Zanna, M. P. & Fong, G. T. (1996). Why common sense goes out the window: Effects of alcohol on intentions to use condoms. *Personality and Social Psychology Bulletin, 22*, 763-775.

Moran, J. D., & McCullers, J. C. (1984). A comparison of achievement scores in physically attractive and unattractive students. *Home Economics Research Journal, 13*, 36-40.

Pedersen, D., Keithly, S. & Brady, M. (1986). Effects of an observer on comformity to handwashing norm. *Perceptual and Motor Skills, 62*, 169-170.

Poincare, H. (1996). *Science and method.* Herndon, VA: Books International. (Original work published in 1952).

Porch, A. M., Ross, T. P., Hanks, R. & Whitman, D. R. (1995). Ethnicity, socioeconomic background and psychosis-proneness in a diverse sample of college students. *Current Psychology: Developmental, Learning, Personality, Social, 13*, 365-370.

Reinholtz, R. K. & Muehlenhard, C. L. (1995). Genital perceptions and sexual activity in a college population. *Journal of Sex Research, 32,* 155-165.

Rind, B. (1997). Effects of interest arousal on compliance with a request for help. *Basic and Applied Social Psychology, 19*, 49-59.

Rosch, E. H. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.) *Cognitive development and the acquisition of language*. (pp. 111-144). New York: Academic Press.

Saal, F. K., Johnson, C. B., Weber, N. (1989). Friendly or sexy: It may depend on whom you ask. *Psychology of Women Quarterly, 13,* 263-276.

Saal, F. K. & Moore, S. C. (1993). Perceptions of promotion fairness and promotion candidates' qualifications. *Journal of Applied Psychology, 78,* 105-110.

Shaw, J. I. & Skolnick, P. (1995). Effects of prohibitive and informative judicial instructions on jury decision making. *Social Behavior and Personality, 23*, 319-325.

Spilich, G. J., June, L. & Remer, J. (1992). Cigarette smoking and cognitive performance. *British Journal of Addiction, 87*, 1313-1326.

Stein, J. A., Newcomb, M. D., & Bentler, P. M. (1996). Initiation and maintenance of tobacco smoking: Changing personality correlates in adolescence and young adulthood. *Journal of Applied Social Psychology, 26,* 160-187.

Stoneberg, C., Pitcock, N. & Myton, C. (1986). Pressure sores in the homebound: One solution , alternating pressure pads. *American Journal of Nursing, 86,* 426-428.

Stuart, D. L., Gresham, F. M. & Elliot, S. N. (1991). Teacher rating of social skills in popular and rejected males and females. *School Psychology Quarterly, 6*, 16-26.

Swanson, E. A., Maas, M. L., & Buckwalter, K. C. (1994). Alzheimer's residents' cognitive and functional measures: Special and traditional care unit comparison. *Clinical Nursing Research, 3,* 27-41.

Talaga, J. A. & Beehr, T. A. (1995). Are there gender differences in predicting retirement decisions? *Journal of Applied Psychology, 80*, 16-28.

Tingstrom, D. H. (1994). The Good Behavior Game: An investigation of teachers' acceptance. *Psychology in the Schools, 31*, 57-65.

Tulving, E. & Kroll, N. (1995). Novelty assessment in the brain and long-term memory encoding. *Psychonomic Bulletin and Review, 2*, 387-390.

Wegner, D. M., Schneider, D. J., Carter, S. R., & White, T. L. (1987). Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology, 53*, 5-13.

Winograd, E. & Soloway, R. (1986). On forgetting the locations of things stored in special places. *Journal of Experimental Psychology: General, 115*, 366-372.

Zajonc, R. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35*, 15l-175.

Chapter 3.

Annis, R. C. & Corenblum, B. (1986). Effect of test language and experimenter race on Canadian Indian children's racial and self-identity. *Journal of Social Psychology, 126,* 761-773.

Bahrick, H. P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *Journal of Experimental Psychology: General, 113*, 1-26.

Balzer, W. K. & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology, 77*, 975-985.

Bandura, A. & MacDonald, F. (1994). Influence of social reinforcement and the behavior of models in shaping children's moral judgments. In B. Paka (Ed.) *Defining perspectives in moral development. Moral development: A compendium, Vol. 1*. (pp. 136-143). New York: Garland Publications.. (Original work published in 1963).

Barron, F. & Harrington, D. M. (1981). Creativity, intelligence and personality. *Annual Review of Psychology, 32*, 439-476.

Bollen, K. (1979). Political democracy and the timing of development. *American* Sociological *Review, 44*, 572-587

Capaldi, D. H., Crosby, L. & Stoolmiller, M. (1996). Predicting the timing of first sexual intercourse for at-risk adolescent males. *Child Development, 67*, 344-359.

Colwell, J. C., Foreman, M. D., & Trotter, J. P. (1993). A comparison of the efficacy and cost-effectiveness of two methods of managing pressure ulcers. *Debubitus, 6(4*), 28-36.

