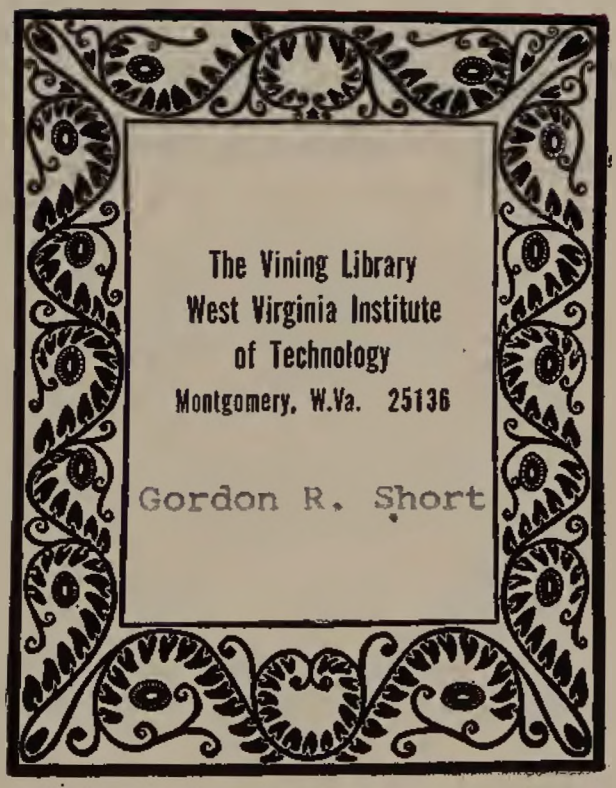
INFLUENCES on HUMAN DEVELOPMENT

Second Edition

Urie Bronfenbrenner
Maureen A. Mahoney



INFLUENCES ON HUMAN DEVELOPMENT

INFLUENCES ON HUMAN DEVELOPMENT

SECOND EDITION

Edited by

Urie Bronfenbrenner
Cornell University

Maureen A. Mahoney
Cornell University

The Dryden Press
Hinsdale, Illinois

Library of Congress Catalog Card Number: 74-6720
ISBN: 0-03-089413-1
Printed in the United States of America
5 6 7 8 9 090 9 8 7 6 5 4 3

CONTENTS

INTRODUCTION

This collection of readings reflects a new theoretical perspective in the study of human development and behavior. The change has not yet, and may never, become the norm, but it has significantly altered and enriched our understanding of the forces that shape the development of the human being.

The nature of the change is reflected in a superficial sign. Ten years ago, it would have been difficult to find research literature with a rigorous study that was not couched in terms of S and E. S was of course the subject, and E the experimenter. The latter, and often the former, had or needed no further identity—no name, no age, no sex, no role in life other than to participate in an experiment that ended on the same day, or even in the same hour, in which it began. In fact, in most experiments, the only participant besides S was a graduate student, whose prior relationship with the child was nonexistent, or, if existent, trivial in character. Indeed, it can be said that *much of American developmental psychology, even today, is the science of the behavior of children with adult strangers.*

More correctly, we should say it is a study of the behavior of a child with *one* strange adult. Existing theoretical models in human development typically assume a two-person system only. This assumption continues to be true even when the other person is a familiar figure—such as a parent, teacher, or therapist. Even if more than one is included in the research, e.g. mother and father, they are still treated separately. Three-person models are found in theory (e.g. Parsons & Bales) but rarely in practice. The S-and-E model also reflects the fact that, with a few exceptions, the process taking place is viewed as unidirectional. One is concerned, for example, with the effect of the experimenter's behavior—or that of the parent, teacher, or therapist —on the child, not the reverse.

Inevitably in a two-person model, and typically in the rare N-person (multi-person) system attention is limited to direct effects, i.e. the influence of A on B. There is neither interest in nor, often, the possibility of examining how the interaction of A and B (mother with child) might be affected by a third party C—say the father, or a second child, a grandparent, or teacher. One might call this the *second-order effect.* Only one substantial body of research in our field focuses on second-order effects; again, the other participant is a stranger. We refer to the growing literature on the effect of a stranger on the interaction of a child with his mother.

Finally, and most important of all, in much of our research the two-person system exists, or is treated as if it existed, in isolation from any other social context that could impinge on or encompass it.

These features so common in our research are hardly characteristic of the situations in which children actually live and develop. Thus in the family, the day care center, preschool, play group, school classroom, or neighborhood:

1. There are usually more than two people.
2. The child invariably influences those who influence him.

3. The other participants are not strangers but persons who have enduring roles and relationships vis-à-vis the child.
4. Finally, the behavior of all these persons is profoundly affected by other social systems in which these same persons participate in significant roles and relationships, both toward the child and each other.

The contrast between the conditions that have traditionally prevailed in our experiments and those that exist in the child's everyday life situation points up the fact that much of our research has been *ecologically invalid*. By removing the child from the environment in which he ordinarily finds himself and placing him in another setting which is typically unfamiliar, short-lived, and devoid of the persons, objects, and experiences that have been central in his life, we are getting only a partial picture both of the child and his environment. As a result, the potentialities of each to influence the other may be substantially greater than we have thus far seen.

Furthermore, existing theoretical models in human development typically focus attention on processes occurring within a single setting (e.g. family, day-care center, classroom, peer group). An ecological orientation points to the additional importance of relations *between systems* as critical to the child's development (e.g. the interaction between home and school, family and peer group).

Present theoretical orientations also tend to be limited to those ecological systems that actually contain the child himself (e.g. family, preschool, classroom, peer group); they seldom include the adjacent or encompassing system which may in fact determine what can or cannot occur in the more immediate context. Such encompassing systems include the nature and requirements of the parents' work, characteristics of the neighborhood, transportation facilities, the relation between school and community, and the role of television (not only in its direct effect on the child but in its indirect influence on patterns of family and community life). These, along with a host of other ecological circumstances and changes determine with whom and how the child spends his time: for example, the separation of residential and business areas, the disappearance of neighborhoods, zoning ordinances, geographic and social mobility, child labor laws, moonlighting, super markets, welfare policies, age segregation, the growth of single parent families, the abolition of the apprentice system, consolidated schools, commuting, the working mother, the delegation of child care to specialists and others outside the home, urban renewal, or the existence and character of an explicit national policy on children and families.

The emphasis, then, in this new theoretical perspective, is on the enduring environment of the child. This enduring environment, which we shall refer to as the *child's ecology,* consists of two concentric layers, the first superimposed upon the second.

A. The upper layer and the most visible is the immediate setting actually containing the child—home, school, street, playground, camp, etc. Every setting, in turn, is viewed along three dimensions:
   1. Design of physical space and materials.

2. People, in differing roles and relationships toward the child.
3. Activities in which the people are engaging—both with each other and with the child—including the *social meaning* of these activities.

B. The supporting and surrounding layer, in which the immediate setting is embedded, limits and shapes what can and does occur within the immediate setting:
1. Geographic and physical, for example, a housing project in which people live.
2. Institutions—the social systems which affect what can occur in the immediate setting—not just social class—but much more explicit systems such as health services and homemaker services.

But perhaps even more important are systems not directly or primarily focused on the child—shopping facilities, public transportation, parental working hours, traffic regulations, and a variety of other arrangements and customs that determine where children can be, and what activities they engage in with what kinds of people. In some countries, but not in our own, such arrangements grow in part out of an explicit national policy on children and family life.

The readings which follow reflect our concern with the scientific investigation of the child in his enduring environment and are examples of the new direction of research in human development. By and large, the participants in the experiments and investigations cited have lost their anonymity. They are not merely passive strangers, but active and familiar agents who play significant roles in each other's lives as family members, friends, pupils, teachers, and fellow workers. They come from particular segments of society and share common experiences. Indeed, the experiments in which they participate, whether made by man or nature, often continue over a period of days, weeks, or months—or, on occasion, even a lifetime. And these experiments are conducted not only in the laboratory, but also in the setting in which the activities under investigation normally occur, such as the home, play group, school, or community. In short, the new approach combines rigorous scientific method with the study of the developing person in the contexts in which he lives.

Our broad perspective on human development is reflected in the range of sources of material. The investigations here are drawn not only from psychological research, but also from biology, human genetics, sociology, and anthropology. The principal concern is the interplay of biological factors, human interaction, social structures, and cultural forces in shaping the individual. Thus we have selected articles which document specific influences—biological and social—on the process of development. Special emphasis is given to the implications of science for social policy, and in turn, to the manner in which social policy shapes, and sometimes stifles, scientific endeavor.

Some of the readings, like the articles of Spitz, and Skodak and Skeels, have stood the test of a quarter of a century of discussion; others are as yet untried, having been published only in the past year or specially written for this volume. But we commend them all to our readers as invitations to a new—ecological—perspective on the understanding of the process of human development.

The editors wish to express appreciation to Lorraine Sharpe, who typed, telephoned, and generally kept track of the proceedings. Nancy Burston also deserves special acknowledgment for her patience and judgment in preparing the index. Above all, our thanks go to the authors represented in these pages for making the real contribution—first in research and now in teaching.

Ithaca, New York
January, 1975

Urie Bronfenbrenner
Maureen Mahoney

INFLUENCES ON HUMAN DEVELOPMENT

part one
SCIENTIFIC METHOD IN THE STUDY OF HUMAN BEHAVIOR

Whatever the subject matter, the aims and methods of science are fundamentally the same: to formulate and verify general principles through the joint application of logic and objective, controlled observation. Science begins with a question and proceeds to a tentative answer. This tentative answer is referred to as a hypothesis. A hypothesis is usually derived from observation, theory, or a combination of both.

But, while there are no restrictions on the source of a hypothesis, there are rather rigorous technical requirements on the form that a hypothesis must take and on the conditions that must be met if the hypothesis is to be considered as verified or, more properly, as not rejected. Unfortunately, these technical matters are typically presented in abstract terms, without reference to concrete research problems. This is particularly true in discussions of the structure of hypotheses, the logic of verification, and the problems of measurement in the study of human behavior. For this reason, a special chapter has been written to deal with these topics.

1.1

The Structure and Verification of Hypotheses

Urie Bronfenbrenner
Maureen A. Mahoney

I. THE STRUCTURE OF HYPOTHESES

What Is a Hypothesis?

The scientific process begins with the asking of a question about the nature of objective reality. The scientist then tries to obtain an answer to his question through observing, classifying, and relating what he sees. He tries to arrive at tentative answers. Such tentative answers are called *hypotheses*.

Actually, in science, the term *hypothesis* is used with several different meanings. If we seek a general definition, one that encompasses all uses of the term, we can say that *a hypothesis is any supposition about a fact*.

Let us examine what forms such suppositions may take.

Forms of Hypotheses

Some years ago, Dr. Louis DiCarlo, then the director of a speech clinic in Syracuse, New York, was surprised by the unexpectedly high proportion of cases of cleft palate coming from certain sparsely populated counties in upstate New York. He was so struck by the phenomenon that he reported it to the district office of the U.S. Public Health Service directed by Dr. John Gentry. Gentry responded by doing what public health physicians have done for decades; he started putting up pins on a map, in this instance a map of New York State, one pin for each case, not only of cleft palate, but of all reported congenital malformations (which are deformities present at birth). When all the pins were in place, they made a pattern that Gentry found familiar. Where had he seen it before? After some effort, he remembered. It was in a geology course, on a map of igneous rock formations in New York State. Igneous rocks are those that were originally extracted from within the earth's surface. They are found in mountainous areas and glacial deposits. What is more, some of these rocks emit natural radiation, and, as Gentry knew, radiation had been suspected as a possible source of cleft palate and other deformities present at birth.

In short, Gentry had arrived at a hypothesis about DiCarlo's original observation. He had a "supposition about a fact." Indeed, he had several. In his published paper, Gentry (1959) investigated a series of hypotheses. Let us examine three of them:

1. Igneous rocks are radioactive.
2. Rates of congenital malformation are higher among residents of mountain areas than of plains and valleys.

3. An increase in the amount of natural radiation increases the rate of congenital malformation in the population living in the area.

Each of these three statements is a supposition about a fact, but the propositions differ in kind. To begin with, note how the first statement can be distinguished from the other two. The latter both postulate a relation between variation in one factor and variation in another, whereas the former involves no variation at all; it merely claims the presence of a particular characteristic in a class of objects, in this case of radioactivity in igneous rocks. A supposition that simply postulates the presence of a phenomenon we shall refer to as an *attributive* hypothesis. It merely asserts that a particular entity or event exists and can be observed. Witness the following examples:

1. In 1869, Mendelyev (1869), using his periodic table as a basis, predicted the existence of three then-unknown elements and specified their properties. The elements were subsequently discovered and found to have precisely the properties Mendelyev had attributed to them.
2. In 1929, the German psychiatrist, Hans Berger (1930), found evidence for the hypothesis that the human brain exhibited constant electrical activity, even during sleep.
3. In 1956, the Swedish biologists, Tjio and Levan (1956), using a new method for handling cells, obtained photographic evidence that the normal number of chromosomes in man was not 48, as had been thought previously, but 46. This hypothesis was subsequently confirmed by numerous investigators.
4. For over a quarter of a century, the American psychologist, J. B. Rhine (1964), has claimed validity for the hypothesis of *extrasensory perception*—the ability to receive information without benefit of the known physical senses (for example, thought transference, clairvoyance, etc.). Although much evidence has been adduced in support of the phenomena, most psychologists remain unconvinced, because of flaws in the methods used (for a critical review, see Gerden, 1962).

