From Abstract Ideas to Concrete Instances

Some Guidelines for Choosing Natural Research Settings

PHOEBE C. ELLSWORTH *Yale University*

ABSTRACT: *This article addresses the general problem of the translation of abstract concepts and hypotheses into concrete instances that are likely to permit a valid empirical test. Considerations that should guide the choice of a research setting include (a) the history of research and the current state of knowledge in the area, (b) the types of variables to be studied, (c) the form of the hypothesized relationship among variables, and (d) the most plausible rival hypotheses. It is argued that research settings should be chosen with the whole experimental design in mind—that the availability of appropriate controls (groups or occasions for observation) should be as important as the suitability of the treatment group.*

Some people, including many social scientists, are most comfortable when negotiating in the realm of abstract ideas. They think in terms of conceptual variables and constructs, such as "attitude," "commitment," and "frustration," examining the research literature for relationships among variables and exceptions to these relationships, seeking logical consequences not previously considered, generating hypotheses, and arranging abstract concepts and relationships into theories intended to account for past and future observations. This kind of thinking is justifiably admired by scientists, and it is essential for the development of theory.

Those who think in this abstract manner tend to do research of the hypothesis-testing variety. Having generated predictions that provide the impetus for research, their next task is to find empirical realizations of their conceptual variables in order to test these predictions (Carlsmith, Ellsworth, & Aronson, 1976). The more general the concepts, the more concrete instances there are to choose from. And yet, within many of the broad provinces of social science, very few of the available paths are followed; the abstract ideas turn out, in practice, to encompass only a very narrow range of empirical realizations. The mental dexterity demonstrated in dealing with abstractions often seems to vanish at the translation stage, as the old standard treatments and measures are used and reused with little consideration of their suitability for the task at hand (Webb, Campbell, Schwartz, & Sechrest, 1966, chap. 1).

Of course, not all researchers think this way; many start with a concrete situation or behavior and study it (see Sidman, 1960, for an excellent defense of this less prestigious approach).[1] They do not face the problem of finding the concrete instance that best embodies the abstract idea, since they have started with the instance. The suggestions presented here are not intended for them. Other researchers have managed to excel at both levels. The dissonance theorists, for example, working in the service of an abstract set of hypotheses, have happily exploited the psychological laboratory, the small group of religious zealots (Festinger, Schachter, & Back, 1950), the discount house (Doob, Carlsmith, Freedman, Landauer, & Tom, 1969), and the race track (Knox & Inkster, 1968). But many social psychologists are untrained and uncertain when it comes to choosing a concrete version of an abstract question, and they

This article was supported in part by the Manpower Research and Advisory Services division of the Smithsonian Institution and in part by the Office of Naval Research. A longer version, written in collaboration with Eugene J. Webb, was presented at the Smithsonian–Navy Conference on Survey Alternatives, Santa Fe, New Mexico, April 22–24, 1975.

The author is extremely grateful to Roger E. Tourangeau, who read and criticized an earlier version of this manuscript and provided a good deal of necessary clarification. David Kenny and G. Evelyn Hutchinson also made useful suggestions.

Requests for reprints should be sent to Phoebe C. Ellsworth, Department of Psychology, Box 11A, Yale University, New Haven, Connecticut 06520.

[1] It is clear that this distinction has much in common with the distinction between deductive and inductive science. The validity, value, and very possibility of each of these approaches are admirably disputed elsewhere, and in the present article I express no opinion at all on this controversy.

cling to the settings and techniques that have been used before. Courses on research methods generally do not pay much attention to the problem of finding concrete instances that fit abstract questions, and while the research literature contains many examples of the ingenious use of natural settings, it is difficult to know how to profit from these examples unless they represent the same variables one wants to study. We can admire, but we cannot emulate. The purpose of this article, then, is to provide a few tentative suggestions and rough guidelines for the location of natural instances of conceptual variables.

I should say at the outset that I am not interested in devising ways of finding settings in which hypotheses are likely to be *confirmed.* The aim of research is to arrange or locate a setting in which a hypothesis can be tested, or a question fairly asked. Selecting a setting on the basis of the adequacy of the test of the hypothesis, however, implies that the investigator will believe a disconfirmation. If he[2] won't, it is no kind of test. A theory or a hypothesis is useful only insofar as it is falsifiable (Popper, 1961), and thus the aim of the scientist is to find a context in which the hypothesis can be falsified to his satisfaction. The only time that it is useful to seek a setting in order to confirm a hypothesis is when everyone believes that the hypothesis is wrong or the predicted outcome impossible. In general, the goal of the hypothesis tester is to find a context in which the hypothesis may or may not be confirmed and in which he would believe a disconfirmation. A secondary goal is to pick a context that is informative enough to generate ideas about why the hypothesis was disconfirmed.

Choice of a Setting Based on Progress So Far

At the early stages of research, the investigator may not yet have a hypothesis, but simply a general area of interest. Some areas of interest imply definite concrete settings: One may be interested in trials, mental patients, blind people, or policemen. In these cases the selection problem is simple; for the initial exploratory stages the researcher may simply pick the courtroom, hospital, institute for the blind, or police station that is most convenient, first checking around a little to make sure that the convenient instance is not a notoriously peculiar one.

Other general areas of interest are abstract: One may be interested in persuasion, deindividuation, helplessness, or role conflict, again without yet having a specific hypothesis. In this case, there is often a much wider range of settings to choose from. Persuasion, for example, can be studied in courtrooms, in mental hospitals, in institutes for the blind, or in police stations. Before moving into the field, the researcher will already have thrown out certain possible instances of his concepts as irrelevant, and in this process the concepts themselves are refined.