Comrey, A. L. (1993). *EdITS Manual for the Comrey Personality Scales*. San Diego, CA: Educational and Industrial Testing Service.

Day, N. E. & Schoenrode, P. (1997). Staying in the closet versus coming out: Relationship between communication about sexual orientation and work attitudes. *Personnel Psychology, 50*, 147-163.

de Weerth, C. & Kalma, A. F. (1993). Female aggression as a response to sexual jealousy: A sex role reversal? *Aggressive Behavior, 19*, 265-279.

Doctor, R. S., Cutris, D. & Isaacs, G. (1994). Psychiatric morbidity in policemen and the effect of brief therapeutic intervention. *Stress Medicine, 10*, 151-157.

Eliopoulos, C. (1993). *Gerontological nursing,* 3rd. Ed. Philadelphia: J. B. Lippincott.

Francis-Felsen, L. C., Coward, R. T., Hogan, T. L., & Duncan, R. P. (1996). Factors influencing intentions of nursing personnel to leave employment in long-term care settings. *Journal of Applied Gerontology, 15*, 450-470.

Gillings, V. & Joseph, S. (1996). Religiosity and social desirability: Impression management and self-deceptive positivity. *Personality and Individual Differences, 21*, 1047-1050.

Gordon, R. A. (1996). Impact of ingratiation on judgment and evaluations: A meta-analytic investigation. *Journal of Personality and Social Psychology, 71,* 54-70.

Gresham, F. M. & Elliot, S. N. (1990*). Social skills rating system manual*. Circle Pines, MN: American Guidance Service.

Guida, F. V. & Ludlow, L. H. (1989). A cross-cultural study of test anxiety. *Journal of Cross-Cultural Psychology, 20*, 178-190.

Hart, S. D., Forth, A. E. & Hare, R. D. (1990). Performance of criminal psychopaths on selected neuropsychological tests. *Journal of Abnormal Psychology, 99*, 374-379.

Hernstein, R. J. & Murray, C. (1996). *The bell curve: Intelligence and class structure in American life.* New York: The Free Press.

Hodson, R. (1989). Gender differences in job satisfaction: Why women aren't more dissatisfied. *Sociological Quarterly, 30*, 385-399.

Hogan, J. & Hogan, R. (1989). How to measure employee reliability. *Journal of Applied Psychology, 74*, 273-279.

Hom, H. L. , Berger, M., Duncan, M. K., Miller, A. & Belvin, A. (1994). The effects of cooperative and individualistic reward on intrinsic motivation. *Journal of Genetic Psychology, 155*, 87-97.

Hutchinson, S. J. & Turner, J. A. (1988). Developing a multidimensional turnover prevention program. *Archives of Psychiatric Nursing, 2*, 373-378.

Kerlinger, F. N. & Pedhazur, E. (1973). *Multiple regression analysis in behavioral research*. New York: Holt, Rinehart and Winston.

Kounin, J. & Doyle, P (1975). Degree of continuity of a lesson's signal system and the task involvement of children. *Journal of Educational Psychology, 67*, 159-164.

Kumar, K. & Beyerlein, M. (1991). Construction and validation of an instrument for measuring ingratiatory behaviors in organizational settings. *Journal of Applied Psychology, 76*, 619-627.

Lester, D. (1989). Attitudes toward AIDS. *Personality and Individual Differences, 10*, 693-694.

Little, S. G., Sterling, R. C. & Tingstrom, D. H. (1996). The influence of geographical and racial cues on evaluation of blame. *Journal of Social Psychology, 136*, 373-379.

Luhtanen, R. & Crocker, J. (1992). A collective self-esteem scale: Self evaluation of ones' social identity. *Personality and Social Psychology Bulletin, 18*, 302-318.

Margenau, H. (1977). *The nature of physical reality.*. Woodbridge, CT : Ox Bow Press. (Original work published in 1950).

Martens, B. K., Hiralall, A. S., & Bradley, T. A. (1997). Improving student behavior through goal setting and feedback. *School Psychology Quarterly, 12*, 33-41.

Meiksin, P. F. & Watson, J. M. (1989). Professional autonomy and organizational constraint: The case of engineers. *Sociological Quarterly, 30*, 561-585.

Michel, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, NJ: Erlbaum.

Murphy, J. M., Olivier, D. C., Monson, R. R., Sobol, A.M. (1991). Depression and anxiety in relation to social status: A prospective epidemiological study. *Archives of General Psychiatry, 48(3),* 223-229.

Newcomb, T. (1978). The acquaintance process: Looking mostly backwards. *Journal of Personality and Social Psychology, 36, 1075-1083.*

Norman, D. (1976). *Memory* and *attention: An introduction to human information* processing. (2nd Ed.) New York: Wiley.

Northrop, F. (1983). *The logic of the sciences and the humanities*. Woodbridge, CT: Ox Bow Press. (Original work published in 1947).

Oldani, R. (1997). Causes of increases in achievement motivation: Is the personality influenced by prenatal environment? *Personality and Individual Differences, 22*, 403-410.