If we now turn our attention to hypotheses involving two variables, we observe a difference in the two examples given from Gentry's research. Strictly speaking, the first simply posits a statistical relationship between two variables, rate of congenital malformation on the one hand, geographical environment on the other. In contrast, the last proposition explicitly goes beyond a merely statistical association in claiming a cause and effect relationship: natural radiation is presumed to contribute to the development of congenital malformations.

To distinguish one kind of statement from the other, we shall designate the first as an *associative hypothesis,* and the second as a *causal hypothesis.* In other words, an *associative* hypothesis is one that stipulates a statistical relationship between two

variables without explicitly asserting that one of these variables influences the other. A *causal* hypothesis does add this additional specification; it asserts that variation in one factor produces variation in the other.

It is true that many associative hypotheses, when proved correct, add weight to an existing causal hypothesis, or suggest a new explanatory principle. We have an instance of the latter when confirmation of the purely geographical hypothesis proposed by DiCarlo led Gentry (1959) to identify natural radiation as a possible cause of congenital malformation.

But the associative hypothesis may not imply or suggest any explanations at all; it may merely call attention to a phenomenon to be explained. Consider, for example, the hypothesis proposed by Bronfenbrenner (1958) that over recent decades child rearing practices in the United States have become more permissive. Certainly one cannot conclude that the sheer passage of time makes successive generations more lenient in bringing up their children. We are left with the question as to what factors lead parents to treat their own children differently from the way in which they themselves were treated. In reply to this question Bronfenbrenner suggests that one explanation is to be found in the changing pattern of advice given to parents in successive editions of such widely read publications as the Children's Bureau bulletin on Infant Care (1951) and Benjamin Spock's pocketbook on the same subject (1957).

The foregoing example illustrates the nature and uses of the associative hypothesis. What this type of hypothesis does is to raise the question of a possible pattern among the variables observed. If the presence of the pattern is confirmed, this fact in time may serve one of three purposes: call attention to a new research problem, suggest a new causal hypothesis for investigation, or provide evidence in support or rejection of an already existing one.

The Causal Hypothesis Analyzed

We have said that a causal hypothesis differs from a purely associative one in stipulating a cause-and-effect relationship. What do we mean by this term?

The answer to this question turns out to be rather complex. We may begin with a formal definition. A cause-and-effect relationship exists when *a change in one variable is a necessary or sufficient condition to produce a given effect on another variable*.

Independent and Dependent Variables

Let us now look at the definition in greater detail. Notice first of all that the hypothesis postulates a relation between two factors, one which produces the change, the other the one in which the change is produced. The first, the one that produces the change, is called the *independent* or *antecedent* variable. The other, the one that is changed, is called the *dependent* or *consequent* variable, because it hangs (depends) on the independent variable and follows from it. Thus in Gentry's causal hypothesis the amount of natural radiation is the independent variable, the rate of congenital malformation the dependent variable.

Usually, the causal hypothesis is a one-way street. It works in only one direction. For example, changing the amount of natural radiation increases congenital malformation, but a change in the malformation rate, say by medical intervention, cannot have the slightest effect on the radioactivity of the rocks. But there can be causal relationships which operate in both directions. Contemporary theories of interpersonal relations (Homans, 1950; Heider, 1958; Newcomb, 1953) contain many hypotheses of this character. For example: the more similar two people are, the more likely they are to have positive feelings toward one another. But the process also operates in the reverse direction: the more two people like each other, the more they take on each other's characteristics. In other words, in this particular instance, independent and dependent variables are interchangeable. For this and other reasons, statements of causal hypotheses can be deceptive and may require careful analysis to ascertain which variable is which.

A helpful device in this connection is to conduct a hypothetical experiment in your mind—what the German psychologists call a Gedanken Experiment, a "thought experiment." Try changing each of the variables, and see what happens to the other one. If a change in A leads to a change in B, then A is an independent variable and B dependent. Conversely, if A varies with the change in B, it is A which is the dependent variable. If both change, we have what may be called a *reciprocal* causal relationship.

This does not mean, however, that a directional relationship, in which a change in one variable is followed by a change in the other, necessarily implies cause and effect. A case in point is provided by Bronfenbrenner's hypothesis, already cited, that over the past 25 years American child rearing practices have become more permissive. Clearly, it is the practices that have changed with time, rather than the reverse. In other words, the passage of time is the independent variable, parental behavior the dependent. But one cannot regard time as the factor which makes successive generations more lenient.

Necessary and Sufficient Conditions

This brings us to the main feature which distinguishes the causal hypothesis from its purely associative counterpart: the former always stipulates that one variable actually *influences* the other. This requirement may take two different forms, specified in the definition by the terms "necessary" or "sufficient." What does each of these terms mean?

Necessary means that *the effect cannot occur except under the* specified condition. For example, the disease process known as tuberculosis cannot occur in animal or man in the absence of a creature known as *Microbacterium tuberculosis,* the bacterium which we say "causes" the illness. It is a *necessary* condition.

Although the tubercular bacillus is necessary for the development of the disease, it is quite possible for a person to be a carrier of this baccillus and yet be perfectly healthy, because the human organism develops antibodies which keep the bacillus under control. In other words, the bacillus is a necessary condition, but not a *sufficient*

one. A sufficient condition is one that produces a given effect: for example, stick a healthy baby with a pin, and it cries. Of course inflicting pain is not the only condition that will make a baby cry. He will also cry if deprived of food. But sticking him with a pin will do the job. It is a *sufficient* condition.

An example of a less obvious sufficient condition is provided by a study conducted by Rheingold, Gewirtz, and Ross (1959) at the National Institute of Health. These investigators were interested in what it takes to make a baby vocalize more —that is, make more sounds. They showed that the rate of vocalizing among three-month-old infants could be increased markedly by such simple actions on the part of the adult as touching the baby's abdomen with the finger every time the infant made a sound, smiling at it, and clucking at it ("tsk, tsk, tsk") in return. When the adult experimenter reacted in this way, the baby's vocalization rate just about doubled. None of these things was *necessary* to make the baby vocalize more, but they were *sufficient.*

Notice that in actual practice, a sufficient condition may not produce the given effect in every case. A baby will not vocalize every time an adult touches, smiles, or makes noises at it. It may be tired, fearful, or simply distracted by some other event. Nor will an infant cry every time it is stuck with a pin; it may have a bad case of laryngitis, the pin may not hit a pain spot, or the baby may be crying already. In other words, something can interfere.

To put it in another way, a given phenomenon can occur only under certain *boundary conditions.* These boundary conditions are of two kinds. First, all *necessary* conditions to make possible variation in the dependent variable must be satisfied. Thus a baby cannot cry without a functioning voice box (larynx); it cannot smile if the facial nerve is injured; it can do neither if it is seriously ill or exhausted. In other words, a sufficient condition cannot be effective until all necessary conditions are met. But even if all necessary conditions for producing the effect are satisfied, the infant may not be able to perceive the stimulus, and hence not make the response. He may not hear the "tsk, tsk," see the smiling face, or feel the touch or the pin because the stimulus doesn't hit a functioning touch receptor. These are conditions that must be met—not for the effect to occur in the dependent variable (a blind, deaf, or tactilely insensitive baby may be perfectly able to smile and cry) but for the independent variable to be functional in producing the effect. In other words, we are dealing here with a class of variables which limits not the dependent variables but the *relation* between the independent and dependent variables. To distinguish such requirements from what we have called necessary conditions—those that apply to the dependent variable itself—we shall refer to them as *contingent* conditions. Given an independent variable x and a dependent variable y, a contingent condition is one that is not necessary to produce a given effect in y but is required *for x* to produce the given effect in y. Boundary conditions, then, include all necessary and contingent conditions.

We are now in a position to offer a definition of a sufficient condition: *within a given set of boundary conditions, x is a sufficient condition if a change in x produces a change in y.*

Interrelation of the Three Types of Hypotheses

A causal hypothesis, then, is one that stipulates one variable as a necessary or sufficient condition for another. Note also that the causal hypothesis always implies an associative hypothesis as well, since a necessary or sufficient condition inevitably produces a statistical relationship between independent and dependent variables. Finally, an associative hypothesis assumes that each of its components is an observable phenomenon; in other words, it implies two attributive hypotheses. In short, the three types of hypotheses fall into a nested arrangement, like a set of Russian dolls, with the causal hypothesis containing the associative hypothesis, and the associative containing two attributive hypotheses.

If one asks whether there exists still a higher order construct encompassing more than one causal hypothesis, the answer is of course found in the concept of a *theory.* Although this term has no rigorous definition or use, it usually refers to a body of interrelated hypotheses such as Freudian theory, learning theory, dissonance theory, or role theory, all of which have been employed in the attempt to understand the development of human behavior.

As we shall see, the hierarchical structure of the causal hypothesis dictates the steps to be followed in proving a cause-and-effect relationship. The investigator begins by specifying a procedure for observing each of the separate variables in the investigation. By demonstrating that such observations can be made, he in effect confirms the attributive hypothesis for each factor. He then proceeds to demonstrate a statistical association between two of the variables under circumstances which require such an association to be the product of a cause-and-effect relationship.

We turn next to an examination of the principles and procedures involved in the process of hypothesis testing.

II. THE LOGIC OF VERIFICATION

In our discussion we shall focus attention on the causal hypothesis since, as we have seen, this incorporates all other types as well. Given our definition of a cause-and-effect relationship, the task of verification becomes that of demonstrating that a given condition is necessary or sufficient. As we shall see, the proof begins somewhat differently for the two cases, but ends with a similar exacting requirement.

The Logic of Demonstrating a Cause-and-Effect Relationship

Let us begin with the case of a necessary condition. Since necessary means that the dependent variable (B) cannot occur without the independent variable (A), one must show first of all that all possible instances of (B) are preceded or accompanied by (A). The situation with a sufficient condition involves a different requirement. Here variation in one factor must produce a given effect on the other. In other words, a change in A must be followed by the specified effect in B.

Now let us suppose that one or the other of the above requirements is fulfilled. Does this mean that a necessary condition has in fact been demonstrated in the one case, and a sufficient condition in the other? Unfortunately not. If that were all there was to it, the work of the scientist would be much easier than it is, and much less interesting.

The inadequacy of either of the above demonstrations is brought home by the tale of a man who obviously understood exactly what he was doing. This chap discovered that every time he drank a highball of scotch and ginger ale, he became drunk. The next time he tried gin and ginger ale and got the same effect. Then he experimented with rye and ginger ale and you know what happened. What to do? "Aha,' " he said, "I know what does it." And the next time he eliminated the ginger ale.

This example shows that, to demonstrate a necessary or sufficient condition, it is not enough to show the required kind of relationship between the independent variable A and the dependent variable B. One must also establish that no other factor X is functioning as an independent variable. For if it is, then the results are *confounded;* that is, one cannot separate the effect of A from the effect of X. And if X is actually responsible for the effect, then A is neither necessary nor sufficient.

Demonstrating that a given effect is produced by A, and not by some other factor, can turn out to be a fairly complicated task, especially if people in general—and scientists in particular—are sure they already know what the necessary or sufficient condition is.

The Goldberger Story

To show how complicated it can be, let us look at the case history of a hypothesis that is all inclusive, since it postulates a condition which is both necessary and sufficient at the same time. The case history begins back in the early 1900's.[1] At that time they were having trouble down in the South—Virginia, South Carolina, Mississippi, and Georgia. It was a different kind of trouble. But they did what they so often do when there is trouble in the South; they sent a man from Washington. This man's name was Joseph Goldberger.

In the period 1900 to 1914, 100,000 people died every year in the American South from a disease called *pellagra.* In Italian, pellagra means "rough skin," one of the symptoms of the disease. Other symptoms, which come on gradually, include running sores, foul odor, vomiting, diarrhea, terrible pain, nervous and mental disturbances, and finally that ultimate symptom—death.

In 1912, just before Goldberger was sent down from Washington, a top medical commission had surveyed all the available evidence and had concluded that pellagra was "a specific infectious disease communicable from person to person by means at present unknown." Goldberger got off the train at Spartanberg, South Carolina, where there was a pellagra hospital. As an investigator he saw his first task as one

of observing. Pellagra cases were everywhere; not only in the hospital but all round the countryside. Goldberger saw emaciated bodies, sallow sunken faces, cases of insanity.