So far, I have been more or less assuming that the investigator is interested in an abstract concept but that he is at the exploratory stages of research in an area. When this is the case, the investigator has maximal freedom in his choice of instances, since when nothing is known the information gain is bound to be great, no matter what the starting point. Virgin territory is rare, however, and additional considerations come into play in the secondary growth of follow-up studies. At this stage, the investigator must ask, How has this topic, or phenomenon, been studied before? Unfortunately, many investigators regard this question as the equivalent of the question: How does one study this topic? and they model their contexts as closely as possible on those of their predecessors. For some purposes this is advantageous, for others (such as direct replication) even necessary, but for other purposes it may be disadvantageous. In any case, the decision should not be an automatic one.

All settings have unfavorable quirks and characteristic sources of error, and often one's aim in asking about previous research is to select an instance that is free of the problems of the earlier instances. In what is perhaps the typical case for the social psychologist, the answer to this question will be "in the laboratory," and the characteristic disadvantages will be demand characteristics (Orne, 1962), evaluation apprehension (Rosenberg, 1969), use of subject populations that restrict the range of variation on the dimensions being studied, and treatments that are weak. Some topics, such as interpersonal distance regulation, may have a history of research in nonlaboratory settings and may need the advantages of the laboratory, such as the opportunity it provides to uncorrelate normally correlated variables. No single instance is a perfect

[2] The pronouns *he, his,* and *him* are used throughout this article to refer to both sexes.

embodiment of a set of abstract concepts, and in choosing a way of operationalizing a research question, the investigator always compromises, sacrificing some methodological advantages for the sake of others.

In choosing a new context in which to ask a question that has been asked before, one aim is to achieve a compromise that is as different as possible from the type of compromise typical of that line of research. Thus, if the techniques used in the past allowed great precision of measurement and careful, parametric specification of the treatment levels at the price of a restricted range of both treatments and subjects (as in the case with most psychological research on attitudes, to cite the usual example), a new instance that is characterized by treatments falling outside the previously studied range and by a heterogeneity of subjects is desirable, even if it means that the assessment of treatments and outcomes is relatively crude.

There are some properties that are desirable in any research setting. In brief, one wants an instance that is capable of disconfirming the hypothesis, that allows for fairly precise specification of both independent and dependent variables, that is free of serious confounds, and that is informative, allowing the investigator to collect supplementary data that will be helpful in understanding the results. These criteria apply to both laboratory and field research, and they are rarely or never completely satisfied in a single instance. In choosing a setting, however, it is often wise to look out from one's own small corner, to consider the whole history of research on the topic, and to try to choose a new instance that will help the field as a whole to approximate the fulfillment of these criteria more closely.

It often—perhaps usually—happens that one is working within a tradition but that the specific question addressed is a new one, without even a single laboratory study devoted to it. In this case, the choice of an instance should still be responsive to the needs of the general research tradition, but it should also, and primarily, be responsive to the question itself. The investigator's first task (in any research) is to understand what it is that he is asking.

The Substantive Question: Variables

Most hypotheses can be reduced to the basic form, "Whenever X occurs, Y will happen." X may be very simple, as in the hypothesis, "Whenever fear is aroused, behavior change will be facilitated" (Janis & Feshbach, 1953); moderately complex, as in the hypothesis "Whenever fear is aroused *and* the recommended behavior is easy and totally effective, behavior change will be facilitated" (Leventhal, Singer, & Jones, 1965); or highly complex, as in hypotheses involving very specific higher order interactions among many variables (for examples of such hypotheses in the area of fear arousal and behavior change, see Janis, 1967; Leventhal, 1970). In any case, the investigator's problem is to find an instance of X where Y is measurable. The simpler and more general X is, on the whole, the wider the range of settings the investigator has to choose from. As X becomes more clearly defined and more complex, many settings are ruled out. As I pointed out earlier, generating a list of settings may in itself reveal one's hidden assumptions about X.

The choice of setting for the general hypothesis "Whenever X occurs, Y will happen" is determined by X. The investigator does not look for settings in which both X and Y occur, for to do so would be to load the research, making it more like an exercise in "finding," in McGuire's (1973) sense of the term. The aim is to find a setting that provides a good realization of X, and the *possibility* of Y. In considering Y, one simply wants to assume that no extraneous factors are preventing Y from occurring or the investigator from measuring it. Thus, if the hypothesis is that fear of tetanus leads to getting a tetanus shot, there should be some place in the region where tetanus shots are given and where the records of who gets tetanus shots are available to the investigator; if the hypothesis is that fear leads to affiliation, the setting should include other people; if the hypothesis is that fear leads to compliance, the setting should provide a request or command. Otherwise, one concentrates on finding the most nearly perfect X.

Some hypotheses require a natural field setting because they are about that kind of setting; other, more abstract hypotheses may lead to the field because they involve variables not easily translated into laboratory terms. On the independent-variable side, the example most often given is the high-impact variable (Carlsmith, Ellsworth, & Aronson, 1976). If the question or hypothesis concerns powerful or highly arousing events or strong feelings, the psychological laboratory is usually not the most appropriate place to test it. While considerable ingenuity has been devoted to creating laboratory treatments that approximate the impact of analogous real-world events (e.g., Ax, 1953; Milgram, 1973), relatively little has been devoted to choosing settings where nature provides the powerful manipulation. If one is interested in fear, conflict, grief, or love, the laboratory may be one of the least appropriate settings, because the possible and permissible range of these variables is at best restricted to fairly low levels. Other independent variables such as subtle inconsistencies of information, evaluation apprehension, and distraction may be easy to study in the laboratory. In general, variables for which precise control is more important than impact are particularly suited to the laboratory.

