Human-Computer Interaction: Toward the Year 2000

Written and Edited by
Ronald M. Baecker
Jonathan Grudin
William A. S. Buxton
Saul Greenberg
methodology matters: doing research in the behavioral and social sciences

joseph e. mcgrath
psychology, university of illinois, urbana
23 october 1994

"Doing research" simply means the systematic use of some set of theoretical and empirical tools to try to increase our understanding of some set of phenomena or events. in the social and behavioral sciences, the phenomena of interest involve states and actions of human systems — of individuals, groups, organizations, and larger social entities — and the by-products of those actions.

the meaning of research evidence, in any area of science, is inherently tied to the means or methods by which that evidence was obtained. hence, to understand empirical evidence, its meaning, and its limitations, requires that you understand the concepts and techniques on which that evidence is based.

this chapter is about some of the tools with which researchers in the social and behavioral sciences go about "doing" research. it raises some issues about strategy, tactics and operations. especially, it points out some of the inherent limits, as well as the potential strengths, of various features of the research process by which behavioral and social scientists do research.

some basic features of the research process

doing research, in the behavioral and social sciences, always involves bringing together three sets of things:

(a) some content that is of interest,
(b) some ideas that give meaning to that content, and
(c) some techniques or procedures by means of which those ideas and concepts can be studied.

for example, the contents of a study might involve the behavior of a jury, conversations in a family about buying a new car, the voting behavior of members of a community, voting in a park, championship patterns in a small town, and so forth. the ideas might include the concept of conformity, the notion that education affects political preferences, the concept of conformity, the hypothesis that groups whose members like one another perform tasks better than groups whose members do not like each other, and so forth. the techniques might include a questionnaire to assess individual attitudes toward a car or a candidate or group mates; a set of procedures for observing family discussions about cars and money; a means to gather election returns; a plan to evaluate the quality of group task products; and so forth.

i will refer to these three sets of things more formally, as three distinct, though interrelated, domains:

(a) the substantive domain, from which we draw concepts that seem worthy of our study and attention;
(b) the conceptual domain, from which we draw ideas that seem likely to give meaning to our results; and
(c) the methodological domain, from which we draw techniques that seem useful in conducting that research.

furthermore, research always deals with several levels of phenomena: with relations between units or elements within a context or embedding system. the elements, relations, and embedding systems have different forms in each of the three domains (see figure 1).

 substantive domain

in the substantive domain, i will call the units or elements phenomena, and the relations among them patterns of phenomena. these phenomena, and patterns of them, are the object of our study. for the behavioral and social sciences, the phenomena of interest involve the states and actions of some human systems — individuals, groups, organizations, communities, and the like — and the conditions and processes that give rise to and follow from those states and actions.

another way to say this is to say that the behavioral and social sciences study "actors behaving toward objects in context." an example would be "an individual casting a vote in a county election." another example would be "the number of units produced in the week of april 12th by group 32 of the production division of the danville plant".

it must be understood that both "actors" and "context" here refer to human systems at any of a number of levels of organization, individual, group, community, organization, and so on. different behavioral and social sciences specialize in the study of different human systems — that is, in the study of phenomena and patterns at different levels and of different kinds. the rest of this book presents material that illustrates many of the substantive phenomena and patterns that have been studied within the field of human-computer interaction.

conceptual domain

for the social and behavioral sciences, the elements of interest in the conceptual domain are properties of the states and actions of those human systems that are the focus of study — properties of "actors behaving toward objects in context." these might include such familiar ideas as "attitude," "cohesion," "power," "social pressure," "status," as well as many others that are used in social and behavioral science research.

relationships in the conceptual domain refer to any of a variety of possible ways in which two or more elements can be connected. some of those ways are viewed as "causal" connections. some simply are chronological relations. for example, two elements can be equal or unequal, they can be related linearly or non-linearly, one can be a necessary or sufficient cause of the other, one can include the other, the relations between them can be one way or reciprocal, and many more. materials from the conceptual domain — properties, and relations among those properties — are the "ideas" that can give meaning to the phenomena and patterns that we study in the substantive domain.

methodological domain

in the methodological domain, elements are methods. i will call the methods modes of treatment (of properties of phenomena). modes of treatment are different ways by which a researcher can deal with a particular feature of the human systems that are to be studied.

one set of such modes of treatment include various techniques for measuring some feature (that is, for assessing the state or magnitude of some property of some actors-behavior-context), so that the researcher can determine what value or level that feature has for each "case" to be studied. measurement methods include such things as: a questionnaire, a rating scale, a personality test, instruments for observing and recording communications, techniques for assessing the quality of some products resulting from individual or group task performance, and the like. (more is said about kinds of measures near the end of this chapter.)

modes of treatment also include various techniques for manipulating some feature of a research situation (that is, some property of an actor-behavior-context), to carry out an experimental manipulation of a feature of the situation (sometimes referred to as "manipulating a variable") means making that feature have one particular predetermined value or level for certain "cases" to be studied and another specific predetermined value or level for certain other "cases," so that the effect of differences in that property can be assessed by comparing these two sets of "cases." for example: you might want to study the effectiveness of a particular human-computer system by studying two sets of work groups, one set of groups working with that computer system and the other set doing the same tasks "manually." social psychologists have tried to manipulate features of the systems they study by a number of techniques, such as:

(a) giving instruction to participants (e.g., trying to motivate them to try hard by telling them that there will be a valuable prize for the best product);
(b) giving instructions on the work to be done;
(c) selecting materials for use (e.g., trying to produce differences in task difficulty by giving some participants very difficult word problems to complete, and giving other participants easier problems of the same type);
(d) giving feedback to participants (e.g., trying to induce feelings of success or failure by telling some participants they did well, and telling others they did poorly, on a previous task);
(e) using experimental conditions (e.g., trying to establish different degrees of liking for group members by having an experimental assistant who is pretending to be a normal participant work very hard in some groups and act indifferent in others).

(more is said about techniques for manipulating variables near the end of this chapter.)

modes of treatment of variables also include a set of techniques for controlling the impact of various "extraneous" features of the situation — features that are important but that are not going to measure or manipulate in a particular study. these include: techniques for experimental control, by which you make certain features take the same predetermined value for all cases in the study (e.g., study only 6-year-olds to control on
RESEARCH METHODS AS OPPORTUNITIES AND LIMITATIONS

Methods are the tools—the instruments, techniques and procedures—by which a science gathers and analyzes information. Like tools in other domains, different methods can do different things. Each method should be regarded as offering potential opportunities not available by other means, but also as having inherent limitations. You cannot pound a nail if you don’t have a hammer (or some functional equivalent). But if you have a hammer, that hammer will not help you much if you need to cut a board in half. For that you need a saw (or the functional equivalent). And, of course, the saw would not have helped to drive the nail. So it is with the tools or methods of the social and behavioral sciences.