But he noticed other things too. It was cotton country. In the mill towns and villages everyone was trying to live by cotton, and not succeeding. Families were attempting to survive on fifteen dollars a week, and they looked it. If they paid the rent and bought the barest necessities in clothes, there wasn't half enough left for food. And it was the same in Georgia, in Florida, Alabama, and Virginia. In fact, you can still see it—not the pellagra, but the poverty. And it still takes a toll, both physically and psychologically.

But back to Goldberger. As he visited hospitals and institutions, he noticed a strange thing: none of the nurses, attendants or other employees, who were in daily contact with the pellagra cases, had ever developed the disease. He also observed that in an orphanage he visited, there were virtually no pellagra cases below the age of six or above the age of twelve, but in the middle group, practically all the children had pellagra. Goldberger knew that he had run onto a critical natural experiment. He was hot on the trail of a hypothesis. In the institutional setting, it didn't take him very long to find what he was looking for. It had to do with food.

1. The middle group ate only the regular institutional diet.
2. The little ones got a regular milk supplement.
3. The older group supplemented their diet by foraging on their own.

And what did the regular institutional menu consist of? Traditional Southern dishes—but low cost ones—biscuits, hominy grits, corn mush, syrup, molasses, gravy, sowbelly. Plenty to eat, but no milk, no eggs, no butter, no meat—in short, no animal proteins.

It would seem he had found the answer. All that remained was to publish it to the world. But Goldberger didn't. Instead, he went to another institution with lots of pellagra cases, instituted dietary changes, and the pellagra went away. Just like that.

Then Goldberger published. And what was the reaction? Leaders of the medical profession were unconvinced. Goldberger had not found the agent of contagion, and "as is well known, pellagra is a contagious disease."

So Goldberger decided to do an experiment that would be convincing. He would show that he could produce pellagra in healthy human beings. But he needed volunteers. So he went to the Governor of Mississippi with a request. Would the Governor grant pardons to 12 convicts with long-term sentences in state prison if they volunteered to stay in Goldberger's experiment for six months. The risk: pellagra, the gain: freedom.

The Governor of Mississippi was an understanding man, Goldberger got his volunteers—twelve enterprising fellows—embezzlers, highwaymen, murderers—and in the best of health.

When the experiment started, the volunteers were sure they had a good thing going. Instead of the regular prison fare, they were being given special meals—and what meals! The food was well cooked, tasted fine, and every man got all he wanted. For breakfast there was: biscuits, fried mush, syrup. For dinner: corn bread, cabbage, sweet potatoes, grits. For supper: rice, gravy, fried mush, coffee with sugar.

Eighty additional convicts, who were designated as a control group, lived under the same conditions as the volunteers—except for diet. After a few weeks, the volunteers began to doubt that they had made such a fine bargain. They began to feel queer—headache, stomach ache, dizziness. By the fifth month, the skin began to scale, and all the classical symptoms of pellagra were present.

The convicts were pardoned, released, and offered a cure, but they were too frightened to take it. Goldberger hastened to report his findings at a meeting of public health experts in Washington. At the meeting, the Chairman had some remarks to make. He said he wished "to enunciate certain beliefs concerning pellagra, these beliefs being backed by research done by good American citizens such as Siler, Garrison, and MacNeal." The Chairman then went on to state that pellagra was "a specific infectious disease, communicable from person to person by unknown means."

In short, Goldberger's argument was unconvincing in the light of prevailing views. How did Goldberger react? With another experiment. He recognized that he had not yet proved his point. He still had a requirement to fulfill. He had built a case for his own independent variable—diet, but he had not shown that another independent variable, X, had not influenced the results. Specifically, he had not dealt with the accepted explanation for pellagra, and for virtually all the other diseases known at that time—contagion. He hadn't really answered the principal argument of his opponents.

Goldberger set out to eliminate the contagion hypothesis once and for all. This is how he did it.

He drew an ounce of blood from a pellagra patient suffering an acute attack. A colleague then injected five cubic centimeters of the patient's blood into Goldberger's shoulder. In addition, secretions from the patient's nose and throat were swabbed into Goldberger's nose and throat. Finally, he selected two patients—one with scaling sores and the other with diarrhea. He scraped the scales from the sores, mixed the scales with four cubic centimeters of urine from the same patients, added an equal amount of liquid feces, and rolled the mixture into little dough balls by the addition of a few pinches of flour. The pills were then taken voluntarily by him, by his assistants and by his wife. Publication of the experiment was withheld for five months to make sure that there were no delayed effects. The report mentioned no names. It simply spoke of 15 men and a housewife.

The experiment made its point. Finally, the "infectionists" were silenced. Years after Goldberger's death the substance contained in animal protein necessary to prevent pellagra was named Vitamin G. But the name didn't stick. Personal tribute had to give way to scientific progress. In 1936, the critical substance was identified more specifically as nicotinic acid.

Proof as a Psychological Problem

As Goldberger's experience demonstrates, the process of proof may demand more than satisfying logical requirements. It may also necessitate overcoming psychological barriers. Verification does not take place in a vacuum; it occurs in the minds of men. The human mind is not always capable of seeing objectively, let alone thinking objectively. Both perception and thought can be distorted by conviction and desire. Nor are scientists any more immune to such distortions than other men. The medical scientists who heard and read Goldberger's reports were neither stupid nor evil. They simply *knew* that pellagra was a contagious disease, and—what is more important—they knew this was the opinion of the "best authorities."

Since one does not have to be a social scientist to regard himself or be regarded by others as an authority on problems of human behavior, the researcher in this sphere frequently finds himself confronted with a task of psychological as well as logical persuasion.

Some Principles of Verification

Nevertheless, it is logic that lies at the heart of the matter. To make explicit the principles involved, let us analyze the steps in the argument as exemplified in Goldberger's work.

What was Goldberger's line of reasoning? Where did he begin?

First he established that every case of pellagra he observed ate a certain diet; persons not eating the diet never contracted pellagra.

In other words, Goldberger showed that two variables were related statistically. He confirmed what we have called an associative hypothesis.

What does this demonstration accomplish?

Does it prove a necessary condition? He had shown that every case of pellagra had eaten the same kind of diet.

Could some other factor have accounted for the same results?

Yes, contagion—through infected food; through sewage.

Does his demonstration prove a sufficient condition? He had shown that a difference in diet was related to pellagra, but could some other factor have accounted for the same results? Again the answer is yes. The food could have been infected. This brings us to our first principle.

Principle I

A causal hypothesis is not proved so long as an alternative hypothesis can be offered to explain the same findings.

In other words, no causal hypothesis can be regarded as confirmed until all plausible alternative hypotheses have been eliminated. This requirement points up a critical limitation of any purely associative hypothesis, such as the one we have just been considering. The demonstration that a particular kind of association exists of course does not eliminate the possibility that this association is the product of some third factor. This is one of the reasons why—

Principle II

The demonstration that a particular kind of association between two variables exists cannot, by itself, prove that one of these variables is a necessary or sufficient condition for the other.

So what did Goldberger's first step accomplish? Could he draw any inferences from it at all?

What was his associative hypothesis? Was it simply that diet and pellagra tend to occur together?

Suppose Goldberger had found the following:

	PERCENT PELLAGRA CASES
Eating diet	85%
Not eating diet	5%

Do we have an association? Yes.

Do we have a necessary condition? No.

Putting the matter more generally—in establishing a necessary condition, *all* positive instances of the dependent variable must be associated with *presence* of the independent variable. There can be no exceptions.

And Goldberger didn't find any. The actual percentage among those not eating the diet was zero.

So the results for an associative hypothesis can be useful after all. They can be used to *disprove* a hypothesis about a necessary condition. We may put this as a general principle.

Principle III

Results for a hypothesis of association can disprove a hypothesis of necessity so long as there is even a single case in which the dependent variable occurs without the independent variable.

Can results for a hypothesis of association disprove a hypothesis of sufficient condition as well?

The answer to this question turns out to be somewhat more complicated. To demonstrate the point, we must take a different example. Suppose that instead of working with pellagra, Goldberger had been working with a still unsolved mystery, the common cold, and was investigating the hypothesis that exposure to damp weather increases susceptibility. In studying case records of visits to general practitioners by persons complaining of colds, he finds the following:

	% OF ALL PATIENTS WITH COLDS
Persons working mainly indoors (businessmen, secretaries, etc.)	50%
Persons working mainly outdoors (laborers, policemen, etc.)	50%

In other words, no association. Does this mean that exposure to dampness does not increase susceptibility to colds? No.

Persons working outdoors may be healthier.

Persons working indoors may take their complaints to a doctor more often.

Principle IV

A negative result for a hypothesis of association cannot, by itself, disprove a hypothesis of sufficiency, since it does not eliminate the possibility that an association in fact exists but is counteracted by some third factor.

Since results for a purely associative hypothesis can neither prove nor disprove that a variable is a sufficient condition, do they have any utility at all in investigating hypotheses of sufficiency?

Did it make any difference to Gentry that the association between congenital malformation and rock formations came out as it did?

Of course; the association suggested that he was on the right track.

Since an association between two variables can imply a possible causal relation—

Principle V

A positive result for an associative hypothesis increases the probability that the implied causal hypothesis can be sustained; a negative result decreases that probability, eliminating it completely in the case of a hypothesis of necessity.

In view of the above principle, tests of associative hypotheses are very useful in the early stages of research, not only to narrow down the independent variables that are to become the foci of further investigation, but also tentatively to identify or eliminate other factors that may have to be controlled in order to establish the hypothesis in question.

We had just identified one such factor in relation to the first step in Goldberger's investigation; namely, we pointed out that the observed association between diet and pellagra did not rule out the possibility of contagion.

Let us now eliminate that possibility and see where we stand then. Suppose Goldberger had done both the first step and the last; that is, he

1. Showed an association between diet and pellagra, with all cases of pellagra eating a deficient diet,
2. Showed by his experiment that pellagra was not contagious.

Would he have proved his hypothesis?

Remember: a hypothesis is not proved until no alternative hypothesis can be given to explain the observed findings.

Can an alternative explanation be offered?

Who were the people who ate the deficient diet? Poor cotton workers. What about the following possibilities:

1. Contact with some poison associated with processing cotton.
2. Excessive exposure to sunlight.
3. Exposure to natural radiation.
4. Exposure to extremes of temperatures.

Notice that in the groups that Goldberger studied these are factors that were associated with eating a deficient diet—that is, with the independent variable of the hypothesis.

Does this mean that you have to rule out any factor that happens to be associated with the independent variable of the hypothesis?

What about the evil eye? Some people think you can get sick from being given the evil eye. And poor people are more likely to have such beliefs than the well-to-do and well educated.

Is it necessary to rule out the evil eye as a possible influence?

No. Because there is no scientific basis for believing that an evil eye could have such an effect.

So what alternative hypotheses do you have to rule out? Any and all? When must a variable be considered as a potentially confounding factor?

Principle VI

In testing a hypothesis, a variable must be considered as a potentially confounding factor when it is associated with the independent variable of the hypothesis, and scientific ground exists for believing that the variable in question could influence the dependent variable.

Poison encountered either in the course of cotton processing or—more important—simply in the diet eaten by poor families, would satisfy both of the above conditions. And in Goldberger's time, perhaps so could exposure to dampness, sunlight, or extremes of temperature.

How would one rule out the possibility that such factors could explain the observed results?

Would one have to disprove each of them the way Goldberger did with contagion?

That would have been an awful lot of work. Did Goldberger do all that?

No, he didn't.

Then Goldberger didn't prove his case after all? We have all been taken in—and so have all the scientists—then and today.

What did Goldberger do that eliminated all of these alternative hypotheses—and some others as well—at one fell swoop?

He supplemented the diet of pellagra cases with animal proteins and the pellagra disappeared.

Notice that this procedure immediately eliminates the possibility that poison, climate, sleep, etc. could influence the results. Since the same individuals are involved throughout, and they remain in the same environment, all of the above variables—and a host of others—are held constant. Only one factor is allowed to vary —the independent variable of the hypothesis—change in diet. If, under the circumstances, the dependent variable then also changes, the change can be attributable only to the independent variable.

We have just illustrated the fundamental principle for verifying causal hypotheses.

Principle VII

To establish a cause-and-effect relationship, one demonstrates that variable x has an effect on variable y by allowing x to vary, holding constant other sources of variation for y, and then showing that y varies in a specified fashion as a function of the variation in x.

There are a number of different methods for varying x while holding constant other sources of variation for y; these are discussed in textbooks on experimental design. One more step remains to complete our examination of principles of verification as illustrated by Goldberger's classic investigation.