Onwuegbuzie, A. J. & Seaman, M. A. (1995). The effect of two constraints and statistics test anxiety on test performance in a statistics course. *Journal of Experimental Education, 63*, 115-124.

Orpen, C. (1996). Construct validation of a measure of ingratiatory behavior in organizational settings. *Current Psychology: Developmental, Learning, Personality, Social, 15(1),* 38-41

Oshagan, H. & Allen, R. L. (1992). Three loneliness scales: An assessment of their measurement. *Journal of Personality Assessment, 59*, 380-409.

Peng, S. S. & Wright, D. (1994). Explanation of academic achievement of Asian-American students. *Journal of Educational Research, 87*, 346-352.

Richter, M. L. & Seay, M. B. (1987). Anova designs with subjects and stimuli as random effects: Application to prototype effects in recognition memory. *Journal of Personality and Social Psychology, 53*, 470-480.

Scott, K. S., Moore, K. S. & Miceli, M. P. (1997). An exploration of the meaning and consequences of workaholics. *Human Relations, 50,* 287-314.

Shoffner, L. B. (1990). The effects of home environment on achievement and attitudes toward computer literacy. *Educational Research Quarterly, 14(1),* 6-14.

Silverman, S. (1993). Student characteristics, practice and achievement in physical education. *Journal of Educational Research, 87*, 54-61.

Skinner, B.F. (1945). The operational analysis of psychological terms. *Psychological Review, 52*, 270-277.

Smeltzer, S. C. & Bare, B. G. (1992). *Brunner and Suddarth's textbook of medical surgical nursing*, 7th ed. Philadelphia: J. B. Lippincott.

Somers, M. J. (1996). Modeling employees withdrawal behavior over time: A study of turnover using survival analysis. *Journal of Occupational and Organizational Psychology, 69*, 315-326.

Steele, C. M., Spencer, S. J. & Lynch, M. (1993). Self-image resilience and dissonance: The role of affirmational resources. *Journal of Personality and Social Psychology, 64,* 885-896.

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of facial feedback hypothesis. *Journal of Personality and Social Psychology, 54*, 768-777.

Strom, B., Hocevar, D. & Zimmer, J. (1990). Satisfaction and achievement antagonists in ATI research on student-oriented instruction. *Educational Research Quarterly, 14(4),* 15-21.

Strutton, D., Pelton, L. E. & Lumpkin, J. R. (1995). Sex differences in ingratiatory behavior: An investigation of influence tactics in the salesperson-customer dyad. *Journal of Business Research, 34*, 35-45.

Swanson, E. L., Maas, M. L. & Buckwalter, K. C. (1994). Alzheimer's residents' cognitive and functional measures: Special and traditional care unit comparison. *Clinical Nursing Research*, *3(1),* 27-41.

Tolman, E. (1951). *Behavior and psychological man.* Berkeley, CA: University of California Press.

Torgerson, W. (1985). *Theory and methods of scaling.* Melbourne, FL: Krieger. (Original work published in 1958).

Torrance, E. P. (1982). "Sounds and Images" production of elementary school pupils as predictors of creative achievements of young adults. *Creative Child and Adult Quarterly*, *7*, 8-14.

Underwood, B. (1957). *Psychological research.* New York: Appleton.

Warner, W. & Lunt, P. (1941). *The social life of a modern community.* New Haven: Yale University Press.

Wilson, G. D. & Reading, A. E. (1989). Pelvic shape, gender role conformity and sexual satisfaction. *Personality and Individual Differences, 10*, 577-579.

Chapter 4.

Barber, T. X. (1976). *Pitfalls of human research: Ten pivotal points*. New York: Pergamon.

Braud, L. & Braud, W. (1972) Biochemical transfer of relational responding. *Science, 176*, 942- 944.

Camel, J. E., Withers, G. S. & Greenough, W. T. (1986). Persistence of visual cortex dendritic alterations induced by post weaning exposure to a "superenriched" environment in rats. *Behavioral Neuroscience, 100*, 810-813.

Comrey, A. L. & Lee, H. B. (1995). *Elementary statistics: A problem-solving approach.* 3rd Ed. Dubuque, IA: Kendall-Hunt.

Guida, F. V. & Ludlow, L. H. (1989). A cross-cultural study of test anxiety. *Journal of Cross-Cultural Psychology, 20*, 178-190.

Holtgraves, T. (1997). Politeness and memory for wording of remarks. *Memory and Cognition, 25*, 106-116.

Prokasy, W. F. (1987). A perspective on the acquisition of skeletal responses employing the Pavlovian paradigm. . In I. Gormezano, W. F. Prokasy, & R. Thompson (Eds.), *Classical conditioning,* 3rd. Ed. Hillsdale, NJ: Lawrence Erlbaum.

Chapter 5

Barkan, J. D. & Bruno, J. B. (1972). Operations research in planning political campaign strategies. *Operations research, 20,* 926-936.