All research methods should be regarded as bounded opportunities to gain knowledge about some set of phenomena, some substantive domain. Knowledge in science is based on use of some degree of substantiation, construction and methods. The notion of that knowledge, and the confidence we can have in it, are both contingent on the methods by which it was obtained. All methods used to gather and to analyze evidence offer both opportunities not available with other methods, and limitations inherent in the use of those particular methods.

One good example of this dual nature of methods—both opportunities for gaining knowledge and limitations to that knowledge—is the widespread use of questionnaires and other forms of self-report in behavioral and social research. On the one hand, self-report measures (questionnaires, interviews, rating scales, and the like) are a direct way, and sometimes the only apparent way, to get evidence about certain kinds of variables that are worthy of study: attitudes, feelings, memories, perceptions, anticipations, goals, values, and the like. On the other hand, such self-report measures have some serious flaws. For example: Respondents may try to appear competent, to be consistent, to answer in socially desirable ways, to please (or frustrate) the researcher. Sometimes respondents are reactive on such self-report measures without even being aware of it. These flaws limit, and potentially distort, the information that can be gained from such self-report measures. Other approaches to data collection, such as observation of visible behavior, may be difficult or impossible to use when studying particular kinds of variables. For example: How do you measure sadness, or some other emotion? In any case, while such methods may avoid some of the particular weaknesses of self-reports, those methods will have other weaknesses.

Such is the dilemma of empirical science: All methods have inherent flaws, though each has certain potential advantages. You cannot avoid those flaws; but you can bring more than one approach, more than one method, to bear on each aspect of a problem. If you only use one method, there is no way to separate out the part that is the “true” measure of the concept in question from the part that reflects merely the method itself. All other methods (if they are to be of any use at all) will add strength and weaknesses, the methods can add strength to one another by offsetting each other’s weaknesses.

much doubt we may have about the two sets of evidence depends on what else is known about the problem and the methods from still other studies. On the other hand, if all of the judgments are based on only these methods, then that body of information is very much contingent on, and limited by, the flaws of those methods. Such a body of information must be regarded with some skepticism until you know whether it holds for a broader array of methods.

It should be noted, however, that no one investigator is apt to be trained in the use of all methods, nor to have access to the resources needed for all of them. For example, some researchers have access to use of expensive and well-equipped laboratory facilities and are well versed in the use of those facilities but do not have ready access to the resources needed for a full scale sample survey, or for an elaborate field study. Other researchers may be in the reverse situation, with poor or no laboratory facilities but ready access to extensive survey facilities and field study opportunities. All these factors must be considered in any evaluation of the evidence. Given researcher abilities and access to all methods on his or her research problem, but rather that the field as a whole make such use of diverse methods on each of its key problem areas. The fundamental principle, in behavioral and social science is that credible knowledge requires consistency or convergence of evidence across studies based on different methods. These issues and their implications for behavioral and social science are discussed further in the parts of this chapter to follow, along with more detailed descriptions of strategies, comparison techniques, designs and methods.

RESEARCH STRATEGIES: CHOOSING A SETTING FOR A STUDY

Research evidence, in the social and behavioral sciences, always involves somebody doing something, in some situation. We can always ask about three facets. Who (which individuals? in which context? and where (which environment)? The terms "actor," "behavior," and "context" are used here as technical terms with meanings somewhat different from ordinary usage. Actor refers to those human systems, at whatever level of aggregation (e.g., individuals, groups, organizations) that are the subjects of investigation. Behavior refers to all aspects of the states and actions of those human systems that might be of interest for such study. Context refers to all the relevant temporal, locational and situational features of the "surround" within which those human systems are embedded.

When you gather a batch of research evidence, you are always trying to maximize three desirable features or criteria:

A. Generalizability of the evidence over the populations of Actors.
B. Precision of measurement of the behaviors that are being studied (and precision of control over extraneous factors that are not being studied).
C. Realism of the situation or Context within which the evidence is gathered, in relation to the contexts in which the behaviors of interest are likely to be studied.

Although you always want to maximize all three of these criteria, A, B, and C usually, you cannot do so. This is one fundamental dilemma of the research process. The very things you can do to increase one of these few features reduce one or both

ty; techniques for statistical control by which you try to nullify the effects of varia-
tions in a given property within a study by "removing" those variations by statistical
means; and techniques for distributing the impact of a number of features of the sys-
tem and its context—without directly manipulating or controlling any one of them—so
that such impact can be taken into account in a interpretation of results. The most promi-
nent means for distributing impact of a number of features is called randomization, and
refers to procedures for the allocation of "tases" among various conditions within the
study. Those Modes for dealing with various features of the human systems to be stud-
ed—measuring, manipulating, controlling and distributing impact—are the basic
set of elements or "tools" by which social and behavioral scientists systematically
gather empirical information.

The methodological dilemma have to do with the application of various
Comparison Techniques. These are methods or techniques by means of which the
researcher can assess relations among the values of two or more features of the human
system under study. Such comparisons involve three sets of features of the systems
under study: (a) the features that have been measured, and that are regarded as mea-
sures of the phenomena of interest (these are sometimes called "dependent variables"); (b) the features that have been measured and manipulated, and that are regarded as poten-
tial covariates of, or antecedents to, the phenomena of interest (these are sometimes
called "independent variables"); and (c) all of the other features of the system that are
relevant to the relations of interest (between dependent and independent variables), and
that you have (or have failed to) control, or whose impact you have (or have failed to)
distribute or otherwise take into account. Generally, comparision assess the consistency of the asso-
ciation between the values of the first two sets (the dependent and independent vari-
bles), against the backdrop of the third set (i.e., other relevant features that were not
studied directly but that nevertheless are a part of the meaning of results).

Most of the rest of this chapter will deal with features of the research process that
emphasize the methodological domain, without much systematic consideration of other
conceptual or substantive matters. The reader should keep in mind, though, that the
research enterprise, like a well-regulated school, always depends on materials from all
domains—every domain, ideas, and techniques.
of the other two. For example: The things you can do to try to increase the precision with which you can measure behavior and control related variables (B) (for example, conducting a carefully controlled laboratory experiment) will intrude upon the situation and reduce its "naturalness" or realism (that is, reduce C), and will also reduce the range of actors (A) to whom the findings can be generalized. Conversely, the things you can do to try to keep high realism of context (C) (for example, conducting a field study in a natural situation) will reduce both the range of populations to which your results can be applied (A) and the precision of the information you generate (B). As a third example, the things you can do to try to establish a high degree of generalizability over actors (A) (for example, conducting a well-designed sample survey) will reduce realism (C) by obtaining the measures out of context, and will reduce precision (B) both by having measures of only a limited number of behaviors, and by failing to control or otherwise take into account extraneous factors that may affect results.