Suppose Goldberger had carried out only the second and the last steps of his research. That is, he had shown that

Step 2. Supplementing the diet of pellagra cases with animal proteins made the pellagra disappear.

Step 4. Pellagra could not be contracted through contagion.

Are these two steps enough to prove that deficient diet is a necessary condition for pellagra?

To prove a hypothesis of necessity you have to show two things:

1. that without the independent variable (diet), the effect cannot take place,

Has this been shown? Yes.

2. and that no other independent variable can explain the obtained results.

Does any alternative explanation remain? Note that contagion was eliminated by the experiment in Step 4. Such things as amount of sleep, exposure to sunlight, natural radiation, etc. are held constant by having the same individuals be sick and well in the same environment.

In other words, the combination of Step 2 and Step 4 has established the hypothesis of necessity. What about the hypothesis of sufficiency. Do Steps 2 and 4 confirm that as well?

To prove a hypothesis of sufficiency, one has to demonstrate that a change in the independent variable A produces a given effect B, and that no other independent variable can explain the obtained results. Thus far we have shown only that diet A is a necessary condition for producing pellagra and that a normal diet is sufficient to

cure pellagra. What we still have to do is prove that diet A is sufficient to produce pellagra. How to do it?—by means of Goldberg's experiment producing pellagra in convict volunteers. Other variables were controlled by feeding a comparable group in the same institution on a normal diet. Here change in A produces the effect in question. And no alternative hypothesis remains.

Quod erat demonstrandum!

Taken together, the results of Goldberger's three experiments "prove" his hypothesis. By producing pellagra through the removal of animal protein from the diet, by curing pellagra through adding this same ingredient to the diet, and by showing that pellagra was not transmitted through contagion, Goldberger established that the necessary and sufficient condition for pellagra is absence of animal proteins in the diet. The cause of the disease—what medical scientists call its etiology—was now fully known.

The Problem of Levels of Analysis

But was it really? Perhaps a medical scientist would be satisfied, but what about a chemist? After all, subsequent investigation showed that pellagra was actually caused not by protein deficiency in general but by the absence in the diet of a specific chemical substance known as nicotinic acid. And if biological hypotheses have to be reduced to chemical ones, what about psychological explanations, or sociological ones—must each be reduced ultimately to the level of charged particles?

Clearly not. The scientist is free to choose the level of analysis at which he wishes to work. His independent and dependent variables may be at the same or at different levels. The only restriction upon him is that he may not claim conclusions beyond the levels at which he has worked, although his results may suggest new problems and hypotheses at these other levels.

Thus we shall find students of human behavior working at different levels of analysis. Some attempt to relate the behavior of the individual to its physiological and even chemical substrata. Others seek to explain the actions and attitudes of one person as a function of the behavior of others acting individually or in concert as groups, communities, and societies.

But at whatever the level the causal hypothesis is couched, the logic of proof remains the same. We may summarize this logic by offering a paradox: *the process of proof is actually one of disproof.* The scientist never really demonstrates that a hypothesis is true; what he does is to eliminate all other possible explanations. In sum, scientific truth is established by default.

III. THE MEASUREMENT OF VARIABLES

We have examined the *logic* of science—the principles involved in formulating and testing hypotheses. What about the method? How does one translate the logic into actual research operations?

Operational Definition

Since variables constitute the building blocks of every hypothesis, the first step calls for expressing these variables in some concrete form, in terms of some operation which enables the investigator to determine a change, or lack of change, in the variable in question. Thus, Goldberger had to have some way of knowing when the subjects in his experiments developed pellagra and when they were cured of this affliction. For this purpose, he used as an index the symptoms of the disease, primarily the appearance of the characteristic rash for which the condition was named. In other words, the measure of the dependent variable in his hypothesis was simply the diagnosis made by a physician as to whether pellagra was present or absent. Similarly, Goldberger's independent variable was defined by presence or absence of animal proteins in the diet fed to his research subjects.

Of course, we are often interested in intermediate points between complete absence and full development of a variable. For example, as his measure of the relative degree of congenital malformation in a given area, Gentry used the rate of such cases per thousand births as determined from entries in the infants' birth and death certificates of abnormalities observed by the attending physician.

In science, the procedure one employs to determine the degree to which a particular variable is present is called *operational definition*. It is also referred to as *indexing* or just plain *measuring*.

Scales of Measurement

As the foregoing examples illustrate, measuring always involves specification of the category into which a particular observation falls. Categories may differ from each other in one of two ways, in quality or in quantity. Examples of the former are classifications of disease, nationality, occupation, or sex. When a system of classification is based on qualitative distinctions without any implication of order among the categories, we refer to such a system as a *nominal scale*. In a nominal scale, the classes differ by name and not by number; that is, they do not fall into any fixed sequence.

Where the categories do fall into a regular order, but the interval between steps is not fixed in terms of a constant unit of measurement, we speak of an *ordinal scale*. The most simple example of an ordinal scale is a ranking. Ranks, however, have the disadvantage that their significance depends on the number of persons ranked. Thus the student ranking 10th in a class of 10 is clearly not comparable to one who ranks 10th in a class of 100. For this reason, distributions of ordinal position are often broken up into divisions with an equal number of cases in each division. Thus one speaks of a measurement in the top *quartile* (upper fourth of all the cases), *decile* (tenth) or *percentile* (hundredth). All of these are examples of ordinal scales.

The chief limitation of ordinal measurements is that the distance between successive ordinal positions may not be equal (e.g., the difference between the tallest and the second tallest may not be the same as that between the second tallest and

third tallest). It is of course much more convenient when the units of measurement are stable. When this condition is satisfied we have what is called an *interval scale,* illustrated by the common thermometer. Notice that the location of the zero point on such a scale is usually quite arbitrary; on an ordinary thermometer, zero does not mean the absence of any temperature at all. It is for this reason that one cannot say that a temperature of 40° C is twice as hot as 20° C. To be able to make such proportional statements, it is necessary to have an absolute zero point, as in measurements of weight, time, and distance. When an interval scale has this property it is referred to as a *ratio scale.* The scale of cardinal numbers—the one we use to count a series of objects—is of course a ratio scale.

Except for their use in counting people or frequencies of an event, ratio scales are a rarity in the behavioral sciences, since it is difficult to establish an absolute zero point for psychological characteristics. It is usually possible, however, to construct interval scales for most aspects of human behavior. This is fortunate, since the interval scale has many advantages for scientific work. In particular, because of its equally spaced intervals, it permits the calculation of stable indices which summarize the characteristics of a whole series of measurements.

But before considering how and for what purpose measurements themselves can be manipulated and summarized, we must confront a prior problem regarding their basic soundness.

The Problem of Validity

Whenever one undertakes to translate theoretical variables into concrete indices, one must take into account an omnipresent danger—the danger of mistranslation. The index may not be *measuring what it is supposed to measure.* In scientific terminology, it may not be *valid.* For example, in measuring brain waves, the meter may be plugged into the wrong circuit, with the result that what is being recorded is not the perturbations of electric current in the brain but in the overhead light fixture.

The foregoing example may not be so outlandish as it may seem. In the age of computers, the possibility that one set of data has been substituted for another is hardly negligible. Take the experience of one of the writers. A dozen years ago, in a Presidential address (Bronfenbrenner, 1958), I reported some fascinating findings on the effects of different types of parental treatment on the behavior of the child. The results were rather complex, some seemingly contradictory, but I managed to show how they all really fitted together into a single theory. It was rather impressive.

Unfortunately, at the next convention of the American Psychological Association a year later I had to present the same material again to much the same audience (Bronfenbrenner, 1959). I was able to assure my listeners, however, that no one would be bored. You see, through a misunderstanding at the computing center, the signs on all the computations had been reversed, so every relationship that I had reported a year before was in fact exactly backwards. In my second address, I had to set everything right side up. And of course, I came up with a new theory that fitted the "new" results, but somehow it was not so impressive any more.

Here, then, was an instance in which the measurements, as they were being used, were completely *invalid*. The more typical case, however, is that of partial validity—the index reflects the variable in question plus other factors as well, which may account for much if not most of the variation. For example, years ago, before the French psychologists Binet and Simon (1905) invented an "objective method" for testing functional intelligence, the evaluation of a child's mental ability was made simply by asking some adult who knew him—usually the teacher—to make a judgment. We know now that although teachers' judgments show a positive relation to more objective measures of the child's intellectual performance, they are also influenced by other factors such as the child's social class level, how he behaves in school, or—perhaps most importantly—the degree to which the teacher likes him. To the extent that these other factors affect the teacher's judgment, her evaluation of the child's mental ability is an invalid index.

Of course, so-called objective measures—such as paper-and-pencil tests with predetermined scoring schemes—also present problems of validity. For example, group tests that purport to measure intelligence, or achievement in a particular subject like history, biology, or even mathematics, may actually be measuring little more than speed of reading. Another source of confounding arises with paper-and-pencil tests of personality, for here the respondent often gives not the real answer but what he thinks he ought to say—the socially desirable response.

There are various ways for getting around such difficulties more or less satisfactorily, but these are technical problems which need not concern us at the moment. The point we wish to make here is that the investigator must always consider the issue of validity with respect to each of the variables included in his investigation and provide some evidence that the procedures he is using—his operational definitions—do in fact measure what they are supposed to measure.

Validation Against an Outside Criterion

How does one establish that his methods of measurement are valid? When the senior author was himself a student, this question was simply answered. To show that your technique was valid, you simply tested it against some *external criterion* presumed to measure the same variable. This outside criterion usually took the form of a judgment by a person or persons deemed to be experts on the phenomenon in question. For example, the Stanford-Binet, the best and most widely used individual test of intelligence,[2] was originally validated against teachers' judgments. In other words, to see whether the test was measuring what it was supposed to, Lewis M. Terman, the famous Stanford psychologist, examined the degree of correspondence between the results of the test and the ratings of each child's intelligence made by his classroom teacher. There was a positive relationship between the two sets of measures, but there were also some exceptions. Some pupils whom the teachers rated high in intelligence turned out low on the test and *vice versa.* Which measure was right? We know now that the Stanford-Binet is usually a more valid measure of intellectual performance than a teacher's rating, but this could not be determined

simply from the degree of association between two measures where the validity of each was in question.

The foregoing consideration points to the principal limitation of the *outside criterion* method for evaluating validity; namely, the method is limited by the *validity* of the outside criterion. This approach, therefore, is most applicable in those situations where a valid criterion already exists. But then why develop a new method? Actually, there may be very good reason to do so. The existing valid method may be excellent but expensive of time and resources. The Stanford-Binet is a case in point. It can only be given to one child at a time, requires a trained examiner, and takes anywhere from thirty minutes to an hour and a half to administer. In contrast, group tests of intelligence can be given to an entire classroom within a specified period of time by persons without a high degree of specialized training. Such group tests are invariably validated against the Stanford-Binet as an external criterion.

But what if the outside criterion is itself of inadequate validity, or at least of poorer validity than is desired of the new instrument? Does testing against the outside criterion then have any utility at all? Clearly, yes, provided this criterion is believed to have *some* validity, for, then, as in the case of teachers' ratings, it provides some reassurance that the investigator is on the right track. But equally clearly, additional evidence of validity is required, especially in those instances where no external criterion is available, as would occur when the variable was being measured for the first time. How can one establish the validity of a measuring technique without relying on some external index of the variable in question?

Construct Validity

The answer to the foregoing question is suggested by the following fact. The ability of the Stanford-Binet to predict school grades was also offered as evidence for its validity. The argument ran as follows: since intelligence is necessary for academic achievement, there should be a positive relation between measures of intelligence and measures of achievement. Since scores on the Stanford-Binet show such a positive association, this fact is *consistent with* the position that the Stanford-Binet does in fact measure intellectual capacity. Notice that, taken by itself, the existence of the expected relationship does not *prove* the validity of the measure of intelligence, it is merely *consistent* with the presence of such validity. But if one could identify a variety of such expected relationships, and if all of these expectations were in fact fulfilled, this would obviously increase the probability that the index was actually measuring what it was presumed to measure.

Here we have the guiding principle of the process known as *construct validation*. It may be stated somewhat more formally in the following terms: *a measure has construct validity if it shows a pattern of relationships with other variables that is to be expected from a theoretical analysis of the phenomenon.*

To make clear what is implied by this definition let us take an example of construct validation carried out by Richard Christie and his associates (Christie, 1964; Geis, Christie, and Nelson, 1963; Geis, 1964; Geis and Christie, 1965; Christie and

Geis, 1968; Christie and Geis, 1970) at Columbia University. These investigators posit the wide prevalence in contemporary American society of a personality trait which they call Machiavellianism. The origin and nature of this characteristic are described in the following excerpt.