Cervone, D. (1987). Chi-square analysis of self-efficacy data: A cautionary note. *Cognitive Therapy and Research, 11*, 709-714.

Comrey, A. L. & Lee, H. B. (1992). *A first course in factor analysis, 2nd ed.,* Hillsdale, NJ: Lawrence Erlbaum Associates.

Congressional Quarterly, (1993) Volume 51, pp. 3497 (No. 266) and (No. 290).

Edgington, E. S. (1980). *Randomization tests*. New York: Marcel Dekker.

Edgington, E. S. (1996). Randomized single-subject experimental designs. *Behaviour Research and Therapy, 34,* 567-574.

Feller, W. (1967). *An introduction to probability theory and its applications*. (3rd. Ed.) New York: Wiley

Kemeny, J. (1959). *A philosopher looks at science*. New York: Van Nostrand Reinhold.

Kirk, R. E., (1990). *Statistics: An introduction, 3d. ed*. Fort Worth, TX: Holt, Rinehart and Winston.
Lee, H. B. & MacQueen, J. B. (1980). A k-means cluster analysis computer program with cross-tabulations and next-nearest neighbor analysis. *Educational and Psychological Measurement, 40*, 133-138.
Norusis, M. J. (1992). *SPSS/PC+ base system user's guide. Version 5.0.* Chicago: SPSS, Inc.
Poincare, H. (1996). *Science* and *method.* Herndon, VA: Books International. (Original work published in 1952).
Rand Corporation (1955). *A million random digits with 700,000 normal deviates*. New York: Free Press.
Simon, C. W. (1987). Will egg-sucking ever become a science? *Human Factors Society Bulletin, 30,* 1-4.
Stilson, D. W. (1966). *Probability* and *statistics in psychological research* and *theory.* San Francisco: Holden-Day
Walter, R. (1998). *The secret guide to computers*. (24th Ed.) Somerville, MA: Russ Walter.
Williams, B. (1978). *A sampler on sampling*. New York: John Wiley.

Chapter 6

Brandt, A. M. (1978). Racism and research: The case of the Tuskegee Syphilis study. *Hastings Center Report, 8,* 21-29.
Dawes, R. M. (1994). *House of cards. Psychology and psychotherapy built on myth*. New York: Free Press.
Erwin, E., Gendin, S. & Kleiman, L. (1994). *Ethical issues in scientific research: An anthology.* New York: Garland.
Ethical principles of psychologists. (1990, March). *American Psychologist*, p. 395.
Gould, S. J. (1981) *The mismeasure of man*. Norton: New York.
Jensen, A. R. (1992). Scientific fraud or false accusations? The case of Cyril Burt. In D. J. Miller & M. Hersen (Eds.). *Research fraud in the behavioral and biomedical sciences* (pp. 97-124). New York: Wiley.
Kamin, L. J. (1974). *The science and politics of IQ*. New York: Wiley.
Keith-Spiegel, P. & Koocher, G. P. (1985). Ethics in psychology: Professional Standards in Cases. Hillsdale, NJ: Erlbaum.
Milgrim, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67*, 371-378.
Miller, N. E. (1985). The value of behavioral research on animals. *American Psychologist, 40*, 423-440.
Saffer, T. H. & Kelly, O. E. (1983). *Countdown zero*. New York: Putnam.
Shapiro, K. J. (1998). *Animal models of human psychology: Critique of science, ethics, and policy.* Seattle, WA: Hogrefe & Huber Publishers.
Shrader-Frechette, K. (1994). *Ethics of scientific research*. New York: Rowman & Littlefield Publishers.
Smith, R. W. & Garner, B. (1976, June). *Are there really any gay male athletes*? Paper presented at the Society for the Scientific Study of Sex Convention, San Diego, California.

Chapter 7

Abbott, R. D. & Falstrom, P. M. (1975). Design of a Keller plan course in elementary statistics. *Psychological Reports, 36*, 171-174.
Amabile, T. (1979). Effects of external evaluation on artistic creativity. *Journal of Personality and Social Psychology. 37*, 221-233.
Bates, J. A. (1979). Extrinsic reward and intrinsic motivation: A review with implications for the classroom. *Review of Educational Research, 49*, 557-576.
Bergin, D. A. (1995). Effects of a mastery versus competitive motivation situation on learning. *Journal of Experimental Education*, 63, 303-314.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences, 2nd. Ed.* Hillsdale, NJ: Lawrence Erlbaum Associates.
Daniel, C. (1975). Applications of statistics to industrial experiments. New York: Wiley.
Deci, E. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality* and *Social Psychology, 18*, 105-115,
Eisenberger, R. & Cameron, J. (1996).Detrimental effects of reward: Reality or myth? *American Psychologist*, 51, 1153-1166.
Fisher, R. (1951). *The Design of Experiments, 4th ed*. New York: Hafner
Jaccard, J. & Becker, M. A. (1997). *Statistics for the behavioral sciences, 3rd Ed*. Pacific Grove, CA: Brooks-Cole.
Lepper, M. & Greene, D. (Eds.), (1978). *The Hidden Costs of Reward.* Hillsdale, N.J.: Erlbaum.
McCullers, J. C., Fabes, R. A. & Moran, J. D. (1987). Does intrinsic motivation theory explain the adverse effects of rewards on immediate task performance? *Journal of Personality & Social Psychology, 52*, 1027-1033.
Ross, L. L. & McBean, D. (1995). A comparison of pacing contingencies in class using a personalized system of instruction. *Journal of Applied Behavior Analysis, 28*, 87-88.
Senemoglu, N. & Fogelman, K. (1995). Effects of enhancing behavior of students and use of feedback-corrective procedures. *Journal of Educational Research, 89*, 59-63.
Sharpley, C. F. (1988). Effects of varying contingency and directness of rewards upon children's performance under implicit reward conditions. *Journal of Experimental Child Psychology, 45*, 422-437.