The ethical requirement of weak treatments, with its attendant restriction of range, creates one major limitation on the kinds of question that can be studied successfully in the laboratory. Time constraints impose a second restriction. In general, laboratory studies are limited to studying the acute or reactive form of variables such as self-esteem, liking, or commitment. Chronic self-esteem, lasting friendship, or long-term commitment may have quite different effects from their reactive counterparts—they may not even involve the same underlying psychological processes; thus laboratory results based on acute manipulations may have very limited generality.

Occasionally, laboratory experimenters make an effort to get around this problem by adding pencil-and-paper tests designed to measure the chronic version of their manipulated variable, as in Aronson and Mettee's (1968) research on self-esteem. While this is some improvement over ignoring the problem altogether, it still usually suffers from the other problem of restriction of range in the homogeneous, college-student sample. If the investigator intends his hypothesis and results to apply to the chronic version of the variable or if he does not distinguish between acute and chronic, the study of an appropriate natural population is desirable and often necessary, at least as a supplement.

Weakness and brevity are perhaps the major restrictions typical of treatments administered in the laboratory. On the dependent-variable side, the problem is one of sweetness and light. Subjects who know that they are being observed by psychologists are motivated to look healthy, normal, tolerant, and intelligent, to turn the other cheek when attempts are made to provoke them to anger, and to judge others as they would be judged (Carlsmith, Ellsworth, & Aronson, 1976; Rosenberg, 1969). Hypotheses that involve socially undesirable behavior may often be better tested outside of the laboratory, ideally in a situation where subjects do not know that they are being observed by a social scientist.[8]

These substantive reasons for choosing a setting are first of all questions of construct validity, and a careful definition of the constructs involved is probably the best guideline to the choice of a setting. The chronic, process version of a variable may resemble the acute, reactive form in name only. An occasion of lowered self-esteem may have nothing to do with a lifetime of low self-esteem. Construct validity may be just as much a problem in cases where the problem is apparently a simple question of restricted range. Semantics aside, there is no a priori reason to assume that the same processes or consequences are characteristic of the "same" variable at different intensities: Mild anger may be qualitatively distinct from rage. For any given variable, of course, the question is an empirical one. Acute anxiety may resemble chronic anxiety in many important ways; faint anxiety may be linearly related to intense anxiety. The problem is that in most cases we haven't a particle of evidence that this is true, and thus to conduct research on the unconsidered assumption that the name of the variable guarantees that it is the thing we are interested in is foolhardy at best.

The second substantive reason for choosing a natural setting is that one ultimately wishes to apply one's results to similar settings; this is the problem of generality. In some ways, the standard philosophy-of-science line treats all hypotheses as universal hypotheses; we start out phrasing a hypothesis in very abstract terms, as though we expected it to be universally true. This leads to a concern for unconfounding variables and testing the abstract hypothesis in the purest way possible: The most nearly perfect X is the one that best fits the abstract conceptualization. For the researcher who only wants to generalize to a limited range of settings, this may not be the best strategy. It may be better for this purpose to sacrifice some of X's purity for the sake of a setting that falls within this range (Cook & Campbell, 1976). Much of the research in the field of psychology and the law,

[8] When I say "ideally" I am speaking solely in methodological terms. Unfortunately, the "best" methods are not always the most ethical ones.

for example, suffers from the problem of these conflicting goals. In a classroom simulation study, the experimenter can prepare materials in which race of the offender, type of crime, past record, income, and amount of bail set are completely independent, as the "scientific study" of their relative contributions seems to require, but such an exercise does not accomplish very much if the experimenter intends to generalize to jury decision making. The noninteractive, large-group setting, the hypothetical nature of the case, and the homogeneous student-subject population make generalization to real juries extremely risky. Of course, if the researcher is interested in how people—any people—combine somewhat colorful information in a cognitive task, the classroom setting may be perfectly appropriate. It is often difficult to have it both ways. Again, it is a matter of choosing a setting that reflects the investigator's concerns: Is the question about abstract conceptual variables, or is it about juries?

There is no reason, incidentally, that a study has to be a laboratory study *or* a field study. Clearly there are a great many research questions that are most clearly illuminated by a series of studies, some of which take place in the laboratory, and others in natural settings. This point is frequently made and needs no further elaboration here. In addition, however, a single experiment may have *both* laboratory and field components. There is no general reason that the treatment and the measures must occur in the same context. The experimenter may take advantage of one of nature's powerful or long-term treatments to select subjects and may then bring these subjects into the laboratory to observe their behavior in a situation that is as controlled and structured as he wishes. Rubin's research on romantic love (e.g., Rubin, 1970) is a good example of this strategy; realizing that he was not likely to be able to create true love, Rubin chose couples who had been together for several months and observed their interaction in the laboratory, where very precise measurements were possible. Similarly, it may be possible to induce a treatment in the laboratory and then send the subjects back out into life, observing their behavior over a longer period of time than is possible in the usual laboratory experiment. This strategy has been followed by Janis (1975), who exposes overweight subjects to his independent variables in the laboratory and then uses weight loss as a long-term behavioral measure of the effects of these variables.