You can appreciate this dilemma better by examining some of the major research strategies used in the behavioral and social sciences. Figure 2 shows a set of eight alternative research strategies, or settings for gathering research information. In that figure, the eight strategies are shown as lying in a circular arrangement in relation to two underlying dimensions: the degree to which the setting used in the strategy is universal or abstract vs. particular or concrete; and the degree to which the strategy involves procedures that are obtrusive, vs. procedures that are not obtrusive, with respect to the ongoing human systems (the actor-behavior-context units) that are to be the object of study. The four strategies on the right side of the circle involve fairly concrete or particularistic procedures, while the four on the left side use fairly universal or abstract procedures. The procedures used in the four strategies in the lower half of the circle can be fairly unobtrusive. The four strategies in the top half of the circle necessarily use procedures that are fairly obtrusive, that is, they disturb the ongoing human systems (the actor-behavior-context units) that are being studied.

Figure 2 also shows where, among the strategies, each of the three desired features, or criteria is at its maximum. Criterion A, generalizability with respect to the populations of Actors, is potentially maximized in the sample survey and in formal theory. Criterion B, precision with respect to measurement and control of behaviors, is potentially maximized at its maximum in the laboratory experiment and in judgment studies. Criterion C, realism of context, is potentially at its maximum in the field study. The geometry of figure 2 emphasizes the differences just discussed, namely: strategies that maximize one of these are far from the maximum point for the other two. The very same changes in research procedures that would let you move toward the maximum of any one of these criteria—A, B, or C—at the same time would move you away from the maximum point of the other two. It is not possible, in principle, to maximize all three criteria simultaneously. Thus, any one research strategy is limited in what it can achieve.

Research done by any single strategy is flawed, although the various strategies are flawed in different ways.

The eight strategies listed in Figure 2 are shown as four pairs, each occupying one quadrant of the circle. Quadrant I contains research strategies that involve observation of ongoing behavior systems under conditions as natural as possible. Quadrant II contains research strategies that are carried out in settings conceived for the purpose of the research. Quadrant III contains research strategies that involve gathering responses of participants under conditions in which the setting is mixed or made more. Quadrant IV contains research strategies that are theoretical, rather than empirical, in character. The two strategies in each of these quadrants will be described and illustrated briefly in the following paragraphs.

QUADRANT I: THE FIELD STRATEGIES

The two research strategies in quadrant I are the Field Study and the Field Experiment. In a field study, the researcher sets out to make direct observations of natural, ongoing systems, while intruding on and disturbing those systems as little as possible. Much of the ethnographic work in cultural anthropology would exemplify this strategy, as would many field studies in sociology and many "case studies" of organizations.

A field experiment is a compromise strategy in which the researcher gives up some of the unobtrusiveness of the plain field study, in the interest of gaining more precision in the information resulting from the study. Typically, a field experiment also works within an ongoing natural system as unobtrusively as possible, except for intruding on that system by manipulating one major feature of that system. Field experiments use a manipulation of one important feature of the system in order to be able to assess the causal effects of the differences in that manipulated feature on other behaviors of the system. Many types of work organizations, such as the famous Western Electric or Hawthorne studies (Roethliger & Dickson, 1939), would exemplify the field experiment. Such studies introduce a major change in one feature of the organization (for example, a change in the formal communication structure), and study the changes that occur elsewhere in the organization subsequently. Sometimes such research also studies an unchanged but otherwise comparable organization, as a basis for comparison.

The essence of both of the strategies in quadrant I, the field study and the field experiment, is the same in so far as the system is studied as it is, in the sense that it would occur whether or not the researcher were there and whether or not they were being observed as part of a study. The two strategies of quadrant I differ in that the field study remains as unobtrusive as it can be (although no study is ever completely unobtrusive), at a cost in the ability to make strong interpretations of resulting evidence; whereas the field experiment attempts to gain the ability to make stronger interpretations of some of the results (for example, that a behavior difference associated with the experimental manipulation may have been caused by the experimental variables involved in the manipulation), but does so at a cost in obtrusiveness, hence in the naturalness or realism of the context.

QUADRANT II: THE EXPERIMENTAL STRATEGIES

The best known of the two strategies in quadrant II is the laboratory experiment. In that strategy, the investigator deliberately conceives a situation or behavior setting or context, defines the rules for its operation, and then induces some individuals or groups to enter the conceived system and engage in the behaviors called for by its rules and circumstances. In this way, the researcher is able to study the behaviors of interest with considerable precision (e.g., the investigator can be better prepared to measure certain behaviors because he or she can be confident about where and when those behaviors will occur), and to do so under conditions where many extraneous factors (that might be important but that are beyond the scope of the researcher's present interest) have been eliminated or brought under experimental control. The potential gain in precision in the measurement and control of behavior, which is the lure of the laboratory experiment, is paid for by increased obtrusiveness, hence decreased realism of context, and by a narrowing of the range of potential generalizability of results.

The other strategy of Quadrant II is the experimental simulation. In this strategy, the researcher attempts to achieve much of the precision and control of the laboratory experiment but to gain realism by concocting a situation or behavior setting or context, as in the laboratory experiment, but making it as much like some class of actual behavior setting as possible. One example would be research using ground-based flight simulators such as those used by the U.S. Air Force and commercial airlines to train pilots for instrument flying. Another would be research that uses auto-driving simulators like those sometimes used to train nephew drivers. Still another would be research using military training exercises, or involving franzo-social practice games by an athletic team. Still another could be a monopoly game, or a strategy game, or other similar board game, if they were used for research purposes and with some degree of control over "extraneous variables".

Here, the key idea is that the researcher wants to create a system under his or her control, but at the same time have that system operate in a manner that simulates the operations of some particular class of naturally occurring system—the flight of airplanes, the steering of autos, the flow of "battle" in various sorts of two-sided combat or contests, or the operations of a "marker" involving both strategic choices and chance factors.

The experimental simulation is a compromise strategy that attempts to retain the precision of the laboratory but at the same time to not give up much realism of context. risks introducing so much realism that precision of measurement and control are weakened, on the one hand, or retaining so much control that it becomes "artificial" as the laboratory experiment on the other hand. An example of the former would be use of a military training exercise, in which the opposing "armies" are allowed to carry out their missions, anywhere and in any order, and thus make it impossible to observe and record the action for research purposes. An example of the latter would be to make such a "combat exercise" so stylized, and simplified in its flow—in the interest of good measurement and control—that all of the "realism" is nullified—that is, the system actually operating in the study does not function like the systems supposedly being simulated (that is, actual combat). The two strategies in Quadrant II, in contrast to those of Quadrant I, involve concocted rather than natural settings. That is, the laboratory experiment and the experimental simulation are strategies that involve "actor-behavior-context" systems that
would not exist at all were it not for the researcher's interest in doing the study. The distinction here is not between "real" and "unreal." The context of the laboratory experiment and the experimental simulation are certainly "real." For the participants once they are in the lab or simulation chamber, and the behaviors performed by the participants are certainly "real." Participants' behaviors are undoubtedly influenced by features of the experimental setting, but that is also the case in any other setting, natural or concocted. In fact, the exploration of such situational influences is, in large part, the point of naturalistic and social science research.