Simon, C. W. (1976). *Economical multifactor designs for human factors engineering experiments*. Culver City, CA: Hughes Aircraft Company, 172 pp.

Simon, C. W. (1987). Will egg-sucking ever become a science? *Human Factors Society Bulletin, 30(6*), 1-4.

Simon, C. W. & Roscoe, S. N. (1984). Application of a multifactor approach to transfer of learning research. *Human Factors, 26*, 591-612.

Thompson, S. (1980). Do individualized mastery and traditional instruction systems yield different course effects in college calculus? *American Educational Research Journal, 17*, 361-375.

Walster, E., Cleary, T. & Clifford, M. (1971). The effect of race and sex on college admission. Sociology of *Education,* 44, 237-244.

Chapter 8

Anastasi, A. (1958). *Differential psychology,* 3d Ed. New York: Macmillan, pp. 203-205.

Campbell, D. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*, 297-312.

Campbell, D. & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Dane, F. C. (1990). *Research methods*. Pacific Grove, CA: Brooks-Cole.

Edmondson, A. C. (1996). Learning from mistakes is easier said than done: Group and organizational influences on the detection and correction of human error. *Journal of Applied Behavioral Science, 32*, 5-28.

Garfinkle, P. E., Kline, S. A. & Stancer, H. C. (1973). Treatment of anorexia nervosa using operant conditioning techniques. *Journal of Nervous and Mental Disease, 157*, 428-433.

Graziano, A. M. & Raulin, M. I. (1993). *Research methods: A process of inquiry*, 2nd Ed. New York, NY: Harper Collins.

Hurlock, E. (1925). An evaluation of certain incentives used in schoolwork. *Journal of Educational Psychology, 16*, 145-159.

Johnson, J. D. (1994). The effect of rape type and information admissibility on perception of rape victims. *Sex Roles, 30*, 781-792.

Keith, T.Z. (1988). Research methods in school psychology: An overview. *School Psychology Review*, *17*, 502-520.

Skinner, B. F. (1968). *The technology of teaching*. New York: Appleton-Century-Crofts.

Solomon, R. (1949). An extension of control group design. *Psychological Bulletin, 46*, 137-150.

Stone-Romero, E. F., Weaver, A. E. and Glenar, J. L. (1995). Trends in research design and data analytic strategies in organizational research. *Journal of Management, 21*, 141-157.

Stouffer, S. (1950). Some observations on study design. *American Journal of Sociology, 55*, 355-361.

Thorndike, R. (1963). *Concepts of over-* and *underachievement.* New York: Teachers College Press, pp. 11-15.

Nesselroade, J., Stigler, S. & Baltes, P. (1980). Regression toward the mean and the study of change. *Psychological Bulletin, 88*, 622-637.

Walster, E., Cleary, T. & Clifford, M. (1970). The effect of race and sex on college admissions. *Journal of Educational Sociology,* 44, 237-244.

Wilson, F. L. (1996). Patient education materials nurses use in community health. *Western Journal of Nursing Research, 18*, 195-205.

Chapter 9

Boring E. (1954). The nature and history of experimental control. *American Journal of Psychology, 67*, 573-589

Campbell, D. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*, 297-312.

Campbell, D. & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Christensen, L. B. (1996). *Experimental methodology.* 6th. Ed. Needham Heights, MA: Allyn & Bacon.

Collins, L. M & Horn, J. L. (Eds.) (1991). *Best methods for the analysis of change: Recent advances, unanswered questions, future directions.* Washington, DC: American Psychological Association.

Cronbach, L. & Furby, L. (1970). How should we measure 'change'-or should we? *Psychological Bulletin, 74,* 68-80.

Friedenberg, L. (1995). *Psychological testing: Design, analysis and use*. Boston, MA: Allyn & Bacon.

Harris, C. W. (Ed.) (1963). *Problems in measuring change*. Madison, WI: University of Wisconsin Press.

Isaac, S. & Michael, W. B. Handbook in research and evaluation. 2nd Ed. San Diego, CA: EDITS

Jones, S. & Cook, S. (1975). The influence of attitude on judgments of the effectiveness of alternative social policies. *Journal of Personality and Social Psychology, 32*, 767-773.