The Formal Question: Relationships

The substantive aspects of the experimenter's question lead him to find instances that seem to typify his conceptual variables. Additional guidelines for the choice of a setting are imposed by the form of the question or hypothesis. The form of the hypothesis defines the plausible rival hypotheses (Campbell & Stanley, 1966) and thus the necessary control groups. In choosing a natural setting, it is important to find a setting which has not only a treatment that fits the construct but also an appropriate control group. The nature of the requisite control groups should be clearly defined and a setting chosen in which these groups exist and in which their behavior can be measured in the same terms as that of the treatment group. Thus, if the hypothesis says that the mere presence of X will lead to Y, it is important to choose a setting that also includes some subjects who get no X at all. The contrast is between some (perhaps undefined) amount of X and a total absence of X. If the hypothesis says that Y increases with X, then a setting that includes a zero-baseline group may be less relevant than a setting that provides a *range* of X great enough to allow for the discrimination of enough levels to test the monotonic hypothesis. If there are plausible rival hypotheses, it is important to try to find a setting in which these hypotheses can also be tested or ruled out. Obviously, no setting can provide complete control or assessment of all alternative explanations, but often some settings are much better in this respect than others. Again, if there is a tradition of research that has been strong on one type of control and weak on another, the investigator may aim for a new setting that is strong on the latter type.

There are many different ways of defining the form of a question, and occasionally the line between form and substance is arbitrary or vague. Basically, I am talking about the abstract properties of the relationship among variables. Statements such as "When X, then Y" or "When more X, then more Y" impose certain formal requirements for adequate experimental design regardless of the particular Xs and Ys involved. A given question may be categorized within a number of different formal frameworks, each of which may have implications for the most appropriate method of seeking an answer. Some of these frameworks are discussed in the following sections.

First, one may consider the scope of the hypothesis or question: It may be universal, existential, general, or particular. A particular hypothesis states that on this occasion, a particular event will happen. The setting is defined by the question, and the main problems involve the measurement of the event. For example, one might predict that "on the day that P's mother dies, P will have a psychotic breakdown." Such hypotheses are rare in psychology, and are confirmed more rarely yet, as we are still far from the ability to specify all the variables operating in the individual case.

An example of a universal hypothesis is the hypothesis that all people recognize certain facial expressions as characteristic of certain emotions (Ekman, Sorenson, & Friesen, 1969). An existential hypothesis simply states that a phenomenon can exist, once, without regard for frequency: There is a philosopher's stone, a female genius, a culture in which a frown is recognized as a sign of delight. This last example points up the intimate relationship between universal and existential hypotheses: The disconfirmation of a universal hypothesis is the confirmation of an existential one. If the hypothesis is that X is universal, its disconfirmation is the demonstration of the existential hypothesis that not-X exists. In choosing a setting in which to prove an existential hypothesis, one searches for a setting in which it is most likely that X will occur, since the single occurrence of X is the proof of the hypothesis. The strategy for testing a universal hypothesis is conceptually the same: One looks for a setting in which not-X is likely to occur. If a single not-X instance can be found, the hypothesis is disconfirmed. (That an existential hypothesis can be confirmed but never disconfirmed, while a universal hypothesis can be disconfirmed but never confirmed, is congently argued by Popper [1961] and provides a further indication of the misconception of "finding" confirmatory settings [McGuire, 1973], at least for universal hypotheses.) To take a concrete example, in seeking a context in which to test the universal hypothesis that facial expressions are associated with specific emotion labels, Ekman, Sorenson, and Friesen (1969) chose to study a New Guinea highlands tribe that had had minimal contact with Europeans. This was a setting in which not-X was likely to occur: It was very different from all other settings in which the expression-labeling relationship had been demonstrated, and it also ruled out many rival hypotheses that might have accounted for the relationship observed in other settings, but especially that of a shared, arbitrary code acquired via contact with Europeans or their communications media. In practice, choosing a setting that maximizes the probability of not-X may be equivalent to choosing a setting which maximizes the influence of all plausible, nonuniversal causes of X. It is a favorite technique of anthropologists in response to the glib universal statements of psychologists: One culture that lacks the Oedipus complex is sufficient to reduce that hypothesis from a universal one to a merely general one.

In fact, most hypotheses in social science are general, stating that on the whole, X is true. Many such hypotheses may be stated in the abstract *as though* they were universal, but typically the issue of universality is not a primary concern of the researcher; the investigator who has demonstrated that "residential proximity leads to marriage" is likely to greet travelers' tales of exogamous social groups with a tolerant smile, but not to abandon his research. For many researchers, the scope of the phenomenon is not a matter of explicit concern, beyond the rough assumption that it is reasonably general within some vaguely defined universe.

The issue of universality does arise, at least implicitly, when a hypothesis is disconfirmed or a finding fails to be replicated. The fact that most researchers take these events as a sign that something must be modified indicates that universality is at least an unspoken concern. Following a disconfirmation, an investigator can revise the theory (perhaps hoping that the new version will be universally true), can limit the conditions in which the relationship is expected to obtain (hoping that the universality of the hypothesis can be maintained if the universe is restricted), or can discount the failure (hoping that the failure really has no implication for the original statement, which is still universally true). Theoretically then, once the hypothesis is refined and all the conditions are specified, the hypothesis—now greatly qualified—may approximate universality.

But exogamy among the Bozo tribe is still not much of a concern—researchers are often willing to pay lip service to the appropriate limiting condition without much thought. The scope of the phenomenon usually becomes a matter of serious concern only when it is restricted in theoretically interesting ways. Thus, the researcher who predicts that fear leads to affiliation may not be concerned about situations in which affiliation is physically impossible, or by vague reports of the inhabitants of some tropical isle who isolate themselves when afraid, or even by another researcher who fails to replicate this effect in a setting similar to his own; he becomes concerned mainly when the failure to replicate is associated with some conceptual variable claimed to limit the scope of the theory. Then he searches for (or creates) a setting in which it is possible for this usurper variable to vary, so that he can discover whether or not that particular variable imposes a general restriction on the theory. So, for example, if it is claimed that fear will not lead to affiliation if the person is ashamed of his fear, the researcher may study the responses of older and younger children to a frightening situation (predicting that younger children will be less ashamed of their fear and will affiliate more), or males and females (predicting that at any given age, fear is considered less shameful for females, who should therefore affiliate more), or people with common versus uncommon phobias (predicting that those with uncommon phobias will feel more deviant and ashamed, and will therefore affiliate less). The choice of settings in which to test the validity of the vaguely general hypotheses characteristic of social scientists is often determined by the plausible rival hypotheses advanced by other social scientists.