The distinction here, between the field research of quadrant I and the experimental research of Quadrant II, has to do with whether the situation exists prior to and independently of the investigator, versus being having been constructed by the researcher; and therefore whether the participants are taking part in an ongoing manipulation or are part of a research endeavor. The issue is not one of reality, although much discussion of research in social sciences mistakenly treats it as such. Rather, the issue is one of the construction of behavior systems in the behavior system under study.

Note that the difference between adjacent strategies are matters of degree. You can find experimental simulations (for example, varieties of strategy games) for which the task is so abstract that it becomes very close to a laboratory experiment. It also should be noted that few studies are "pure" examples of one strategy. These types of strategies represent a set of possibilities for carrying out research, rather than a description of concrete studies.

Q U A D R A N T  I I I :  T H E  R E S P O N D E N T  S T R A T E G I E S

In a sample survey, the investigator tries to obtain evidence that will permit him or her to estimate the distribution of some variables, and/or some relationships among them, within a specified population. This is done, typically, by careful sampling of actors from that population (thus potentially gaining a lot of generalizability, criterion A), and by systematically eliciting responses from those selected actors about the matters of interest. The many public opinion surveys on voting intentions, political preferences, voting intentions, and the like exemplify this strategy. While there is much emphasis on sampling, the researcher is seeking to control a minimum of variables and often little opportunity for much precision of measurement. Hence, this strategy is low on criterion B. And since the responses are gathered under conditions that make the behavior setting irrelevant, the question of realism of context is made moot (hence, this strategy is low on criterion C).

In a judgment study, the researcher concentrates on obtaining information about the properties of a certain set of stimulus materials, usually arranged so that they systematically reflect various instantiation of some variable (hence it is low on criterion B). This strategy would be exemplified by any of the various general theories in behavioral and social sciences. Such theories are based on earlier empirical evidence (it is to be hoped, and they lead to the generation of hypotheses that are representative of or empirical — that is, it does not involve any "actors behaving in context".

The other non-empirical strategy in Quadrant IV is called Computer Simulation. It is like the experimental simulation strategy of quadrant II in that the researcher attempts to model some general system (for example, a battle, a market, an aircraft in flight). But it is quite different from that strategy too. The computer simulation is a complete and closed system that models the operation of the concrete system without any behavior by any system participants (hence it is low on criterion A). Thus, the model is a complete and logical closed. All the important components of the system are specified by the investigator, and as such all of the relations among those components. Then, when the researcher starts a "run" of the system, all that ensues is the relatively predictable resultant of features built into the system. Such models are based on behavior in the sense that they must have all behavior parameters specified in advance, and this is often done on the basis of evidence from prior empirical research (at least for the parts about which the investigator has some conjecture). So this strategy is very low on criterion C. At some time, it is potentially high on criterion B, in the sense that it is an attempt to model some concrete class of real world system (such as the geographic processes going on in connection with the eruption of Mount St. Helens, for example). Computer simulations are then a special type of Superstrategy. But a computer simulation is designed to model some particular class of systems; so the model is likely to have little generality over populations of actors or situations — or, more accurately, the question of generality over populations is moot. So this strategy is low on criterion A, because the to the researcher has as its goal that is a necessary underpinning for any science. Inclusion of these two strategies also gives us the opportunity to note that one of the more powerful general strategies for research, and one that involves the use of multiple strategies on the same problem, is the simultaneous and joint use of one of the behavioral strategies (say, a normative general theory) and one or more of the empirical strategies (for example, a laboratory experiment).

S O M E  S T R A T E G I E S

Within the other chapters of this book, you will find studies done by all or most of the strategies discussed here. You should view that substantive material with two substantive strategies in mind: First, each strategy can contribute weaknesses, although each also has certain potential strengths. These weaknesses and strengths become part of the meaning of any evidence gathered with those strategies. So, an adequate interpretation of the available evidence on any given topic or problem should take those methodological strengths and weaknesses into account. The first strategy is low on criterion B, and the second strategy is low on criterion A, therefore the combination of both is required to understand the data, and is the reason why you have been asked to read this book. The second strategic issue you are encouraged to address, with regard to the substantive material presented in the rest of this book, therefore is: To what extent is the research evidence on each problem or topic based on use of only a single research strategy and therefore limited by the weaknesses of that strategy; and to what extent is that body of evidence based on use of multiple, complementary strategies, with agreement or congruence among the findings strengthing the validity of the strategies here? The answers to the two issues are important indicators of how the research interaction has become a viable science with a cumulative body of credibly interpretable evidence.

S T U D Y  E X I S T ,  C O M P A R I S O N  T E C H N I Q U E S ,  A N D  V A L I D I T Y

In every empirical study, observations must be gathered, those observations must be aggregated and purified, and some comparisons must be made within that set of data. The comparisons to be made are the heart of the research. They reflect the relations that are the central focus of the study. What comparisons are included in the study at the element level from all three domains are: (a) what was being compared, (b) what comparisons in several domains (what phenomena, what properties, what modes for treatment of variables have been used), (c) what is being compared is true or false (this is a substantive system is being studied, what you are examining is a substantial variable system of study being drawn upon), (d) what conceptual relations have been posted for the phenomena of interest (e.g., a certain pair of properties, X and Y, are causally
linked, with X causing Y; and, especially, (d) what comparison techniques are available, within the methodological domain, to ask such relational questions. This section will deal with some general features of the comparison techniques that are most commonly used within the current methodology of the social and behavioral sciences.

**COMPARISON TECHNIQUES: ASSESSING ASSOCIATIONS AND DIFFERENCES**

All research questions can be boiled down to variations of a few basic forms: bivariate, correlations, and differences. The bivariate question asks: How often (at what rate, or what percentage of the cases) is X is it often a very crucial underpinning to the interpretation of other information. A second general form of comparison question, and one that has been given far more attention for the study of human groups and the other organizations in which social relations occur together? That relational question has two major forms, which together subsume most of the questions that are asked in behavioral and social science research: the correlation or covariance question, and the comparison or difference question.

Beyond the occasional interest in the use of correlation, there is a widespread and common basis for deciding whether the rate of Y in some particular case is or is not "notably" high or low.

For example, some researchers recently found that there was a surprisingly high rate of birth defects among fetuses born to women who worked at involving contact of video display tubes. One set of people (Nine to Five, an organization concerned with the rights and well-being of working women) interpreted data as indicating that video display tubes represented a serious health hazard, at least for pregnant women. Another set of people (agents of the organization whose women workers had shown such high rates of birth defects in their pregnancies) inter- preted the same numbers as being indicative of any hazard, arguing that we do know that the rates of birth defects for the pregnancies of working women — of the same age, social class and so for that matter — are not very different. In each case the exposure variables were not controlled by the two groups, not surprisingly, were in sharp contrast to each other. Nine to Five argued some policies that would reduce the exposure of pregnant women to video displays and would probably be job-related at the same time. The management group argued "more research" — presumably in pursuit of the missing bureau information and perhaps in exploration of possible- ings from the video effects — but no other changes in working conditions.