Kirk, R. E. (1995). *Experimental designs: Procedures for the behavioral sciences* (3rd. ed.). Pacific Grove, CA: Brooks/Cole.

Matheson, D. W., Bruce, R. L. & Beauchamp, K. L. (1978). *Experimental psychology: Research design and analysis*. 3rd. Ed. New York: Holt, Rinehart & Winston.

Sax, G. (1997). *Principles of educational and psychological measurement and evaluation*, 4th Ed. Belmont, CA: Wadsworth.

Simon, C. W. (1976). *Economical multifactor designs for human factors engineering experiments*. Culver City, CA: Hughes Aircraft Company, 172 pp.

Solomon, R. (1949). An extension of control group design. *Psychological Bulletin, 46,* 137–150.

Underwood, B. (1957). *Psychological research*. New York, NY: Appleton.

Thorndike, E. (1924). Mental discipline in high school subjects. *Journal of Educational Psychology, 15*, I-22, 83-98.

Thorndike, E. & Woodworth, R. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review*, *8,*, 247-261, 384-395, 553-564.

Chapter 10

Clark, C. & Walberg, H. (1968). The influence of massive rewards on reading achievement in potential school dropouts. *American Educational Research Journal, 5,* 305-310.

Dear, R. E. (1959). *A principal-component missing data method for multiple regression models.* Technical Report SP-86. Santa Monica, CA: Systems Development Corporation.

Dolinski, D. & Nawrat, R. (1998). "Fear-then-relief" procedure for producing compliance: Beware when the danger is over. *Journal of Experimental Social Psychology, 34*, 27-50.

Edwards, A L. (1984). *Experimental* design *in psychological research* (5th. Ed.). Reading, MA: Addison-Wesley.

Elashoff, J. (1969). Analysis of covariance: A delicate instrument. *American Educational Research Journal, 6*, 383-401.

Fisher, R. A (1951). *The* design *of experiments.* {6th Ed.) New York, NY: Hafner

Flowers, M. (1977). A laboratory test of some implications of Janis' groupthink hypothesis. *Journal of Personality* and *Social Psychology,* 35, 888-896.

Freedman, J., Wallington, S. & Bless, E. (1967). Compliance without pressure: The effect of guilt. *Journal of Personality and Social Psychology,* 7,117-124.

Gleason, T. L. & Staelin, R. (1975). A proposal for handling missing data. *Psychometrika, 40*, 229-252.

Hilliard, S., Nguyen, M. & Domjan, M. (1997). One-trial appetitive conditioning in the sexual behavior system. *Psychonomic Bulletin & Review, 4*, 237- 241.

Hoyt, K. (1955). A study of the effects of teacher knowledge of pupil characteristics on pupil achievement and attitudes towards classwork. *Journal of Educational Psychology, 46*, 302-310.

James, W. (1890). *The principles of psychology.* New York: Holt, pp. 666-667

Janis, I. (1971). Groupthink. *Psychology Today,* 43-46, 74-86.

Kerlinger, F. N. & Pedhazur, E. (1973). *Multiple regression in behavioral research.* New York, NY: Holt, Rinehart and Winston

Kirk, R. E. (1995}. *Experimental designs: Procedures for the behavioral sciences* (3rd. ed.}. Pacific Grove, CA: Brooks/Cole.

Kolb, D. (1965). Achievement motivation training for underachieving high-school boys. *Journal of Personality and Social Psychology,* 2, 783-792.

Lindquist, E. (1940). *Statistical analysis in educational research.* Boston, MA Houghton Mifflin.

Miller, N, (1971}. *Selected papers.* New York, NY: Aldine

Miller, N. & DiCara, L. (1968). Instrumental learning of urine formation by rats: Changes in renal blood flow. *American Journal of Physiology, 215*, 677-683.

Pedhazur, E. (1996). *Multiple regression in behavioral research: Explanation and prediction.* (3rd Ed.) Orlando, FL: Harcourt Brace.

Perrine, R. M., Lisle, J. & Tucker, D. L. (1995). Effect of a syllabus offer of help, student age and class size on college students' willingness to seek support from faculty. *Journal of Experimental Education, 64*, 41-52.

Quilici, J. L. & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology, 88*,144-161.

Renner, V. (1970). Effects of modification of cognitive style on creative behavior. *Journal of Personality and Social Psychology,* 14 4, 257-262.

Rosch, E. (1973). Natural categories. *Cognitive Psychology,* 4, 328-350.

Sigall, H. & Ostrove, N. (1975}. Beautiful but dangerous: Effects of offender attractiveness and nature of the crime on juridic judgment. *Journal of Personality and Social Psychology, 31*, 410-414

Simon, C. W. (1976}. *Economical multifactor designs for* human *factors engineering* experiments. Culver City, CA Hughes Aircraft Company, 172 pp.

Snedecor, G. & Cochran, W. (1989}. *Statistical Methods,* 8th ed. Ames, Iowa: Iowa State University Press.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643-662.