Another way of classifying the form of a hypothesis is in terms of the necessity and/or sufficiency of the independent variable. A hypothesis may claim that X is necessary to produce Y, that X is sufficient to produce Y, or both. A necessity hypothesis directs the investigator's attention to settings characterized by the dependent variable Y; since he is predicting that whenever there is Y, X must be present, the disconfirming instance he seeks is one characterized by Y but not X.[4] Looking for settings characterized by X may not fulfill this objective, since X, while necessary for Y, may not be sufficient, and thus the occurrence of X without Y does not disconfirm the hypothesis in the same way that the occurrence of Y without X does. A sufficiency hypothesis directs the investigator's attention to settings characterized by X;[5] there should be no such settings in which Y does not occur (although there may be settings in which Y occurs without X, since other things besides X may also be sufficient to produce Y).

If the hypothesis is that X is both necessary and sufficient to produce Y, then any occurrence of X or Y alone is grounds for rejecting the hypothesis. In testing the hypothesis that X is both necessary and sufficient for the occurrence of Y, one must choose a setting in which it is possible for both X and Y to be either present or absent. The investigator is predicting that of the four possible observable combinations (and they must be possible, or else the test is not a good one)—X and Y, X but no Y, Y but no X, no X and no Y—only the first and last will actually occur. If the investigator hypothesizes that a certain *amount* of X is necessary and sufficient to produce Y—in other words, if he has a threshold hypothesis—any amount less than that threshold can serve as a conceptual zero, and he need not seek a setting that permits a total absence of X.

In general, the researcher who has a presence–absence hypothesis or a well-specified threshold hypothesis looks for a setting in which two clearly differentiated groups can be measured—one of which is above threshold on X and one of which is below (or at absolute zero). In practice, other control groups or measurement occasions will no doubt be necessary to control for plausible rival hypotheses, but in essence the question is a two-group question; often it is embodied in some sort of pretest–posttest design, especially in field settings where nonrandom assignment is common and thus some check on the equivalence of groups is required. Logically, a question involving an X threshold that is not zero takes the same form, but in practice the investigator may not be able to specify the critical level of X in advance and so may seek a situation in which many groups can be tested at different levels of X. In analyzing the data, he will look for a sharp discontinuity in the function and, if brave and confident, may conduct a two-group follow-up study in which he tries to predict the discontinuity precisely.

If, on the other hand, the investigator's hypothesis involves continuous variables, taking the form "the more X, the more Y" or "the more X, the less Y"—in other words, if it is a degree hypothesis—the two-group setting is less appropriate. Instead, the investigator seeks a setting in which a wide range of X is present, so that he can choose Xs at many different levels. The natural setting may be especially useful in this case, as in the case of chronic variables, in that the laboratory is typically ill-suited to produce the heterogeneity of X values necessary for a strong test of a degree hy-

[4] Logically, of course, the researcher could also go out and look for situations *without* X (predicting that no Y will occur), but this strategy is typically less efficient.

[5] Or situations *without* Y (see Footnote 4).

pothesis. If the hypothesized relationship is nonmonotonic, the need for a wide range of X values is even more pressing.

Plausible Rival Hypotheses and Formal Control

Throughout this article the notion of control has appeared, now implicitly, now explicitly, as a kind of leitmotif. One's choice of controls—whether control groups or control observations on the same group (I intend to cover both by my use of the term *control group*)—is equivalent to one's choice of a falsifiability criterion for the hypothesis. For any kind of hypothesis-testing research, the availability of an appropriate control group or groups should be a major consideration in the choice of a natural setting. It is my hunch that many investigators devote a great deal of effort and creativity to finding a setting or population that provides a good embodiment of the treatment variable and that they then cast about hurriedly at the last minute for some sort of control group. The control group is exactly as important as the treatment group in research, and in fact it is impossible to separate the value of one from the value of the other.

Quasi-experimental designs require particular attention to the achievement of appropriate control groups, since the absence of random assignment raises serious dangers of noncomparability of groups and consequent uninterpretability of results. Many studies could be improved by the use of multiple control groups; the more different kinds of control groups (or control observations), the greater the number of rival hypotheses that can be rendered implausible, and the stronger the case for the causal relationship the experimenter has in mind.

But this utopian ideal has serious economic costs. When faced with the realization that 5 or 8 or 10 different control groups (as well as several versions of the treatment group) may be necessary for the best possible test of the hypothesis, the researcher may throw up his hands in despair and decide that his creative impulses are better satisfied elsewhere. Given these constraints, the question then becomes, *Which* control groups are the most important for a good test of my hypothesis? This question may be addressed by imagining the treatment group behaving just as predicted, and then imagining the plausible rival hypotheses most likely to be raised by others to explain one's results and to refute one's conclusions. What kind of information would be necessary to eliminate these alternative explanations? What kind of control observations or control groups would provide that information? The unthinking choice of the most handy no-treatment baseline group may result in information that is not particularly useful.

Just as in the case of the original hypothesis, the likely alternatives or plausible rival hypotheses can be examined either substantively or formally. It is impossible to provide a general listing of alternatives based on the content of the question; these will be different for every question, and only the investigator can decide which of the substantive alternatives create real threats to his interpretation of the results.