> Since the publication of *The Prince* in 1532, the name of its author has come to designate the use of guile, deceit, and opportunism in interpersonal relations. These behaviors are usually conceived as accompanied by congruent perceptual and attitudinal personality dispositions, characteristically including a dispassionate readiness to expect and detect human weaknesses, failings, and foibles, and the willingness to exploit them. More generally, a Machiavellian is one who views and evaluates others impersonally and amorally in terms of their usefulness for his own purposes. The Machiavellian would thus appear to correspond to the ideal type—or stereotype—of the "operator" or "manipulator."
>
> Whatever the labels applied, the syndrome of impersonal, manipulative attitudes and behavior is socially significant. As society becomes more and more organized, increasing proportions of interpersonal contacts become impersonal and means-oriented. As major activity in all areas of society is increasingly conducted by organizations, the ability to influence, direct, and use others effectively becomes increasingly valuable. (Geis, *et al.,* 1963)

To measure the Machiavellian syndrome, Christie and his colleagues have developed a questionnaire of 20 items, of which the following are examples.

> *Never tell anyone the real reason you did something unless it is useful to do so.
> *The biggest difference between most criminals and other people is that the criminals are stupid enough to get caught.
> *Most men forget more easily the death of their father than the loss of their property.
> One should take action only when sure it is morally right. It is safest to assume that all people have a vicious streak and it will come out when they are given a chance.
> Barnum was wrong when he said that there's a sucker born every minute.

The respondent indicates his degree of agreement with each item on a seven point scale ranging from "strongly disagree" to "strongly agree." Starred items are scored in the opposite direction. The measure of Machiavellianism is the sum of the person's score across all 20 items.

How is the validity of such a scale to be established? Clearly an external criterion is hard to come by. This is particularly true of what might be thought of as the ideal validating index—obtaining Machiavelli's own response to the items (although some

would not deny the ultimate availability of this criterion to the originators of the scale!). The authors themselves have sought to demonstrate validity by testing and confirming a variety of hypotheses about differences between high and low scores. Here are some of them:

1. When given an opportunity to deceive others in an experimental situation, high scorers were much more active than low scorers in thinking up and carrying out activities which confused, annoyed, and frustrated the other person.
2. In a sample of Washington lobbyists, high scorers spent more time contacting and entertaining Congressmen than low scorers. The high scorers also had more clients.
3. In group discussions, high scorers were more persuasive than low scorers.
4. In a sample of Hungarian immigrants, high scorers adapted to the American way of life more quickly than low scorers.
5. In an experimental game requiring convincing a partner to join in a coalition at a loss to the partner, high scorers generally won, low scorers lost.
6. In a sample of medical students, high scorers were more likely to choose psychiatry over surgery as a specialty.
7. The more ambiguous the rules in an experimental game, the more likely high scorers were to win it.
8. Mach (short for Machiavelli) score tended to be unrelated to amount of education, socioeconomic status, or level of intelligence.
9. When high and low scorers played strategy games over a period of time, the high scorers tended to win in the early stages, but low scorers won in the final stages. The investigators interpreted these results as supporting the hypothesis that "honesty is the best policy—in the long run."
10. When shown slides of past contestants in the Miss Rheingold contest, high scorers were more successful in selecting the winners for each year than were low scorers.
11. Persons reaching 21 years of age after 1942 had higher Mach scores than those attaining their maturity before that date. In other words, the tendency toward Machiavellianism has increased since World War II.
12. High scores on the Mach scale were associated with a history of disrupted relationships in childhood (parents separated or divorced, many moves from one location to another.)
13. "Graduate students in social psychology are more in tune with Machiavelli than any other aggregate of subjects yet tested." (Christie 1964, p. 14)

Note the following characteristics of this set of findings.

A. The variables with which the Mach score shows positive (Hypotheses 1–8. 10–12), negative (Hypothesis 9), and no relationships (Hypothesis 8) are those for which such relationships would be expected, given a valid measure of Machiavellianism as theoretically defined.
B. The hypotheses involve Machiavellianism both as an independent (Hypotheses 1–8, 9–10), and as a dependent variable (Hypotheses 8, 11, 12).
C. Since the presence of a statistical association can serve as corroborative but not as conclusive evidence for verifying a hypothesis (see preceding section), the data submitted include not only evidence of appropriate statistical relations (Hypotheses 2, 4, 6, 7, 8, 10–13), but also relevant results of controlled experiments (Hypotheses 1, 3, 5, 7, 9).
D. Since a laboratory situation necessarily leaves out aspects of the "real world" in which behavior occurs, validating hypotheses involve events outside the laboratory (Hypotheses 2, 4, 6, 8, 10–13) as well as controlled experiments.

As we see from the foregoing analysis, construct validation involves demonstrating that the antecedents, consequents, and correlates of the variable under consideration are consistent with the presumed nature of that variable. In a sentence, *construct validation involves testing the theory associated with the construct*—the body of interrelated hypotheses involving the variable in question. Confirmation of this set of hypotheses constitutes evidence that the operational definition of the variable does represent what it is supposed to represent, that the variable in question does exist and can be measured. In short, construct validation is a method for confirming what we have called a single variable hypothesis.

Notice that the confirmation is accomplished by taking advantage of the hierarchical, nested structure of a theory and its component causal hypotheses. Specifically, where a set of causal hypotheses have in common a particular variable, either as an independent or dependent factor, then confirmation of these hypotheses is also a validation of their component elements, including the single hypothesis about the existence of the focal variable.

Does this mean that every operational definition must be validated in this comprehensive fashion? If so, this would mean that no measurement could be considered valid until a whole body of hypotheses involving that variable have been confirmed. Actually, this is necessary only in those instances where the operational definition is only remotely or indirectly related to the theoretical variable. Such a state of affairs is likely to occur in two kinds of situations. The first, illustrated by Christie's concept of Machiavellianism as a personality trait, involves a characteristic which is not accessible to direct observation but is inferred as a *hypothetical construct.* Hence the term "construct validity."

A second circumstance in which construct validation is indicated occurs when a theoretical variable could be measured directly, but for reasons of economy the

investigator makes use of some indirect index relatively removed from the original phenomenon but more quickly and cheaply obtained. A case in point is provided by Gentry's research (1959). Here the independent variable, natural radiation, could have been measured directly, through the use of portable Geiger counters, but the cost of carrying out a radiological survey for the entire state would have been prohibitive. In its place, Gentry ingeniously employed the far cheaper alternative of utilizing already available geological survey maps of rock formations. But since his index is indirect, he is under obligation to establish its validity. This he does by demonstrating a chain of relationships as follows:

1. He cites laboratory studies showing that igneous rocks exhibit higher radioactivity than sedimentary rocks.
2. For the few sections of the state where direct field studies of rate of natural radiation had been made, he shows that areas with higher rates are those containing greater concentration of igneous rocks (eg., the Adirondacks).
3. Finally, he shows that the rate of congenital malformation is highest among persons who live in areas with the greatest concentration of igneous rocks and whose way of life involves close contact with rocky soil (e.g., obtaining drinking water from wells and springs).

Notice that where an outside criterion exists, (as in this instance) construct validation includes testing the index against the outside criterion (i.e., laboratory measurement of radioactivity of rock specimens). But, as before, evidence is presented for the validity of the measure outside the laboratory as well. Finally, the ultimate validation of the construct is supplied by confirmation of the causal relationships involving that construct.

The Dangers of Face Validity

But what if the existence of the theoretical variable is not in doubt and its operational definition is fairly direct rather than inferential? For example, suppose the independent variable in our hypothesis is the sex of the child and the dependent variable "crying." Obviously, one does not need to confirm a complete theory of genetics to be sure that one child is a boy or another a girl. Nor does one need to demonstrate the causes or consequences of crying to be assured that Mary or Johnny is shedding tears. Under such circumstances, a variable is said to have *face validity;* that is, the validity of the operational definition is regarded as self-evident.

But even though the validity seems self-evident, it is good practice not to take it for granted but to consider possible sources of confounding. Suppose, for example, that in testing for a sex difference in susceptibility to crying among nursery school children, the dependent variable is measured by the amount of time per hour that each child is observed to be in tears. The problem with such an index becomes readily apparent when we examine the kinds of activities engaged in by boys and girls in the

nursery playroom or out of doors: the boys are at greater risk to physical injury with the result that they may actually cry more than the girls busily playing in the doll corner. In other words, for the measure of crying to be valid, one must control for the degree of instigation in the environment.

The foregoing example illustrates how the problem of validity becomes part of the more general problem of experimental design—that is, controlling for the influence of confounding variables. Such sources of confounding variables are readily overlooked when the face validity of an index is taken for granted. The index may indeed reflect the theoretical variable in question, but one or more extraneous variables as well.

Systematic and Variable Errors

To the extent that a measurement is only partially valid, it is said to be in error. There are two kinds of errors, *systematic* and *random*. A systematic error is one that deviates in one direction more than in another. For example, in a special validating study of the measure of his dependent variable, Gentry found that entries of congenital malformation in birth certificates actually underestimates the true rate of such defects in the population. This fact was established by visiting the family of every n*th* child born in a given region and obtaining more direct information about the presence or absence of congenital defect. The rate of congenital malformation compiled on this basis turned out to be three times as high as that obtained from looking at birth and death certificates. In other words, the information on certificates was often incomplete, producing a systematic underestimate.

The second type of error is illustrated by the following example. The following are estimates of John Jones' height by five of his classmates who observe him as he stands at the front of the class:

5'6", 5'8", 5'4", 4'11", 5'8"—Jones' height actually is 5'4½". This is very close to the average of the above estimates, which equals 5'5". The estimates are obviously in error, but not in any systematic way. The chance of an error in one direction is just about as great as in the other. Errors of this type are known as *variable* or *random* errors and they reflect the degree of precision of the measuring instrument.

It would be easy to increase this precision by using a measuring tape in place of the naked eye, but even so, some degree of variability in successive measurements of the same thing would remain.

Reliability

The extent to which a measuring procedure is free of such random variability, *the degree to which it yields stable or consistent results in measuring the same thing, is referred to as the reliability of the measure*. Reliability, then, is a special form of validity reflecting the extent to which a measuring technique is free of random variation or uncontrolled wobble. Notice that if an index is completely valid, it must also be completely reliable, since a totally valid measure must be completely free of error,

random as well as systematic. A reliable measure, however, is not necessarily valid. For example, at one time it was thought that head size was directly related to mental ability; the larger the head, the more intelligent the person. It is, of course, possible to obtain highly reliable measures of the circumference of the skull, precise to the fraction of a centimeter, but as indices of mental capacity, such exact measurements have virtually no validity.

Reliability is obviously a relative matter, one measuring procedure being more precise than another. For this reason, it is useful to have an index of the degree of reliability of a given measuring procedure. The most direct index would be some indication of the size of the random variations obtained when the same thing is measured several times. For example, in the case of the five estimates of John Jones' height, one could index the extent to which these estimates vary around their average or mean value of 5'5". For this purpose we need a measure of *dispersion or spread.* The most commonly used index of this type is called the *standard deviation,* designated by the Greek letter σ (sigma).

When applied to a distribution of errors in measurement, as is the case in assessing reliability, the index is referred to as a *standard error,* abbreviated as S. E. The standard deviation of a distribution, when extended on either side of the mean, will include about two-thirds of all the observations.

A standard deviation is always expressed in the same units as were employed in making the original measurement. Thus, the standard error in estimating heights would be expressed in inches, that of an intelligence test in I.Q. points, of an achievement test in grade levels, etc. Obviously, it would be desirable to be able to compare the relative reliability of different measuring instruments irrespective of what they were measuring, to be able to determine, for example, whether an intelligence scale is more or less reliable than a personality test being used in the same research. Such comparability becomes possible through the use of a statistic called a *correlation coefficient,* designated by the letter r, which measures the degree of association between two sets of measurements; for example, height and weight. The correlation coefficient varies in magnitude from –1.00 to + 1.00. A value of + 1.00 indicates a perfect and direct association between two variables. The higher the one, the higher the other, with perfect prediction between. A correlation of –1.00 also implies perfect predictability, but in an inverse relationship; as one variable gets bigger, the other gets smaller. A correlation of zero means no association predictability between the two variables. Most observed correlations range somewhere between these two extremes. For example, the correlation between height and weight is about .40.[3]

The correlation coefficient can also be used to measure the extent of correspondence between two sets of measures of the same variable; for example, two versions, or forms, of a test. When used for this purpose, the correlation coefficient is referred to as a *reliability coefficient,* for it measures the consistency of a given measuring instrument when applied more than once to the same set of phenomena. For example, the reliability of a test may be assessed by giving it twice to the same group. This is called "test-retest" reliability. Such a procedure of course, has the disadvantage that the person's responses the second time may be influenced by memory, thus

producing an artificial consistency between the results of the two administrations. To avoid this artifact, the same test may be divided into two parts (for example, odd-numbered items vs. even numbered) and observing the correspondence between them; this is called "split-half reliability." Individually administered measuring instruments, such as the Stanford-Binet, often have reliability coefficients in the .90's. Group administered techniques typically have lower reliabilities, ranging from .40 to .60.