Such differences in interpretation of the same evidence are pervasive throughout the behavioral and social sciences. They are especially apparent in research dealing with various political, economic, and social issues. Those same kinds of disagreements also confront the physical and biological sciences: How much exposure to radiation is "acceptable"? What are tolerable levels of exposure to asbestos, dioxin, and many other environmental contaminants? And, are we looking really "cause" cancer or heart disease, or is the evidence "merely correlation"?

Furthermore, the obvious influence of self-interest on those interpretations holds equally for the so-called "exact sciences" of physics, chemistry, physiology and the like, as well as for the social sciences, which may have an even greater impact.

**Methodological Factors:**

**Randomization and True Experiments**

You can only measure, match, control and manipulate a limited number of variables in any one study, and there are usually many more factors that are potentially important to the phenomena you are studying. You have to do your best with the rest of that rather large set of potentially relevant factors. The main "something else" that you can do is call Randomization, or random assignment of cases to conditions.

Randomization means using a random assignment procedure to allocate "cases" to "conditions." In the above example, you would use the results for assigning individuals in your sample to groups that were to be high in liking and those that were to be low in liking (and to large and small groups of each kind). For an allocation procedure to be random, each case must be equally likely to end up in any given combination of conditions. Using the previous example, for instance, we would say that any given individual is equally likely to be in a "high liking small size condition," a "high liking large size condition," a "low liking small size condition," or a "low liking large size condition." (You must take into account, of course, the difference in numbers of individuals that is used for different conditions, and of differences in the two factors at the same time. For example, you might want to assign different group liking had more of an effect on task performance for small groups than for larger ones. (Questions about the joint effects of differences in Y and X are often referred to in the research literature as "interaction effects").

**Correlation or Covariation**

Correlation or covariation could be just as strong, but more complicated in its form, than for the simple linear case. There are a number of statistical tools that allow the researcher to investigate nonlinear, as well as linear, correlations. Unfortunately, behavioral and social sciences researchers are especially prone to use the former when they should be examining the latter. As the shape of the relation becomes more complicated — e.g., suppose on the average happiness decreased from young child to adolescent, then increased up to about age 45, then decreased again, but flattened out for the over our statistical techniques become more cumbersome to use, and fewer of them are appropriate for the task of assessing such complex forms of relation.

Much research in the social and behavioral sciences makes use of correlations, linear and nonlinear, that involve two, three, or more independent variables. Such multivariate analysis approach requires using measures of X, Y and (of the other variables, if more than two are involved), for a series of "cases" that vary on X and Y (and on the other variables involved). "Cases," here, mean "across behaviors in context," as discussed in the first example. Such approaches are useful for getting a measure of age and of happiness for each of a series of individuals who make up the sample of a given study. The correlation between these two sets of values can tell you whether X and Y go together; but it cannot help you decide whether X is a cause of Y, or vice versa, or both, or neither. That is to say, the correlation comparison techniques cannot assess conceptual relations that imply covariation between (or more variables; but they cannot assess any conceptual relations that are causal in their implications.

**The Difference Question.** Another form of the relational question is the comparison or difference question. The difference question involves asking, essentially, whether Y is present (or at a high value) under conditions where X is present (or at a high value); and whether Y is absent (or at a low value) when X is absent (or low). For example, do groups perform assigned tasks better (Y) when members like each other (X) than when they do not.

You could approach the assessment of this question in either of two ways. One way would be to go out and collect measures of "liking" in groups until you had gathered that a bunch that were being liked (and perhaps a bunch at intermediate levels), and then compare the average task performance scores for those bunches of groups. That kind of study would be, in effect, a messy version of the correlational approach — one that gave up much of the advantage of being able to make a stronger interpretation of causation direction in the results.

A more useful approach to the comparison question would be: To create some groups with partners who do like each other, and other groups with partners who do not like each other; then, to give both groups some common tasks to perform; and, then, to see if the average task performance (Y) of the "high liking" groups (X) is higher than the average task performance of the "low liking" groups (not X). For the comparison to be most useful, you would have to make sure that the two groups were the same, or comparable, on all of the other factors that might affect task performance, such as the difficulty of the task, availability of task materials, quality of working conditions, task-related abilities of members, experience and training of members, and many other environmental factors that sustained reality "cause" cancer or heart disease, or is the evidence "merely correlation"?

Furthermore, the obvious influence of self-interest on those interpretations holds equally for the so-called "exact sciences" of physics, chemistry, physiology and the like, as well as for the social sciences, which may have an even greater impact.

**Methodological Factors:**

**Randomization and True Experiments**

You can only measure, match, control and manipulate a limited number of variables in any one study, and there are usually many more factors that are potentially important to the phenomena you are studying. You have to do your best with the rest of that rather large set of potentially relevant factors. The main "something else" that you can do is call Randomization, or random assignment of cases to conditions.

Randomization means using a random assignment procedure to allocate "cases" to "conditions." In the above example, you would use the results for assigning individuals in your sample to groups that were to be high in liking and those that were to be low in liking (and to large and small groups of each kind). For an allocation procedure to be random, each case must be equally likely to end up in any given combination of conditions. Using the previous example, for instance, we would say that any given individual is equally likely to be in a "high liking small size condition," a "high liking large size condition," a "low liking small size condition," or a "low liking large size condition." (You must take into account, of course, the difference in numbers of individuals that is used for different conditions, and of differences in the two factors at the same time. For example, you might want to assign different group liking had more of an effect on task performance for small groups than for larger ones. (Questions about the joint effects of differences in Y and X are often referred to in the research literature as "interaction effects").
is not likely (though it is possible) that the high X cases were all high on some extraneous factor that caused high X, while the low X cases were all low on that same factor. For example, if you assigned individuals to "tall liking" and "low liking" groups by a random procedure, as in the example given above, and then those groups differed later in their average task performance scores, it is only 1 in 100 times, but not impossible) that they differed because all of the people who happened to end up in the "high" condition on a chance basis had more average task ability, while all the people who, by chance, were assigned to the "low" groups happened to be below average in task ability.

Note that the likelihood or probability of obtaining a particular correlation, not logical certainty. A random allocation procedure does not guarantee an equal distribution of any, let alone all, of the potential extraneous factors among the conditions being compared. Rather, a randomization procedure makes a highly unequal distribution on any one of them highly unlikely (but not impossible). So the reasoning from even a true experiment involves inductive rather than deductive logic, probability rather than certainty.

The effectiveness of randomization for actually rendering the sets of cases in differ-
ent conditions "not different" from one another depends on the number of cases being allocated. Furthermore, you can never know for sure that some one particular factor was not end not up — by the luck of the draw — quite mal-distributed across the conditions of your study. If such a factor is operating, and is also related to the phenomenon you are studying, it will distort your results and you will have no way to know it. This is often called "confounding." Of course, if you had not used a random allocation procedure, such a-mal-distribution across conditions would almost certainly have occurred, per-
haps several variables. You would also not know which ones or in what directions they were getting "confounded" with your results. (Confounding in research evidence is like noise in a com-
munication system. The more noise that is present, the more likely it is that the "sig-
nal" or "true effect" will be obscured or "distorted.")