Assessment of formal alternatives is a less difficult matter. There is only a finite number of general formal relationships that can account for an observed correlation between X and Y. When we say that correlation between X and Y does not imply causation, we usually mean that the correlation does not imply the particular causal relationship "X causes Y"; most correlations reflect some kind of causal relationship, and the problem facing the investigator is to distinguish the causal relationship he favors from the alternatives. Köbben (1970) has provided a succinct and valuable outline of the possible causal relationships between two variables, and this outline is the basis of the guidelines that follow.

In essence, what follows is a list of the formally possible rival hypotheses for the situation in which "X causes Y" is the actual hypothesis. It is meant as a checklist or set of guidelines that can be applied to a hypothesis as a means of defining the most important control conditions. For any given hypothesis, some of the suggested rivals may be completely *implausible;* the investigator will not be able to think of any credible alternative that takes that particular form. In this case, he will typically decide that no control group is necessary to test that particular set of alternatives. Some hypotheses, in their initial form, are not specific enough so that all of the alternatives can be distinguished from the main hypothesis. Sometimes an investigator is mainly interested in prediction and doesn't care about determining the contributing processes with any precision. The investigator must decide not only which alternatives are plausible, but which ones matter. My hope is that some of the alternatives listed, when embodied in a form that makes them pertinent to the investigator's substantive question, may seem important, and it is these that will direct the search for control groups. Finally, I should point out that one person's "plausible alternative" may be another's central question. For simplicity's sake, I start with the notion that the hypothesis of interest is "X causes Y" and that the others are disconcerting rivals, but for a particular investigator, "X causes Y" may be an alternative that he will want to rule out or examine by using an appropriate control group.

I should point out that the arguments presented below will have to be altered and elaborated slightly to accommodate different types of causal statement. While I do not wish to get involved in philosophical discussions of causality, to leave the matter entirely up to the reader's common sense is to leave the matter ambiguous. When I use phrases such as "X causes Y"—and I use them rather loosely—I mean, roughly, that X is sufficient to produce Y, and that in some cases of non-X, non-Y obtains. (For those who think in these terms, the guidelines may serve as an aid to deciding which kinds of non-X are relevant.) If the hypothesis is that X is *necessary* for Y, then the central hypothesis is that not-X produces not-Y, and the alternatives must be rephrased accordingly. If the hypothesis is that X is necessary and sufficient for Y, a slightly different but also fairly obvious set of rephrasings is in order.

Finally, the hypotheses are generally discussed in terms of discrete variables or levels, rather than continuous ones, and some adjustments are necessary to handle hypotheses of degree.

Assume that you wish to test, examine, explore, or question the statement "X causes Y." Randomized assignment is impossible; you may or may not have a baseline control group; and for the time being, you are anticipating positive results.[6] You think forward to the moment when you have achieved these results and imagine the captious criticisms of your colleagues. If your foresight has been accurate, you will have collected information relevant to these objections before they are raised. The hypothesis is

$$X \rightarrow Y.$$

The first possible alternative is

$$Y \rightarrow X.$$

In a true experiment, with random assignment, this cannot be a confound, since the experimenter knows what caused X *in that situation:* he did. In quasi-experimental designs this may or may not be a plausible rival hypothesis. Clearly, if Y is an increase in heart rate and X is an earthquake, plausibility is low. In general, we can weaken the plausibility of this kind of confound by choosing settings in which (a) the temporal relations are such that X clearly precedes Y, and/or (b) we already know what causes X (e.g., shifting of the earth's crust) or at least know enough to know that it has nothing to do with Y. In some settings, particularly situations involving simultaneous measurement of chronic attributes, the "Y causes X" alternative may create a greater threat. Does considerate behavior lead to popularity or does popularity lead to considerate behavior? (Of course either of these hypotheses may be the one entertained, and the other the rival; the formal problems of control are the same.) If this kind of rival hypothesis is plausible, you want to find a situation in which either (a) you can introduce or at least define variables controlling X, or (b) you can sort out the temporal ordering. In the specific case, you can create popularity or considerate behavior, or you can find a situation in which either considerate behavior or popularity clearly preceded the other. (The last of these four possibilities taxes the ingenuity and may in fact be impossible.) If the hypothesis is that popularity leads to considerate behavior, one might arrange for a random sample of freshman women to be sought after, deferred to, courted, and included as stars in all important social activities and see if they are especially considerate (compared to their peers and their earlier selves) at the end of a year. If the hypothesis is that considerateness leads to popularity, one might study transfer students or new employees who have not yet had a chance to build up a reputation of popularity in their setting.

Perhaps somewhat more common, especially in natural settings, is the possibility that

$$X \rightleftarrows Y.$$

This is the vicious (or possibly benign) circle of functional interdependence, perhaps more characteristic of continuous variables than discrete ones. Failure leads to depression and depression leads to failure; popularity leads to charity and charity

[6] Ideally, of course, this exercise should be applied to all possible patterns of significant results, not merely to the hypothesized pattern.

leads to popularity; money leads to power and power leads to money, and so on. When this hypothesis takes the equivalent form of

$$X \rightleftarrows Y,$$

it is the homeostatic hypothesis. In some cases the investigator may not be particularly concerned about this type of rival hypothesis: If all he wants to show is that X results in Y, without caring one way or the other about whether Y also leads to X, he need not bother to control for this alternative. He need only worry about finding a situation in which he can *begin* with an X that was not caused by Y (so as to rule out the possibility that X only enhances Y when Y is present to begin with). This may involve happy people who suddenly fail, considerate people who just moved into town, or poor people who are given power, for example. If, however, the investigator either wants to reject the circular alternative or to demonstrate it, he should find situations that are characterized by non-Y-produced X and situations that are characterized by non-X-produced Y. In addition, he will probably want to follow up these settings over extended periods of time, observing the relative fluctuations in X and Y, or, if the Xs are events in the past, to use some sort of lagged correlational procedure for analysis. Other good settings are settings in which intervention is possible: If at one time one can inhibit or prevent X and observe the effects on Y, and at another time (or in an equivalent setting) one can inhibit Y and observe X, one is in a good position to differentiate the one-way, cause–effect model from the two-way model. In fact, one can differentiate all three possibilities outlined so far, since the controls that work for the functional interdependence hypothesis also work for the $Y \rightarrow X$ alternative.