The measurement of variables is a necessary step in the testing of a hypothesis, but not a sufficient one. The process of proof, which, as we have already seen, is actually a process of disproof, requires the use of an appropriate strategy of analysis, involving statistical methods and research design. These matters are discussed in the next section.

IV. RESEARCH DESIGN

Suppose we have a hypothesis we wish to verify. What do we do next?

Several years ago, the senior author and his colleagues conducted an experiment in a large introductory course, testing a series of hypotheses. We wanted to determine the effects on academic performance of three kinds of grading procedures—letter grades, pass-fail, and guaranteed pass. We were also interested in whether the students' participation in a discussion section influenced their learning. Finally, we wanted to see if attending lectures delivered "live" by the professor produced a different result from seeing the same lectures on videotape in a smaller classroom setting. Effects were examined in terms of performance on examinations (all students took the same tests regardless of grading condition), in academic and career choices, and in attitudes toward the subject matter and the lecturer.

Let us take one of the above variations to demonstrate how research methods are applied to verify a hypothesis, specifically, the hypothesis that students do better in courses in which they see the lecturer "live" rather than on a television screen.

Sampling

The first problem that arises is how to pick our sample. One possibility might be to ask for volunteers. At the first meeting of the class, we might ask who would be willing to view the lecturer on television, and let the remaining students stay in the lecture hall and watch the course "live." What would be wrong with this plan?

The difficulty, of course, is that the groups would probably not be comparable. Those who volunteered for television might be the students more eager to please the lecturer, or less willing to risk being called on. A similar problem occurs in many other research situations, for example in trying to measure the effect of group preschool programs, like Head Start. Many investigators have compared Head Start children

with non-Head Start children from the same neighborhood without regard to the possibility that the parents of the Head Start children might have been more motivated to enroll their children in the program, whereas parents in the control group were less interested or able to make the necessary arrangements. How can we get around this kind of problem? How would we make the two groups comparable?

One strategy is to ask for volunteers willing to participate in the program, and then assign them at random to either the experimental or control group. "Assigning at random" means forming the two groups in such a way that every individual has an equal chance of being chosen for one or the other. This can be accomplished by pulling names out of a hat, or, more effectively, by using a computer which can pull names out of a hat far faster than a human can. The names are fed into the computer, a distinctive number is assigned to each name, and then the computer is asked to select numbers at random either for one or the other experimental condition.

Suppose we had followed this procedure in our course experiment; that is, asked for volunteers from among the students in the class and then assigned these volunteers at random to the TV or non-TV treatments. Would this have been a good way to select our sample?

The answer to this question depends on the population to which we wish to generalize our findings. The nature of the problem is exemplified by a famous error in scientific method that occurred in the 1932 presidential election in which the candidates were Franklin D. Roosevelt and Herbert Hoover. Polls had just come into fashion, and a leading weekly magazine of the period, the *Literary Digest,* conducted a survey to predict the outcome. On the basis of its national polls, the magazine announced that Hoover would win by a substantial margin. On Election Day Roosevelt won by a landslide. As a result, the *Literary Digest* lost subscribers and eventually disappeared. What had gone wrong? The problem was in the selection of the people to be interviewed. All were subscribers to the magazine, and all were contacted by telephone. In retrospect, it became clear that most of the subscribers, especially those who in 1932 owned a telephone, were Republicans. In other words, the sample was biased. The error made by the *Literary Digest* pollsters highlights a fundamental requirement in any research design: *the sample must be representative of the population to which the researcher wishes to generalize.*

How does this requirement relate to our own example? What are the likely biases in asking for volunteers in an introductory course in child development? First, only those most interested in the subject are likely to volunteer. Also, past research has shown that firstborns and females are more likely to volunteer than those born later and males.

How do we solve this problem? How do we obtain a cross-section of university students? Ideally, we could take a random sample from the student directory. But how would we get all the students so selected to take the course, since it is not a university-wide requirement? What we did was to announce in the student daily that only those willing to be in the experiment could enroll in the course, and each participant would have to agree in advance to accept the experimental condition to

which he would be assigned at random. The result was the highest enrollment ever experienced in the history of the course, with representation from all college years and undergraduate divisions.

But even so, to what population can we generalize these results? American college students? Students at a particular university? In a particular course? With a particular professor? There are problems in generalizing even to the latter, because this was the first, and probably the only time that students enrolled in the class were also all volunteers for an experiment. So do we restrict our findings to "student volunteers in this course with this professor in this university"?

If so, and if the goal of science is to establish universal laws, on what grounds was the experiment worth doing at all? Let us consider the answer to this question in terms of one of our major research findings: students in the letter grade condition averaged 30 points higher on exams than students in the guaranteed pass condition, and 9 points more than students in the pass-fail condition. We certainly cannot generalize the results to all students in all courses. But the finding does increase the probability that the implied causal hypotheses will be sustained in similar situations. A skeptic is obligated to come up with a plausible counter-hypothesis explaining why the same relation would not obtain elsewhere. If no reasonable alternative hypothesis can be offered, it is only reasonable to assume that a similar result would be obtained with other students in other courses at other universities.

This means that, with one important qualification, the researcher does not have to test his hypothesis with every group in every kind of situation that he believes the hypothesis covers. It is a scientific contribution to demonstrate that the hypothesis is confirmed in one such group in one such situation, *provided there is nothing about the group or situation which would make the observed relation peculiar to that group or setting*—as occurred in the case of the *Literary Digest* poll.

Incidentally, how could the *Literary Digest* have avoided its fatal error? How does one obtain a representative sample of American voters? A random sample—in which every voter in the United States had an equal chance of being selected—would be too complicated and time-consuming to obtain. Instead, a random sample of census tracts—or other units—is more efficient. This type of sample is not completely satisfactory since some important element, for example, a large city, may be overlooked. To avoid this danger, the researcher can specify the proportions within his sample of known components of the universe to which he wishes to generalize. For example, he may determine ratios in the population by sex, age, race, urban-rural residence, city size, and then select randomly *within* these strata. Thus, one would draw random census tracts within a city, and select households and family members within them, adhering to desired sex, age, and race distributions. This strategy of selection is called *stratified cluster sampling*.

Making use of known characteristics of the sample increases the accuracy of estimation. For example, a stratified sample of 1200 voters can predict election results within one or two percentage points. In addition, one can get information about a critical component group of the stratified sample—for example, the black vote.

Most psychologists however, do not employ stratified random sampling or any other technique for assuming broad representativeness. Typically, the investigator uses what is called an "accidental" sample, composed of the people and situations at hand. He then tries to demonstrate that his hypothesis does obtain for these people and this situation, and leaves it to someone else to show that the relation does or does not hold true for other people somewhere else. As we have already indicated, such a limited research can be altogether legitimate and scientifically valuable provided there are no grounds for regarding the observed finding as a special or restricted case.

To return to the results of our experiment: can our finding that students receiving letter grades performed better on exams than students in other grading conditions be interpreted as a special case? For example, the students in our sample were taking all their other courses for a letter grade; perhaps those with a guaranteed pass in this course simply devoted more time and energy to earning better marks in the other courses they were taking.

Although we did not test this alternative hypothesis in our experiment, a professor at another university did so. In his research, students in the experimental group, took *all* their courses during one semester on a pass-fail basis (Gold, 1971). Their grades were not only lower during that term but also in the next one, when all courses were again marked on a letter-grade basis.[4] This example demonstrated another important fact about the process of research: science is the product of *a community of scholars.* Different people make different and often complementary contributions to the investigation of the same general problem. One researcher's design need not and cannot cover all the possibilities.

Usually the first question to be answered is not whether some relation is true for everyone everywhere but whether it is true at all. For example, in our research, only after we show that TV instruction is superior or inferior in one course, does the question arise whether the same result will obtain in other courses.

Methods of Control

In addition to sampling, a second major problem of experimental design is to devise valid measures for the independent and dependent variables. In our case, this was a fairly easy task. We assigned students at random to the "live" vs. TV condition and assessed course performance through six examinations, three of the essay and three of the multiple choice type.

But now we come to the third and crucial feature of the experimental design. We must be able to examine variation in the dependent variable under different degrees (or conditions) of the independent variable, *in such a way that the possible influence of other factors can be ruled out.* One way of achieving this objective is to employ two groups or treatments; the group in which the independent variable is present or maximal is called the *experimental* group. The group in which the independent variable is absent is called the *control* group. The methodology of experimental design can be extremely complex and ingenious. Consider, for example, some of the prob-

lems we encounter in our course experiment. If we are interested in the effect of TV on course performance, and, say, half of the students are in the TV condition and half not, for what variables would it be important to control? For one thing, it has been repeatedly demonstrated that, from the primary grades through college, women get better grades than men. This means that we need to make sure that the sex of the student is not the confounding variable. One way of accomplishing this is to place an equal number of males and females in each group, or in our case, in each discussion section. This technique is called a *balanced design*.

Actually, the number of women and men in each section need not be equal; it is necessary only that the proportion in each group be the same. Indeed, nowadays even the condition of proportionality can be dispensed with, since computers make it possible to correct for unequal numbers in the design and even for missing cases. Bear in mind, however, that if a design is not balanced or proportional, the researcher must make the appropriate corrections. Otherwise, his results will be confounded.

Sex of subject is only one of the many possible sources of confounding that are encountered in research with human beings. For example, in our experiment one might wish to control for such factors as the socioeconomic status of the student's family, his year in college, field of major study, previous courses in the same area, or the student's age, race, and ordinal position in the family. It would clearly be impossible to assign students to sections that were balanced or proportional with respect to all of these possible confounding variables. How are we to meet this practical problem? Fortunately, there is a method that simultaneously controls not only for all these variables but *all others* that might be a source of confounding, even those which might not have occurred to the investigator. This powerful strategy is one we have already discussed—random assignment. Such a procedure avoids systematic bias on any and all variables in the allocation of subjects to experimental or control groups.

Effective as it is, however, random assignment possesses two limitations. First, it is applicable only to experiments done by man—not those done by nature—for it requires that the experimenter be able to determine who goes into which group. For example, in studying the effects of upbringing in a kibbutz versus a conventional family, the researcher cannot assign families at random to one or another setting. He must take them as they come.

The methods of control applicable in this kind of situation are the same as those employed to deal with a second limitation of random assignment. Even when subjects are assigned to different conditions on a chance basis, the possibility of some bias is not eliminated. In our experiment, for example, one might by accident end up having more psychology majors in the "television" treatment than in the "live" condition. Because of this possibility of bias, it is essential not to leave control of the most important confounding variables purely to chance. Instead, we can resort to the already familiar procedure of classifying subjects in terms of these variables and controlling for them by means of a balanced proportional or other type of statistical design. If this is done, there are two critical conditions that must be met. First, *the same independent and control variables must appear in both the experimental and*

control groups, and in the same combinations. For example, if we have boys and girls in the TV condition, we must also have boys and girls in the "live" condition. The same requirement applies for the various combinations of sex and grading systems. For example, if there were girls receiving a letter grade in the "live" condition but not in TV, this could produce a spurious result for the experiment as a whole.

The second requirement is that *there be at least two cases in each "cell" of the design—that is, at least two cases representing every combination.* This requirement is necessary in order to be able to carry out certain crucial statistical analyses to be described in the next section.

When these two conditions are met, it is possible to analyze the independent effect of each independent variable controlling for all other variables. This very powerful type of experimental design is called *analysis of variance.* Variance is the variation of scores about the average. In analysis of variance, the total variation is divided into components, and each component is evaluated for its independent contribution. For example, in our experiment, TV vs. live, grading condition, and sex were each evaluated for their effect on course performance—independent of the others. Moreover, in this technique we can examine not just the independent effects of each variable, but the variables in combination. For example, we can ask whether the effect of TV was the same for boys and girls. If not—if, say, the boys did better with TV whereas the girls scored higher in the live lecture—we would have what is called an *interaction effect,* a joint effect of two or more variables which differs from that of any of the variables taken separately.

What happens when the available data do not meet the strict requirements of an analysis of variance design? For example, it may be impossible, or just too complicated, to find cases at every level and combination of several independent and control variables. Under these circumstances, the correct strategy is often simply a reduction in scale. For instance, instead of working with both sexes at each of three socioeconomic levels, the investigator limits himself to one sex and one socioeconomic level only (e.g. just middle class girls). If this is done, the researcher must bear in mind that his findings are limited and may obtain only for the restricted population he has sampled.