You can see that true experiments are potentially powerful techniques for learning about causal relations among variables. But, as in all aspects of research methodology, you buy this high power at a high price in two ways. First, you reduce the scope of your study, since you can only vary variables constant, and instead vary only a few variables (your total variables (your X) occur only at a few levels (e.g. high vs. low liking, or 3 person vs. 5 person groups). The results of your study will thereby be limited in the range of variations over which the findings can be generalized. Second, you reduce the realism of your study, since each subject almost naturally creates the groups, the designs, the tasks, and the elicited behavior that serves your — not the participants — purposes.

**Sampling, Allocation and Statistical Inference**

The basis you use for choosing the cases that are to be included in your study, out of a larger population of potential cases, also has a substantial effect on the credibility of the evidence resulting from your study. Most of the ways that social scientists have
to assess correlations and differences rely on statistical reasoning that requires that the cases in the study be a "random sample" of the population to which the results apply. So, your results really apply to that population of which your cases are a random sample. For example, suppose you chose cases by talking to all the people who left a dining hall by a certain door, starting at 12:30 Wednesday. Your results, strictly speaking, would apply to a population that does not include: people who do not eat at that dining hall, people who leave by another door, people who eat quickly, people who have a Wednesday class that goes through lunch hour, and, of course, people who refuse to answer "interviewers" who stop them in public places. The question is not whether you have a random sample or not, but whether the questions for selecting cases, what is the nature of the population of which you actually have a random sample? Is it to that population, and only that one, that your results apply?

There is sometimes a confusion in the literature about random in randomizing in how one goes about choosing a sample of cases from a population, and in discussing how one allocates cases (already selected to be in the study) among the conditions of that study. Both selection and allocation of cases require that there be a random component in the proce-
dure. In the sampling component, the random component is designed to describe which cases, out of some larger population, will be included in a given study. In the allocation case, the pro-
cEDURE is designed to determine which conditions each given case — already selected as part of the study — will be assigned to. The two are alike in that, for both population sam-
pling and allocation of cases to conditions, the term random refers to a procedure, not an outcome. You do not actually "select a random sample." You select a sample by using "a random procedure". There is no guarantee that the resulting sample will be a perfect mirror of the population. That is, you have no guarantee that your random sample (the sample you select with a random procedure) will yield a representative sample (that is, a distribution of cases that mirrors the population from which you sampled). Similarly, you do not actu-
ally "allocate a random set of cases to each condition." You allocate cases among condi-
tions by using a "random procedure." There is no guarantee that the sub-samples in fact be comparable in any respect. But in both cases — sampling from a population and allocating cases to conditions within your study — using a random procedure is your best bet. That is, if you use a random procedure to sample from a population or to allocate cases gives you the best chance that the results in your study will be representative, and that the result allocation of cases to conditions will be unbiased.

The preceding discussion also suggests one reason why the size of samples used (the numbers of cases in each condition) is important in the credibility of exper-
imental results. The larger the number of things to be allocated by some random procedure, the more difficult the distribution of those cases will approach the "ideal-
ized" random distribution. For example, the probability that "heads" or "tails" will result from the flip of a die of a given number of times is 50-50. But because of the size of the flip it would not be particularly surprising if the distribution of 10 flips was other than 5 to 5 — say, 6 to 4 or even 7 to 3. As the number of coin flips is increased (assuming, of course, an honest coin and an honest flip), the more the distribution of heads and tails
determined end up in each of the samples. Just as we can estimate the probability of getting a distribution of coin flips that differs a certain amount from the idealized "random dis-
tribution" of coin flips, given that we know the number of coin flips involved, so we can estimate the probability of getting any given deviation from the expected, Y) between two groups of cases — if only chance factors were operating. When such calculations indicate that a difference as big as the one we actually obtained in our study would only rarely occur if chance alone had been operating, we are justified in concluding that something other than chance is probably involved in the difference (in average Y values) between the group of cases that did have X, and the group that did not have X.

Since we can state the probability of the result (that is, the proportion of times it would have occurred) if only chance were operating, that probability value is a rela-
tively precise and quantitative estimate of how confident we can be that something other than chance was at work. But that probability value in and of itself does not help us determine what that "something other than chance" was. It is then up to the researcher to find explanations for that result that make sense in the context of the data obtained. These explanations are sometimes referred to as plausible rival hypotheses (e.g., that the hypothesis to the presence or absence of X is the main causal factor).

**Validity of Findings**

The idea of validity is central to the research process, yet it is a diffuse concept. One quite comprehensive discussion of validity issues (Cook & Campbell, 1979) proposes four different kinds of validity commonly used to assess construct validity, and external validity. All four are discussed in this section, along with other related considerations.

**Internal validity** has to do with the degree to which results of a study permit you to make strong inferences about the causal relations. That is, how close can you come to asserting that the presence of X (or variations in level of X) caused the altered level of Y values? From the preceding discussion, it should be clear that the more existence of a difference in average Y across levels of X that we observe, the more valid is our conclu-
sion (or that X had a causal effect). But just because a correlation coefficient is present (or was high vs. low) is not a sufficient basis for the conclusion that X caused Y. For one thing, the difference might have arisen just by chance. Some of the consid-
erations involved in determining the non-chance basis of a finding (e.g., sample size) are discussed in the following section. Another important issue that is involved in a study's statistical conclusion validity, which has to do with
whether a given result (such as a difference in $Y$ associated with a difference in $X$) is to be regarded as not due to chance.

There are other reasons why a difference in $Y$ associated with a difference in $X$ does not necessarily imply a causal role for $X$. Some other variables might have been covarying with $X$ and $Y$, and they, rather than $X$, might have produced the change in $Y$. Any such factors, that were neither measured, manipulated, held constant, nor matched across groups in your study, is a candidate for the role of "other variable" that is a plausible alternative or rival hypothesis about the cause of the difference in $Y$. If your study included a random sample of people, then the conditions under which such confounding can help rule out many such plausible rival hypotheses about the cause of differences in $Y$.

How many and which rival hypotheses can be ruled out in any given case depends on the transformation of the experimental design, controls, matching, and other features carried out in your study. The internal validity of your study’s findings depends on how well you can rule out — by the logic of your procedures as well as through certain comparisons in your results — all of the plausible rival hypotheses.

Generally speaking: Have you defined a set of conditions under which your study is valid? How clearly understood are the conceptual relations being explored? How clearly specified are the mappings of those concepts and relations to the substantive and operational procedures by which they are to be accomplished? This form of validity obviously is related to the “fit” of elements and relations from the conceptual domain with those from the other two domains. Just as problems arise in this context depend on which of several alternative study paths are followed in a given case.