Next we come to the familiar third-variable correlation:

$$\begin{array}{ccc} & & X \\ & \nearrow & \\ Z & & \\ & \searrow & \\ & & Y \end{array}$$

If a simple correlational study is performed, X and Y are observed to co-occur, but to conclude that $X \rightarrow Y$ would be erroneous. Domineering mothers may produce shy sons, but it is also possible that weak or absent fathers cause mothers to become more assertive and at the same time cause sons to behave in a dependent manner for lack of a strong male model. Finding appropriate settings in which to rule out this type of alternative requires considerable care and imagination because, although the problems of conclusion drawing are formal ones, requiring settings with and without Z and predicting variability in Y accounted for by X to the same extent in each, one cannot do this without first considering the substantive question of which Zs are plausible contenders, given one's particular question. Clearly there are myriad possible Zs for any hypothesis, and given limited time and resources, the difficulty lies in distinguishing the plausible ones. Again, presenting the hypothesis to one's friends and foes as though it had already been confirmed by the data is often a useful way of eliciting large numbers of both plausible and implausible Zs. Having decided upon the plausible ones, one has several options: (a) to choose a control setting in which Z is absent (or at an improbable level), in which case if Z is essential, X and Y should also be low or absent; (b) to choose a setting in which Z is at a constant level, in which case if Z is essential, X and Y should not vary; (c) to choose settings with and without Z (this of course is necessary if one's hypothesis *is* that Z causes both X and Y); (d) to choose a setting in which it is possible to interfere with Z and examine the effects on X and Y; (e) to choose a setting in which Z is absent and X can be introduced; and (f) to choose any old setting where all three could be present and partial out Z statistically. This last option is less satisfactory, as pointed out by Campbell and Stanley (1966) in their discussion of ex post facto designs, since many important Zs are extraordinarily complex, and partialing out a portion may still leave many spurious correlations. The third-variable correlation problem is one that should generally be considered carefully no matter what form of hypothesis one is testing, as it almost always invalidates any other hypothesis.

The fourth possibility is of less universal concern:

$$X \rightarrow Z \rightarrow Y.$$

That is, X only leads to Y via an intervening variable Z. Many investigators, especially those with practical concerns, may not regard this as a serious problem, and indeed during the early stages of hypothesis formation and research, the hypothesis may not be so sharply delineated that the $X \rightarrow Z \rightarrow Y$ possibility can even be considered a definite "rival." Instead, the investigator may predict that $X \rightarrow Y$ and not yet be concerned with the exact processes by which this occurs. Thus, for many researchers, this hypothesis may be more in the way of a refinement or an explanation than an alternative.

If, however, the investigator wishes to rule out (or demonstrate) the $X \rightarrow Z \rightarrow Y$ hypothesis, he must seek settings (a) in which X does not produce Z (and predict that Y will or will not still follow); or (b) in which something else besides X—preferably a variety of something elses—produces Z (and then if Y still follows, X is really of very secondary importance); or (c) in which X and Z can be prevented independently. As in the case of third-variable correlations, some ingenuity may be necessary to do justice to the range of possible Zs. Any Zs that are perfectly correlated with X may be of minor concern, merely involving a definitional problem of the boundaries between X and Z. The situation where Z is neither directly manipulable nor directly measurable is not uncommon in theory-testing research and is usually dealt with by showing that a variety of otherwise unrelated $X \rightarrow Y$ demonstrations can be predicted by postulating Z, and erased by using Xs that are similar in all respects except that they are unlikely, according to theory, to produce the hypothetical Z.

The next possibility is that X is not enough:

$$\left.\begin{matrix} X \\ Z \end{matrix}\right\} \rightarrow Y.$$

That is, X *and* Z together lead to Y. With discrete variables or a threshold hypothesis, this is a simple interaction; with continuous variables or a hypothesis of degree, it may also reflect two main effects.

In some ways the practical problems of choosing a setting to control for this alternative are similar to those raised by the problem of third-variable correlation. The interaction possibility is more likely to lead the investigator to falsely conclude that X has nothing to do with Y, if he happens to pick a situation with an uncongenial level of Z. However, the problems of choosing a setting are similar, and many of the same settings that can rule out third-variable correlations can rule out interaction effects as well. Again, the first and most difficult step is to guess at the important Zs. The essence of the formal control is to find or create settings or occasions in which X occurs alone, Z occurs alone, X and Z occur together, and—for symmetry and to control for the possibility that X and Z aren't even the right variables—settings in which neither X nor Z occurs; the factorial design is the design for testing interactions. If the third-variable hypothesis is the right one, Y should occur when Z occurs even if X is blocked; if the interaction hypothesis is correct, Y should not occur unless both X and Z are present.

All of these alternative hypotheses have been causal hypotheses, under the loose definition of cause given earlier.