The procedure of controlling for a variable by having it appear at only one level represents one form of a more general technique for equaling experimental and control groups, called *matching.* Matching can be applied not only with qualitative categories, like sex or ethnicity, but also with variables that can fall along a continuous scale, such as age or intelligence quotient. For example, age or intelligence can be controlled by insuring that both experimental and control groups show the same averages on each of these variables. Such a procedure is known as *group matching.*

Matching can also be done at the level of individual cases. For example, each member of the experimental group is compared to a partner from the control group who is of the same sex, age, social class, IQ, etc. This procedure, called *pair matching,* allows for precise control on several variables, but it has a number of disadvantages. First, in pair matching, the several control variables are confounded so that it is impossible to assess their independent effects and interactions. Second, it may

become very difficult to find pairs that match on several variables at once, so that most of the cases remain unused. Finally, the matched sample one ends up with may be so restricted that it is no longer representative of the population to which one wishes to generalize.

Finally, there are some situations to which analysis of variance cannot be applied without distorting or eliminating valuable information. For instance, in all the situations we have considered thus far, the independent variable, and most of the control variables, have been measured in terms of mutually exclusive, qualitative categories (boys vs. girls, letter grade vs. pass-fail, etc.), whereas the dependent variable (e.g. scores on exams) has been quantitative and continuous (i.e., the scores go from low to high). Suppose that the independent and control variables are also quantitative and continuous: for instance, we wish to know whether attendance in the course is related to course grade, or weight at birth is related to a child's psychological development.

A second powerful method can handle this situation. It is one we have already encountered in the form of the correlation coefficient (r) which measures the degree of association between two continuous variables. If two variables are correlated, it means that a researcher can predict one from the other as a linear relation ($y = bx + c$). For example, we can predict the weight of the baby from the mother's food intake. We can also look at the effect of two variables on a third—for example, the mother's weight and food intake as both affect the baby's weight. This method is called *multiple correlation*. Finally, we can look at the relation between two variables controlling for the third—for example, the effect of food intake on baby's weight controlling for the mother's weight. This is called *partial correlation*. This general method is known as correlational analysis, and also as regression analysis.

There is one final situation we need to consider. What does one do when the independent and control variables are mixed and include *both* continuous and discontinuous components; for example, we may wish to examine the effect both of television and of attendance controlling for sex and for age. Under such circumstances one can resort to a procedure called *analysis of covariance,* which combines the strategies of correlation and analysis of variance within the same design.

There are many other statistical designs for achieving control of variables, and they can become rather complex. In fact, there is an increasing danger that methods can become so complex that they depart too far from reality—especially with the availability of computers. Another unfortunate consequence of sophisticated methodology is that researchers can often overlook more effective and direct designs in favor of the complicated ones. It is for this reason that, in our brief description of different strategies in research design, we have emphasized the underlying logic and not the statistical procedures involved. For information on the latter score, the reader can consult readily available textbooks on statistical methods. It is important to bear in mind that the choice of design should depend primarily not on statistical but logical considerations—the nature of the questions being asked, of possibly confounding variables, and of the subjects and situations to which the investigator wishes to generalize.

Testing the Null Hypothesis

One final task remains. Most alternative hypotheses are specific to a particular problem. But there is one alternative hypothesis that applies to every research design, and must be eliminated before the major hypothesis can be accepted.

To take a simple example: suppose we are interested in whether there is any difference in the heights of men and women in a given class. We pick a random sample from the class of one male and one female and measure the difference in heights. The problem is obvious: when the sample is small, the obtained result may differ by chance, to an appreciable degree, from the situation that obtains in the parent population as a whole, in this case the entire class. In such circumstances the observed difference between the small experimental and control samples may be larger or smaller than it really is; it may show no difference when one actually exists; or it can even reverse the direction of the true difference.

Note, however, that as we add more cases to our sample, we get a more stable, more reliable estimate of the average height of boys and girls in the class. If we were to look at the variation of height *within* sexes (from one girl to the next) and that between sexes (boys vs. girls) we would find that the latter was greater than the former and the ratio of the two will be greater than one. In contrast, if there is *no* real difference in height between males and females, the variation between sexes would equal the variation within sexes, and the ratio between the two would on the average, approach unity—sometimes a little less than one, sometimes more as a function of chance fluctuation.

Statisticians have calculated the probability of obtaining a ratio greater than one by chance for samples of different sizes. If this probability is less than 5 percent (or sometimes 1 percent), the researcher concludes that there *is* a difference between the means.

Analysis of variance uses this principle. The experimenter calculates the variance within an experimental group and then the variance among experimental groups. The ratio of the latter to the former is called an F ratio, for Sir Ronald Fisher, an English statistician who developed the method. If we get a ratio of, say, 3.5, we can consult a table of probabilities of F (available in most statistics texts) to see whether such a ratio could have occurred by chance. If not, we have what is called a *statistically significant difference,* one that is not very likely to have occurred simply by chance.

Now we can answer the question as to why analysis of variance requires more than one observation within each cell. Such repeated observations are needed to provide an estimate of *within* group variation against which to test the variation among groups in order to determine whether the latter is significantly larger.

An F test can be applied to any number of means—for example, three grading conditions. If only two groups are involved, it is possible to test the observed difference in means without computing their variance. Such a procedure is known as a *t* test. The logic underlying the *t* and F tests is exactly the same. In fact, in the case of two groups, $F = t^2$. A test of statistical significance can be applied to any measure of association between two variables, such as the correlation coefficient (r).

Another example arises in assessing differences between proportions, when the variable of interest is measured in mutually exclusive categories. For example, we may wish to know whether the incidence of congenital malformation is greater among boys than girls. Under these circumstances we use a technique called *chi square* (symbolized by the Greek letter χ^2), which tells us whether the observed difference between two proportions is greater than what could have occurred by chance.

We can now state the universal alternative hypothesis that must always be tested. We must be able to show that the obtained differences are not due simply to chance. The alternative hypothesis is called the *null hypothesis:* it stipulates that *no* true difference exists, that such variation as we see can be attributed to chance fluctuation. The null hypothesis applies to all hypotheses —not only causal, but also associative and even attributive. We must always consider the possibility that observed differences are a function of random variation.

To illustrate the application of the null hypothesis to a hypothesis of association, we may consider an observation by Pasamanick (1958) that a larger proportion of mentally retarded children in Ohio institutions were born in January and February than in other months of the year. He rejected the null hypothesis in his research—the relationship was not due merely to chance. Rather, he showed that it was related to mean temperature during the summer, when the mother was in the third month of pregnancy.

This example illustrates the point that in science there are few purely associative hypotheses. A causal hypothesis, or at least a search for a causal hypothesis, is usually implied. Variable X, or some variable X_1 associated with variable X, is presumed to be a necessary or sufficient condition for variable Y. In other words, there is almost always an implicit independent variable and the associative hypothesis is a test of a potential causal hypothesis. If the association is not explainable as a chance phenomenon—i.e., if the null hypothesis is rejected, then the original hypothesis is supported, *provided* of course that no other plausible alternative hypothesis remains.

Note that the demonstration of statistical significance always implies some degree of reliability. If men are significantly taller than women, this implies that the difference cannot be due to chance; hence measurements must contain more than error.

In summary:

1. Every causal hypothesis implies one or more associative hypotheses. For example, Madigan's thesis (1957) that women are biologically hardier than men, led him to the associative hypothesis that they would outlive men under constant environments.
2. Every associative hypothesis is challenged by a null hypothesis. Thus Madigan had to reject the possibility that the observed

differential mortality rates between the two sexes could be due to chance.

3. If the null hypothesis is rejected, then the original causal hypothesis is supported, *provided there exists no plausible alternative explanation for the observed relation.* For example, Madigan's finding that nuns outlive monks living under virtually identical conditions of monastic life leaves little room for other explanations besides biological sex differences.

One final caution: failure to demonstrate a statistically significant relation does not necessarily mean that no difference exists—that is, that the means are equal. Suppose when we chose class members, to determine if men are taller than women, we happened to pick persons of each sex who, on the average, were of similar height. The difference would not be statistically significant, even though, in the general population, a difference in fact exists. In other words, failure to reject the null hypothesis does not mean the absence of association; it means only that we have not been able to show a significant effect, perhaps because the N (number of people in the sample) was too small. If, under these circumstances, we had concluded there was no difference in heights, we would have made an error. There are, then, two types of error inherent in experimental designs. A so-called *Type I error* is the result of asserting a difference when none exists. A *Type II error* occurs in claiming no difference when a difference does in fact exist. In the case of our sampling of sex differences in height, we made a Type II error.

For some reason, researchers on human behavior have been much more concerned about Type I than Type II errors. They are more worried about being caught wrong than about failing to state the right. Scientists, for the most part, are conservatives, at least those who have engaged in the study of human behavior. This trend is reflected in the studies which follow, although there are some in which even the possibility of a Type I error has been overlooked. The reader is therefore well advised to be on his guard, or, still better, to do his own thinking, which is the essence of scientific inquiry.

NOTES

[1]The account which follows is condensed from R. P. Parsons, Joseph Goldberger and Pellagra, in *Trail to Light* (New York: Bobbs-Merrill, 1943).

[2]In recent years, serious questions have been raised about the validity of this instrument when applied to persons from a different social and cultural background from that of the predominantly white, middle class samples for whom the test was originally developed.

[3]The correlation coefficient should not be interpreted as a percent. It is simply a number varying between – 1.00 and + 1.00.

[4]Still, one might argue that taking courses of pass-fail grades within a university which otherwise assigns grades for courses may be quite different from taking all courses for pass-fail when all other students are evaluated in the same way. This particular case, so far as the authors know, has not been investigated.

REFERENCES

Bechtold, H. P. Construct validity: A critique. *American Psychologist,* 1959, *5,* (14), 619–629.

Berger, H. Ueber das Elektroenkephalogramm des Menschens, *Archiv fur Psychologie and Neurologie,* 1930, 40, 160–179.

Binet, A., & Simon, T. Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *Annee Pschologique,* 1905, 191–244.

Bronfenbrenner, U. Family structure and development. Presidential Address to the Division of Development Psychology, September, 1958.

Bronfenbrenner, U. Socialization and social class through time and space. In E. E. Maccoby: T. M. Newcomb, and E. L. Hartley (Eds.) *Readings in social psychology.* New York: Henry Holt and Co., 1958. Pp. 400–425.

Bronfenbrenner, U. Parental behavior and adolescent responsibility: a reorientation. Paper presented at the annual meeting of the American Psychological Association, September, 1959.

Christie, R. The prevalence of Machiavellian orientations. Paper presented at the annual meeting of the American Psychological Association, September 7, 1964.

Christie, R., Geis, F. Some consequences of taking Machiavelli seriously. In Borgatta, E. F. and W. W. Lambert, (Eds.), *Handbook of Personality Theory and Research.* Chicago: Rand McNally, 1968, 959–973.

Christie, R., & Geis, F. *Studies in Machiavellianism.* New York: Academic Press, 1970.

Cronbach, L. J. and Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin,* 1965, 52, 281–302.

Geis, F. Machiavellianism and the manipulation of one's fellow man. Paper presented at the annual meeting of the American Psychological Association, Los Angeles, 1964.

Geis, F., Christie, R. Machiavellianism and the tactics of manipulation. Paper presented at the annual meeting of the American Psychological Association, Chicago, 1965.

Geis, F., Christie, R., & Nelson, C. Some Machiavellian manipulations. Unpublished paper, Department of Psychology, Columbia University, 1963.

Gentry, J. P. An epidemiological study of congenital malformations in New York State. *American Journal of Public Health,* 1959, *49,* (4), 1–22.

Gerden, E. A review of psychokinesis. *Psychol. Bull.,* 1962, *59,* 353–388.

Gold, R. C., Reilly, A., Silberman, R., & Lehr, R. Academic achievement declines under pass-fail grading. *The Journal of Experimental Education,* 1971, 17–21.

Grollman, A. *Functional pathology of disease.* (2nd ed.) New York: McGraw-Hill, 1963, 172–176.
This is an account of present day knowledge about pellagra and the chemistry of its effects.

Heider, F. *The psychology of interpersonal relations.* New York: John Wiley, 1958.

Homans, G. C. *The human group.* New York: Harcourt, Brace, 1950.

Knoblach, H., & Pasamanick, B. Seasonal variation in the births of the mentally deficient. *American Journal of Public Health,* 1958, *48,* 1201–1208.

Madigan, Francis C. Are sex mortality differentials biologically caused? *Milbank Memorial Fund Quarterly,* 1957 (April), 35, *2,* 202–223.

Mendelyev, D. I. *The principles of chemistry.* Translated from the Russian (5th ed.) by George Kamensky. New York: Longmans, Green, 1891.

Newcomb, F. An approach to the study of communicative acts. *Psychological Review,* 1953, *60,* 393–404.

Parsons, R. P. Joseph Goldberger and Pellagra. In *Trail to light.* New York: Bobbs-Merrill, 1943. Also reprinted in Rapport, S., and H. Wright, *Great adventures in medicine.* New York: Dial Press, 1952, 586–604.