Suedfeld, P. & Rank, A (1976). Revolutionary leaders: Long-term success as a function of changes in conceptual complexity. *Journal of Personality* and *Social Psychology, 34*,169-I78.

Taris, T. W. (1997). Reckless driving behaviour of youth: Does locus of control influence perceptions of situational characteristics and driving behavior? *Personality* and *Individual Differences, 23*, 987-995.

Thorndike, E. (1924). Mental Discipline in High School Studies. *Journal of Educational Psychology, 15*, 1-22, 83-98.

Tipper, S. P., Eissenberg, T. & Weaver, B. (1992}. The effects of practice on mechanisms of attention. *Bulletin of the Psychonomic Society, 30,* 77-80.

Winter, D. & McClelland, D. (1978). Thematic analysis: An empirically derived measure of the effects of liberal arts education. *Journal of Educational Psychology, 70*, 8-16.

Zakay, D., Hayduk, L. A & Tsal, Y. (1992). Personal space and distance misperception: Implications of a novel observation. *Bulletin of the Psychonomic Society, 30*, 33-35.

Chapter 11

Barlow, D. & Hersen, M. (1984). *Single case experimental designs: Strategies for studying behavior change* (2nd. Ed.). New York: Pergamon Press.

Box, G. E. P. & Jenkins, G. M. (1970). *Time-series analysis: forecasting and control*. San Francisco. CA: Holden-Day.

Campbell, D. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297-312.

Campbell, D. & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Caporaso, J. A. (1973). Quasi-experimental approaches to social sciences. In J. A. Caporaso & L. L. Ross (Eds.), *Quasi-experimental approaches*. Evanston, IL: Northwestern University Press.

Chapman, G. B. & McCauley, C. (1993). Early career achievements of National Science Foundation (NSF) Graduate Applicants: Looking for Pygmalion and Galatea effects on NSF winners. *Journal of Applied Psychology, 78*, 815-820.

Gottman, J. M. (1981*). Time-series analysis: A comprehensive introduction for social scientists*. New York, NY: Cambridge University Press.

Gottman, J. M., McFall, R. & Barnett, J. (1969). Design and analysis of research using time series. *Psychological Bulletin, 72*, 299-306

Gould, R. A. & Clum, G. A. (1995). Self-help plus minimal therapist contact in the treatment of panic disorder: A replication and extension. *Behavior Therapy, 26*, 533-546.

Graziano, A. M. & Raulin, M. L. (1993). *Research methods: A process of inquiry.* 2nd Ed. New York: Harper Collins.

Isaac, S. & Michael, W. B. (1987). *Handbook in research and evaluation*. 2nd Ed. San Diego, CA: EDITS

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York, NY: Oxford University Press.

Kirk, R. E. (1995). Experimental designs: Procedures for the behavioral sciences (3rd. ed.). Pacific Grove, CA: Brooks/Cole.

Nelson, G., Hall, G. B., & Walsh-Bowers, R. (1997). A comparative evaluation of supportive apartments, group homes and board-and-care homes for psychiatric consumer/survivors. *Journal of Community Psychology, 25*, 167-188.

Powell, S. & Nelson, B. (1997). Effects of choosing academic assignments on a student with attention deficit hyperactivity disorder. *Journal of Applied Behavior Analysis, 30*, 181-183.

Ray, W. J. (1997). *Methods: Toward a science of behavior and experience*. (5th Ed.) Pacific Grove, CA: Brooks-Cole.

Rosenquist, P. B., Bodfish, J. W. & Thompson, R. (1997). Tourette syndrome associated with mental retardation: A single-subject treatment study with haloperidol. American Journal of Mental Retardation, 101, 497-504.

Sanford, F. H. & Hemphill, J. K. (1952). An evaluation of a brief course in psychology at the U.S. Naval Academy. *Educational and Psychological Measurement, 12*, 194-216.

Sidman, M. (1960). *Tactics of scientific research*. New York, NY: Basic Books.

Simon, C. W. (1976). *Economical multifactor designs for human factors engineering experiments*. Culver City, CA: Hughes Aircraft Company, 172 pp.

Strube, M. J. (1991). Small sample failure of random assignment: A further examination. *Journal of Consulting and Clinical Psychology, 59*, 346-350.

Tingstrom, D. H (1996). ABAB designs and experimental designs. In T. K. Fagan & P. G. Warden (Eds.), *Historical encyclopedia of school psychology*. Westport, CN: Greenwood Press.