X and Y may also vary together without being causally related at all (or at least not at any level less cosmic than the trepidation of the spheres). Historical trends may coexist with no close interrelationships, and produce significant correlation coefficients when data are analyzed longitudinally. Thus, the development and proliferation of communism may be highly correlated with the development and proliferation of the automobile, with no functional relationship between them. Adding a cross-sectional setting to a longitudinal analysis may clarify the issues involved, as will choosing a smaller unit of time or space in which one of the variables changes drastically. If the other shows corresponding changes as more and more of these smaller settings are examined, the hypothesis of independence becomes correspondingly less and less plausible.

Finally, Y may be a logical implication of X, or in fact Y may *be* X by any other name. That is, being a woman doesn't *cause* having ovaries, nor does being a later-born child cause one to have an older sibling. Here again, one tries to think of settings in which the hypothesis might not be confirmed. If one can conceive of none at all, it is probably a good idea to ask whether there really is an empirical hypothesis.

More and more, we are coming to recognize that interrelations may be causal but much more complicated than we can assess with our usual methods (McGuire, 1973). There may be a plethora of Xs, Ys, and Zs locked together in a system in which all the simple types of relationship so far discussed are represented, perhaps represented more than once. It is in just these instances that the typical laboratory experiment is weakest; so much is held constant that there is no opportunity for this sort of complex causation to manifest itself.

Once again, we should point out that one person's hypothesis may be another person's rival. For some questions, the simple $X \rightarrow Y$ structure may be an alternative explanation. In any case, one starts with the relationship of interest and then scans the logical rivals to see which are plausible rivals in the particular instance; then one chooses one's setting in such a way as to control for or measure the strongest contenders.

Obviously, the same set of procedures may be followed, with minor modifications, if one's hypothesis takes the form $X \rightarrow \bar{Y}$ or $\bar{X} \rightarrow Y$.

If one is arguing that X is necessary and sufficient for Y, one is predicting not only that $X \rightarrow Y$, but also that $\bar{X} \rightarrow \bar{Y}$. In order to make this stronger dual statement, it may be necessary to add $\bar{X}$ control groups across the various rival hypotheses described above.

Still further modifications will be necessary to set up appropriate controls for hypotheses involving multiple independent variables and/or multiple levels of the independent variable, but the logic is the same. Likewise, some adjustments will be necessary if the hypothesis is one of degree and the variables are continuous; here a correlational analysis is an essential addition to the procedure.

In sum, the decision to move to a natural setting is often laudable, but rarely enough. It is but one of a whole series of decisions about the essence of one's question and the logic of its exploration. While finding good test settings in the field raises a somewhat different set of issues than creating good test settings in the laboratory, the essential logic remains unchanged.

REFERENCES

Aronson, E., & Mettee, D. R. Dishonest behavior as a function of different levels of self-esteem. *Journal of Personality and Social Psychology*, 1968, *9*, 121–127.

Ax, A. F. The physiological differentiation between fear and anger in humans. *Psychosomatic Medicine*, 1953, *15*, 433–442.

Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally, 1966.

Carlsmith, J. M., Ellsworth, P. C., & Aronson, E. *Research methods in social psychology.* Reading, Mass.: Addison-Wesley, 1976.

Cook, T. D., & Campbell, D. T. The design and conduct of quasi-experiments and true experiments in field settings. In M. Dunnette (Ed.), *Handbook of industrial and organizational research.* Chicago: Rand McNally, 1976.

Doob, A. N., Carlsmith, J. M., Freedman, J. L., Landauer, T. K., & Tom, S., Jr. Effect of initial selling price on subsequent sales. *Journal of Personality and Social Psychology*, 1969, *11*, 345–350.

Ekman, P., Sorenson, E. R., & Friesen, W. V. Pan-cultural elements in social displays of emotion. *Science*, 1969, *164*, 86–88.

Festinger, L., Schachter, S., & Back, K. *Social pressures in informal groups.* New York: Harper, 1950.

Janis, I. L. Effects of fear arousal on attitude change. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 3). New York: Academic Press, 1967.

Janis, I. L. The effectiveness of social support for stressful decision. In M. Deutsch & H. Hornstein (Eds.), *Applying social psychology.* New York: Basic Books, 1975.

Janis, I. L., & Feshbach, S. Effects of fear-arousing communications on attitude change. *Journal of Abnormal and Social Psychology*, 1953, *48*, 78–92.

Knox, R. E., & Inkster, J. A. Post decision dissonance at post time. *Journal of Personality and Social Psychology*, 1968, *8*, 319–323.

Köbben, A. Cause and intention. In R. Naroll & R. Cohen (Eds.), *A handbook of method in cultural anthropology.* Garden City, N.Y.: Natural History Press, 1970.

Leventhal, H. Findings and theory in the fear communications. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 5). New York: Academic Press, 1970.

Leventhal, H., Singer, R. P., & Jones, S. The effects of fear and specificity of recommendation upon attitudes and behavior. *Journal of Personality and Social Psychology*, 1965, *2*, 20–29.

McGuire, W. J. The yin and yang of progress in social psychology. *Journal of Personality and Social Psychology*, 1973, *26*, 446–456.

Milgram, S. *Obedience to authority: An experimental view.* New York: Harper & Row, 1973.

Orne, M. On the social psychology of the psychological experiment. *American Psychologist*, 1962, *17*, 776–783.

Popper, K. R. *The logic of scientific discovery.* New York: Science Editions, 1961.

Rosenberg, M. J. The conditions and consequences of evaluation apprehension. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research.* New York: Academic Press, 1969.

Rubin, Z. Measurement of romantic love. *Journal of Personality and Social Psychology*, 1970, *16*, 265–273.

Sidman, M. *Tactics of scientific research.* New York: Basic Books, 1960.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. *Unobtrusive measures.* Chicago: Rand McNally, 1966.