External validity refers to whether the findings of your study will hold up under replication, and how confidently you can extend the results of the study over which your findings will hold and the limits beyond which they will not hold. Obviously, some features of a study have a direct bearing on whether that study’s findings are likely to prove generalizable: the site, nature, and mode of selection of the sample of cases used in the study; the degree to which the study involved relatively artificial, versus relatively natural, settings and procedures; and the like. Nevertheless, determining the generalizability of any particular set of findings in any definitive sense requires conducting one or more follow-up studies. No single test of generalizability is valid or invalid, in and of itself. Logical studies may shed some light on the generalizability of its findings; and may have shown some light on the robustness of findings from prior studies.

Threats to validity are ever-present. If Campbell and his colleagues have developed an excellent list of more than thirty major classes of plausible rival hypotheses that are potential threats to these four forms of validity. Which of these different classes of threats to validity are most problematic depends on which type of study design is being used, and the way in which that design provides the statistical context for the study.

Campbell and associates also have developed a classification of some 31 major types of study designs that have been or can be used in the study of a variety of behavioral and social science topics, and indicated which sets of plausible rival hypotheses are and are not frequent problems for each type of design. Some of the design types are

"true experiments" in the sense discussed in the preceding section. Some strong infer-
ences about the $X$-$Y$ relation potentially can be made from these, although even these true experiments do not by any means eliminate all plausible rival hypotheses. Some of the design types are what Campbell and colleagues call "pre-experimental" designs, meaning that they fail to cope with a very large number of potential rival hypotheses. Some of the design types are what Cook and Campbell call "quasi-experimental" designs. As the name implies, these have some but not all the virtues of the "true experi-
m ent." For example, they may have no randomization, but specifies, basis for allocation of participants to conditions. With regard to true experiments, these designs can deal effective-

ly with some, but not nearly so many, of the plausible rival hypotheses, they permit some, though weaker, inferences about the causal status of the $X$-$Y$ relation. Besides these issues in study design, and the issues related to research strategies that were discussed earlier, the design needs to take into account design features that are relevant to the "true experimental" hypoth-

ics that are, or could be, available for measuring, manipulating, controlling or oth-

erwise treating key properties of the human systems that are the focus of our behavioral and social science studies. Some features of these operational level techniques for the "Molds" of Treatment of variables are discussed in the next section of this chapter.

CLASSES OF MEASURES AND MANIPULATION TECHNIQUES

POTENTIAL CLASSES OF MEASURES IN SOCIAL PSYCHOLOGY

Social and behavioral scientists have used a wide variety of techniques to measure the presence or absence of the specific features of the human systems that they wish to study. By far the most widely used type of measure involves questionnaires or other forms of self-

report, of some whose main strengths and weaknesses were noted earlier in this chapter. But researchers have invented a number of other approaches, some of which offset the weakness-

ess of self-reports but, of course, do so at the cost of incurring other weaknesses. Campbell and colleagues have provided a useful access to some of these -- which may be either methodological or methodological, and indicated their major strengths and weaknesses (See Webb, Campbell, Schwartz & Sechrest, 1966). That schema has been extended and elaborated by McGrath and colleagues (See Sackett & McGrath, 1979; McGrath, 1984; McGrath, Martin & Bozak, 1982; Emberg & McGrath, 1985). A brief and simplified form of these ideas will be presented here.

Whenever an investigator wants to obtain a measure of some feature of a system being studied he or she must somehow arrange for a record of that feature to be made for each subject (or to be observed for each subject at every time). The objective is to have access to it later. The information contained in the record is always about the human system being studied (the actor-having-in-context, whether that is an individual, a group, or whatever). And it is always to be used by the investigator (that is, scored, aggregated with other scores, used in computer programs, etc.). But the record of it can be made by any one of three parties: by the investigator, whose behavior is the focus of study (or some representative of the actor when that is a multi-person unit); by the investigator who is conducting the study (for some person or institution serving as

sargeant of the investigation); or by some external third party who is not involved in the research and who makes a record of the behavior for some other purpose (e.g., records of attendance made for administrative purposes).

When such a record of behavior is made — by any of the three recording agents, partic-

ipants, investigators, or some external third party — there are three basic categories of what the behaviors being recorded are aware that the recording process is taking place and that those records will or may later be used for research (or other quasi-public) pur-

poses. When the actors are aware that their behaviors are or may be being recorded, we have a form of recording procedure that is quite apart from any of the aspects of the research methodology, such as use of field or experimental strategies, and quite different from the state of affairs in nonhuman sciences (e.g., physics, chemistry, biology). The investigator must take that potential "unamnessness" of behavior, that has been induced by the measuring procedure, into account in scoring (i.e., use scores, aggregated, analyzes, interprets) that evidence. This problem is sometimes referred to as the reactivity of measures. It is one major way in which social and behavioral science

research often loses realism (criterion C as discussed earlier in this chapter) even when that research is done in natural or field settings.

We can use the two distinctions discussed above — "who makes a record of the behavior?" and "is the participant aware that his or her behavior is being recorded and used for research purposes?"— to classify two of the basic types of recording procedures.

"Self-Reports." Records that the participants knowingly make of their own behavior are called

Self-Reports. Records that participants unwittingly make by their behavior are called

Trace Measures. Records of behavior made by the investigator (or some agent or instru-
mament working for the investigator) are called Observations — and they may be by an observer "visible" to, or hidden from, the participants. Records of behavior by some third party, for non-research purposes, are called Archival Records — and they may be done either with the expectation that the information will be utilized for some purpose other than research or not with the expectation that it will be so utilized. The material to follow examines those six classes of data collection methods — Self-Reports, Trace Measures, Observations by a Visible Observer, Observations by a Hidden Observer, Archival Records, and Private Archives.

The first of these six classes are the self-exports of participants, always done under conditions in which the respondents know that their behavior is being recorded for research purposes. An example would be responses on a questionnaire that the participants were asked to complete.

Observations. A second way to get records of behavior is by means of observations. This term refers to records of behavior (such as a record of the sequence of speakers in a group), made directly by the investigator, or made by someone substituting for the investigator (e.g., an experimental assistant), or made by some physical instrument that

is serving the investigator (e.g., an automatic electronic counter, or a stopwatch).