Tingstrom, D. H., Marlow, L. L. G., Edwards, R. P., Kelshaw, K. L., and Olmi, J. D. (April, 1997). Evaluation of a compliance training package for children. In S. G. Little (Chair), *Behavior school psychology: Time-out revisited*. Symposium conducted at the 29th Annual Convention, National Association of School Psychologists, Anaheim, California

Index

A
ABA design 145
ABAB design 145
A-B-A-B-BC-B-BC design 146
accidental samples 63
active and attribute variables 28
American Psychological Association 72
analysis of covariance 128
analysis of variance 99
approach of science 1
ARIMA (Autoregressive, integrated moving average) 141
Attrition 94
B
Bahrick 13
Barry Marshall 7, 69
basic aim of science 5
Bayesian approach 63
behavioral or observational definition 24
between-groups or experimental variance 41
Blondlot 7
C
Campbell and Stanley 89
Cartesian product of the independent variables and the dependent variable 97
change scores 106
Cluster sampling, 64
common sense 1
complete" designs 98
components of variance 48
compromise design 135
Comrey 40
concept 1
concepts 23
conceptual foundation for understanding research 97
construct 23
continuous and categorical variables 28
contro 78
control group 100
Control of Extraneous Independent Variables 92
CONTROL OF EXTRANEOUS VARIABLES 83
Cook and Campbell 136
correlated-groups' designs 120
Covariance 49
Criteria of Problems and Problem Statements 14
criteria of research design. 91
cross product, 50
D
Debriefing 73
Deception 70
Design 8.1 89
Design 8.2 89
Design 8.3 91
Design 8.4 91
Design 9.1 100
Design 9.2 102
Design 9.3 105
Design 9.4 107
Design 9.5 107
Design 9.6 107
Dewey 8
difference between common sense and science 2
difference scores 106
Dr. Stanley Sue 60
dynamic view 4
E
ecological representativeness 94
error variance 43
ethical guidelines for psychologists 72
ETHICS OF ANIMAL RESEARCH 74
Eugene S. Edgington 64
experimental and nonexperimental approaches 87
experimental mortality 94
experimental operational definition 25
experimental variance 41
experimentally manipulated independent variable 88
external validity 94
F
FACTORIAL CORRELATED GROUPS 125
factorial design 80, 114
Factorial Designs with More than Two Variables 114
four general ways of knowing 3
Frequency Distribution Matching Method 103
G
generality 7
generality and specificity of problems and hypotheses 17
generalizability, 93
Global Assessment Scale 60
good problem statement 13
H
Hawthorne study 73
heuristic view 4
highest form of experimentation 135
history 90
Hurlock 2
Hypothesis 8
hypothesis 14
I
inadequate research designs 87
incomplete" designs 98
independent and dependent variables 28
Instrumentation 94
interaction design 146
interaction hypothesis 80
internal validity 93
interrupted time series designs 135
J
J. Robin Warren 7
K
kinds of samples 62

L
latent variable, 33
longitudinal or time studies 141
M
MacQueen, 64
Manipulated variables 30
matching 103
Matching by Equating Participants 102
Matching by Holding Variables Constant 104
Matching by incorporating the Nuisance Variable into the Research Design 104
Matching versus Randomization 102
maturation 90
MAXIMIZATION OF EXPERIMENTAL VARIANCE 82
mean 40
mean square 40
measured variables 30
method of authority 3
method of intuition 3
method of science 3
method of tenacity. 3
MINIMIZATION OF ERROR VARIANCE 85
MULTIGROUP CORRELATED-GROUPS DESIGNS 125
multiple baseline research designs 147
multiple regression 99
multiple time-series design 141
multivariable nature of behavioral research 18
N
nature of theory 5
non-equivalent control group designs 135
nonexperimental research 87
nonobservables 33
no-treatment control group design 136
N-Rays 7
O
observation 23
observation-test-experiment phase 10
One-Shot Case Study 89
one-way analysis of variance design 97
onprobability samples 62
operational definition 24
P
Participant as Own Control 104
Peirce 3
peptic ulcers 7
plan of investigation 77
population variance 41
posttest only control group design 100
power analysis 79
pretest 89
Probability samples 62
Problems and hypotheses 16
Q
quasi-experimental designs 135
R
Rand Corporation, A Million Random Digits 56
Random sampling 54
random variance 43
RANDOMIZATION 56
Randomization 57
randomization 136
randomized subjects design 111
randomized subjects designs 119
randomness 43, 55
reactive measures 90
Reasoning-Deduction 8
Reasoning-Deduction 8
regression effect 90
Research Design 77
Research design has two basic purposes 77
research ethics 69
research problem 13
response variable 31
S
Saffer & Kelly 70
SAMPLE SIZE 60
Sampling variance 41
sampling without replacement 54
Sampson 3
Sampson's Two Views of the Science 6
science and common sense 1
SCIENCE AND ITS FUNCTIONS 4
SCIENTIFIC APPROACH 8
scientific research 7
Second-order 118
Selection 136
selection 94
selection-maturation interaction 138
Shrader-Frechette 69
Simon 9
Single subject designs 142
Sir Cyril Burt 71
Solomon four group design 109
standard deviations 40
Stanley Milgram 70
stereotype of scientists 4
stratified sampling 63
structure of research 77
T
table of random numbers 55
time designs 140
triple interactions 118
true experiment 135
Tuskegee Study 70
U
unethical research behavior 70
V
Variable representativeness 94
variables 23
variance 39
W
Whitehead 1