Rheingold, H. L., Gewirtz, J. L., & Ross, Helen W. Social conditioning of vocalizations in the infant. *J. Comp. Physiol. Psychol.,* 1959, *52,* 68–73.

Rhine, J. B. *Extrasensory perception.* Boston: Humphries, 1964.

Spock, B. *Baby and child care.* New York: Pocket Books, 1957.

Terris, Milton. *Goldberger on pellagra.* Baton Rouge: Louisiana State Press, 1964. This is the most recent and complete account of Goldberger's pursuit on pellagra including original reports by him and his critics.

Tjio, J. H., & Levan. The chromosome number of man. *Hereditas,* 1956, *42,* 1–6.

United States Children's Bureau. Infant care. (Rev. ed.) Washington: United States Government Printing Office, 1951.

part two
NATURE WITH NURTURE

The articles in this section all speak to the same vital principle: namely, heredity and environment never operate in isolation from each other. With respect to human development, neither factor can exert an influence without the other. This point is beautifully illustrated in the initial presentation by Freedman of a series of researches growing out of a study undertaken when he was still a student. What originally appeared as an elegant demonstration of the impact of different training methods on the development of "conscience" in puppies turned out to be mediated critically by genetic predispositions. The fact that analogous effects have been reported for man and attributed primarily to differential patterns of child rearing serves as a reminder that, for human young as well, patterns of indulgence vs. discipline work on a genetic base.

Along the same line, at a time when observed psychological differences between the sexes are being criticized as products of discriminatory child rearing practices, Madigan's ingenious demonstration of the persistence of sex differences in men and women exposed to highly similar ways of life raises the possibility of innate foundations for psychological differences between the sexes.

Finally, the intricate question of the interplay between genetic and social factors in the development of human ability is examined in three articles. It is in this area that the critical relation between science and social policy is most clearly apparent. Skodak and Skeels' pioneering study of the impact of true vs. foster mothers on the development of children adopted in the first six months of life presents a classic example of the difficulties of gaining acceptance for a new idea before its time. One of the first studies to demonstrate how an enriched environment can enhance intellectual capacity, this research ironically continues to be cited in support of the opposite conclusion.

A similarly unwarranted genetic bias, with even more serious social consequences, is analyzed and demolished by Scarr-Salapatek in her critique of the thesis of innate race differences in intelligence—a thesis promulgated in recent years by Jensen, Herrnstein, Eysenck, and others.

Finally, Bronfenbrenner takes on the major thesis of Jensen and others that intelligence is primarily determined by genetic factors; from a review of the evidence, he concludes that the argument for an 80 percent genetic effect is contradicted by the data, which in fact testify to the power of the environment in enabling the developing child to realize his genetic potential.

2.1

The Origins of Social Behavior

Daniel G. Freedman

The studies that form the basis of this article had their origin on the campus of Brandeis University, near Boston, Massachusetts, in 1953. In the winter of that year the campus mascot, a motley bitch with what looked like a beagle and dachshund background, gave birth to a litter of four. We were not sure who the father was. As graduate students in the newly formed graduate school Norbett Mintz and I had sufficient time to 'play around' and we had the happy notion that puppies might prove enjoyable experimental subjects—as indeed they did.

In considering what we might do with the puppies we had two previous studies in mind. The first was that of J. P. Scott and M. Marston at Bar Harbor who, in 1950, had suggested that three to eight weeks of age was a critical period in the social development of puppies. Although this was little more than an impression then, it had subsequently turned out to be largely correct within the breeds studied. Secondly, John Whiting, a psychoanalytically orientated anthropologist, had been working with adult dogs and had developed a test to measure their 'conscience'. This was assessed by the extent to which they remained obedient in the absence of the handler. The test was attractively simple and merely involved a person punishing a dog for eating food that had been forbidden him. The person then left the room and, watching through a one way glass, recorded the amount of time the animal stayed away from the forbidden food. In this way the extent to which punishment was 'incorporated' received a quantitative score.

Our chore became one of using this information in a meaningful study of development, and when the idea finally came to us it seemed the most natural study possible. As clinical psychologists, Mintz and I were aware of the work of the child psychiatrist, David Levy. As a result of his extensive experience with behaviour disorders in children, Levy had hypothesized that extreme permissiveness on the part of parents could lead to a psychiatric condition called psychopathy. This is characterized by an abnormal inability to inhibit one's impulses. These people are asocial in that their own desires and wishes always take precedence over those of others. Levy found that in many families which produced such individuals the parents allowed themselves to become complete slaves to the tyrannical wilfullness of the child.

With Levy's work in mind we carried out the following study. Two puppies were to act as 'control'. One was given to a family which we decided was typically 'middle class'—in other words, a home in which the puppy would receive plenty of affection, but would also be restricted to certain areas, prevented from biting, be housebroken and so on. A second puppy was raised separately in a room by itself, save for the few minutes each day it took to put food in and take the old bowls out. As for the 'psychopathic' puppies, Mintz and I decided to enlist the aid of the boys in the

Reprinted from *Science Journal*, 1967, *3* 69–73, by permission.

dormitories of which we were proctors. The puppies were to have complete freedom in the dormitory. Urine and faeces were to be cleaned up after them. If they wished to climb on someone's bed or lap, they were to be helped up. If they wished to get off, they were helped off. If they wished to sleep in someone's bed, they were allowed that. (If they nipped an ear in bed, one was allowed to duck under the covers.) They were not to be punished for any offense. Since the study was to last only six weeks (from three to nine weeks of age), all the students agreed to cooperate.

At nine weeks we had two distinctly different groups of dogs, and this was nicely shown by the 'incorporation of punishment' test. Beginning at the ninth week of age and continuing for eight consecutive days we administered the test devised by Whiting. We placed each pup, which had not been fed for at least four hours, in a room with a bowl of meat at its centre. When the pup attempted to eat, the experimenter hit it on the rump with a newspaper and shouted, "No". After the experimenter left the room, the time to eat was recorded. In this early version of the test the experimenter rushed back in and punished the pup each time it ate.

We found that whereas the home reared dog averaged 37.5 minutes and the isolate 9.5 minutes between each transgression, the permissively reared pups averaged a lightning 2.2 minutes. By the eighth day of testing, in fact, the permissively reared pups managed to eat all the meat before the tester had a chance to leave the room. He would smack the pup only to have it immediately circle back through his legs for another bite. This happened so fast that we could not measure the time between transgressions accurately and consequently we overestimated the time between transgressions in the seventh and eighth test sessions.

The subsequent history of these two pups was not a happy one. Although people were initially taken with them because of their uninhibited friskiness, they were passed from home to home as each owner found something else to complain about. They seemed to have become untrainable.

The isolate's behaviour was typical of pups raised in this fashion. She was hyperexcitable and initiated contacts only to run off when a hand was reached out or when another pup tried to play with her. After the fourth session of testing, however, she calmed down somewhat and was able to learn what was demanded of her. We are happy to report that she subsequently became a beloved and loving pet, and we could not help but reflect that a dog with no experiences with people may be preferable to one with the wrong experiences.

Having become completely captured by the study, we decided to try another experiment. David Levy had postulated that a second class of psychopaths are formed by extreme cruelty or great emotional deprivation in childhood. With this second hypothesis in mind, we raised a beagle from four to nine weeks of age under conditions in which all his contacts with people were negative. Each time he tried to make contact he was either ignored or pushed aside, and it is not surprising that he developed a rather depressed, fearful personality. True to Levy's hypothesis, his performance on the test was like that of the overindulged puppies, although the average time between transgressions was higher: 8.14 minutes. Again the controls were a home reared littermate and a littermate kept in isolation. Both quickly learned

to stay away from the food and averaged one transgression each 31.13 minutes. Thus once again Levy's hypothesis, developed from studies of problem children, appeared to hold when applied to puppies.

It was clearly time to do these studies on a larger scale and in a systematic and repeatable fashion, and I proposed such a project for my PhD thesis. After interesting Paul Scott, Chairman of the Division of Behaviour Studies at the Jackson Laboratories, Bar Harbor, Maine, it was arranged that the work be done there.

In the Bar Harbor study, eight litters of four pups each were used. They included two litters each of Shetland sheepdogs, basenjis, wire-haired fox terriers, and beagles. Following weaning at three weeks of age, each litter of four was divided into two pairs which were equated as closely as possible on the basis of sex, weight, activity, vocalizations, maturation of eyes and ears, and reactivity to a startling stimulus. Thereafter, each member of one pair was indulged and each member of the second pair was disciplined. However, because of the numbers involved, this treatment could be given only during two daily 15 minute periods, again from the third to the eighth week of age.

Indulgence consisted of encouraging a pup in any activity it initiated, such as play, aggression and climbing on the supine handler. As before, these pups were never punished. By contrast, the disciplined pups were at first restrained in the experimenter's lap and were later taught to sit, to stay, and to come upon command. When still older they were trained to follow on the leash. I handled all the pups and tested them individually; they lived with the identically treated littermate in isolation boxes the remainder of the time.

A revised punishment test was initiated at eight weeks of age. As before, when the pup ate meat from a bowl he was punished with a swat on the rump and a shout of,"No!" After three minutes the experimenter left the room and, observing through a one way glass, recorded the time that elapsed before the pup again ate. This time the experimenter did not return to the room until the allotted ten minutes were up.

The first breed that went through this experimental rearing and testing was a basenji litter. By the fourth day of testing all basenjis tended to eat soon after the experimenter left the room, the method of rearing having little effect. This was most discouraging, since we had failed to duplicate the results of the pilot experiment, but we decided to carry on.

The second group was a Shetland sheepdog litter and they were equally disappointing, except this time all tended to refuse the food. At least one thing was clear at this point: the breed of dog was a major factor to consider. Second litters of basenjis and of Shetland sheepdogs from our inbred stocks bore this out, for they performed much like the first litters.

The next breed which became available was a beagle litter. We found that their behaviour depended upon the way they were reared but, paradoxically, in a direction opposite to our pilot studies. A second beagle litter, from the same mating, performed exactly the same way and so did the two litters of wire-haired fox terriers.

The conditions of rearing were continued over a second period, when the pups were 11 to 15 weeks of age, and all tests were readministered with essentially the

same results. At this point we were so fascinated by breed differences that we postponed thinking about the contradictions to our pilot work. Our question now was how could the breed characteristics explain the differences in performance?

It was clear that, during training, beagles and wire-haired terriers were strongly orientated to the experimenter and sought contact with him continuously. Basenjis, by contrast, were interested in all phases of the environment and often ignored the experimenter in favour of inanimate objects. Shetland sheep dogs showed yet another pattern; all became fearful of physical contact with the experimenter and tended to maintain distance from him. Thus the two breeds that were highly attracted to the experimenter showed behavioural differences as a result of the mode of rearing, whereas the breeds that exhibited aloofness (basenjis) and excessive timidity (Shetland sheep dogs) did not. Apparently it was the strong constitutional attraction combined with indulgent treatment that enhanced the effectiveness of later punishment.

It should be noted that basenjis and Shetland sheep dogs were not entirely unaffected by the differential treatment. The scores of all indulged animals were significantly different from those of their disciplined counterparts on five of ten tests administered. In general, these tests indicated that the indulged pups were more active, more vocal, less timid (although more easily inhibited with punishment), and more attracted to people than the disciplined pups.

We now had to reconcile the findings of the pilot studies of Brandeis and the major study at the Bar Harbor. In the first we found that overindulged rearing in the dormitories led to 'psychopathic' performance, but in the Bar Harbor work the disciplined animals of all breeds tended to be more disobedient, although this was true of the Shetland sheep dogs and basenjis over only the first three days of testing.

To recapitulate the pilot studies, a regimen of affection and freedom and no discipline produced hyperactive, disobedient pups. A regimen of discipline and no affection likewise produced a disobedient pup, albeit within a depressed and fearful personality. Actually, none of the Bar Harbor findings contradict these 'rules'.

In the Bar Harbor studies the permissive group was given free reign to express affection, to investigate and to bite, but each time they were returned to their boxes they were under enforced control; as their cries to get out attested. On thinking it over, it became clear that they were treated much like normal home reared dogs in that restrictions were regularly imposed on their freedom. We therefore changed their official title from 'over indulged,' to 'indulged', although this was still not a complete description.

The disciplined group, on the other hand, received considerably less affection than do home reared dogs and, in addition, few dogs at home have so much demanded of them at so early an age. When we had initially planned the study we called this group 'normal', but very early we changed to the word 'disciplined'. It was clear that in terms of affection and play they were a *deprived* group and that 'disciplined' was only partially descriptive of the treatment they received.

It was in this way that we attempted to resolve the contradiction between the Bar Harbor results and the Brandeis pilot studies and, setting aside any genetic factors for the moment, the general hypothesis still appeared viable: dogs who did not