Sometimes observations are made under conditions in which the participants know that they are being observed; but other times observations are made without the participants being aware of it. So it is important to carry out two forms of observations: Observations by a Visible Observer and Observations by a Hidden Observer. The accuracy of those distinctions is real between observations known to be taking place (whether the observer is literally in sight or not), versus observations that are known "visible observers" are actually out of site (e.g., working behind one-way mirrors) but their presence is known to the partic-

pants. Sometimes "hidden observers" are not literally hidden, as when data are gath-

ered by eavesdropping or by monitoring records. A third way to get records of behavior to be analytic in existing archives. These are records and documents that have been gathered and/or pre-

served by some third party, external to the research activity, presumably for reasons not related to the research purpose. The information contained in newspaper and periodicals, and the files of other communication media; or in public or orga-

nizations records of births, deaths, promotions, marriages, and the like; or in private documents such as diaries, letters and logs. None of these records were made under conditions where the actors were aware that the behavior was likely to be recorded and these records were likely to be used — not for research purposes but for administrative or political ones. For example, one would presume that the public speeches of politicians made under the expectation that they would become part of the public record. And one would presume that some form of official transactions within organizations—election or appointment to an office, attendance or absence from duty, levels of output and of expenditures — are recorded in databases that are not known to the behavior or its consequences. So these kinds of archival records should be regarded as reactive in a way similar to but not exactly the same as the reactive effects on question-

naire responses and behavior in the presence of visible observers. Other forms of archival records, though they may be known to the behavior and the letters and the like — we might presum-

be have to be made without any expectation that they would be used later for research or other quasi-public purposes (unless the source was a public figure). Solid archival material, that is, material as the material that is gathered on a national product, or the number of highway fatalities on a certain weekend — are clearly not affected, consciously or unconsciously, by the participants’ awareness that their behavior or its results will become a matter of record. So, we can identify two types of archival measures, call them records of "visible" versus records of "the Union" speeches, or of promotions in an organization) and Records of Private Behavior (e.g., behavior contributing to the birth or accident rates, or to GDP).

Trace Measures. One final type of measure has been called Trace Measures (see Wedel, et al., 1966) — records of behavior that are laid down by the behavior itself, but without the actors being aware that they are making such a record. They

include traces of the behavior that are accretions of some sort, and traces that are evidences of emotions. For example: Users of a museum inflict wear on the floor tiles. Other things being equal, there will be more wear in the paths leading to the more popular exhibits. So recorded, a walkable floor could be an unobtrusive index of their use. As still another example: The number and types of liquor bottles in the bar of a particular apartment or household community could be an indicator of drinking and other social habits of the groups. These are like self-reports, in that they are the result of "recording" done by the participants themselves. But they are unlike self-reports in that the participants are presumably not aware that their behavior is being recorded, or that it will be used for research purposes. Hence, trace measures are far less reactive than self-reports — although, of course, they are beset with a number of other weaknesses, some of which will be noted below.

STRENGTHS AND WEAKNESSES OF TYPES OF MEASURES

These six types of measures — self-reports and traces produced by the participants themselves, observations made by hidden or visible observers in the service of the investigator, and archival records of public and private behaviors gathered and preserved by third parties external to the research — subsume virtually all of the techniques by means of which social and behavioral scientists have obtained measures of the features of the "actors - behaving - in - context" that they have studied. Measures of each type have both important advantages and disadvantages for social and behavioral science researchers. As with other aspects of the research process, there is no "right" or "best" way to measure; and exclusive use of any one type of measure can compromise the value of the resulting information.

Self-reports. Self-reports include questionnaire responses, interview protocols, rating scales, paper and pencil tests. They are by far the most frequently used type of measure in behavioral and social science, and there are some very good reasons for that popularity. Self-reports are versatile, both as to the potential sources of data and as to the population to which they can be applied. One can ask questions on a self-report and thereby capture unique status variables with very little — if any — idea that one can express in words. And one can adapt such questions for use with most human except for very young children. Self-reports are relatively low in both initial senescence and in accuracy of the data that they yield. They also have low "drop rates": that is, how much information that is gathered gets discarded (something that is not always true for observations, trace measures, and archival materials). They take relatively little time to construct and to apply. But self-reports have a serious Achilles' heel: They are potentially reactive, since respondents may well be aware that their behavior is being done for the sake of the researcher's purposes. Such knowledge may influence how they respond. Participants may try to make a good impression, to give socially desirable answers, to help the researcher get the results he or she wants (or, alternately, to make the researcher's problem as unclear as possible). Such influences may or may not be deliberate or unwillingly. All self-reports are potentially fraught with potential, though self-reports are nevertheless a very useful form of evidence.

Observations. Observations by a visible observer share with self-reports the serious problem of reactivity. They also are vulnerable to observer errors that derive from the fact that both humans and physical instruments that might be used for the observation and recording of phenomena are fallible. Unlike self-reports, observations can be used only on overt behavior, not on thoughts or feelings or expectations. But within that limitation, they are relatively versatile in their contents and in the populations to which they can be applied. Relative to self-reports, observations are costly in both time and resources, and have a greater variance between 7 p.m. or a pre-post observer - hour basis. Use of a hidden observer may reduce the problem of reactivity even more, but such observations are still vulnerable to observer errors, are still costly in time and money and high in drop rates, and are generally less versatile with regard to both content and population. Furthermore, use of hidden observers raises some rather serious ethical concerns.

Trace measures. Trace measures, physical evidences of behavior left behind as unintended residue or outcroppings of past behavior, offer a sharply contrast to self-reports in both strengths and weaknesses. They are only as strong as the degree to which they are unobtrusive; they do not interfere with the ongoing flow of behavior and events, and they are not likely to be affected reactively by the participants' awareness of the role of the physical evidence in later measures. On the other hand, trace measures are not nearly so versatile as to content or population as are self-reports or observations. They are simply not available for many concepts one might wish to study. Furthermore, they are often quite loosely linked to the concepts they are alleged to measure. For example, specific kinds of trash in a garbage can (such as liquor bottles) may indicate any or all of the features of the life style of the residential: social class, gregariousness, family size, the presence of a drinking problem, and so forth. Much wear on certain floor tiles may indicate differential populatior of a certain exhibit, but it also could mark a path to the rest rooms or the museum calendars, or simply denote the part of the floor that was least recently treaded. Trace measures are often very time consuming to gather and process. They sometimes are costly, and sometimes they have a very high drop rate. Yet, their genuine attractive features — unobtrusiveness and meaningfulness — makes them a very valuable potential class of measures, though a class that has as yet seen relatively little use in social and behavioral science.

Archival records. Archival records refer to such things as census data, production records, court proceedings, medical records, military records, weather records, "gang" records, "news" or "scandal" records, and official administrative records, documents and contracts. Some of them are records of public behavior (such as political speeches, votes in a legislature), and some are records of private behavior (such as bank records or the newspaper, magazine, and other written materials of an individual). They are not subject to the same kinds of questionnaires or as behavior in the presence of a visible observer. Others are records of essentially private behavior (such as birth rates, records of consumer purchases, and the like), and would seem to be as free of reactive biases as are trace measures or data from hidden observers. Both kinds are like trace measures in some of their vulnerabilities:

relatively low versatility of content and population; relatively high drop rates; sometimes only a loose linkage between the record and the concept to be represented by it. They are often far less costly than trace measures, since someone else has already gathered and preserved the data. They are relatively easy to store; and the data are preserved. The records are stored in terms used by Webb, et al. (1966), when you use a set of archival records in research, they are "methodologically consumed" — meaning that there is no opportunity to "cross-validate" your findings by getting more data from another member of the variable cases. They have low "drop rates": that is, how much information that is gathered gets discarded (something that is not always true for observations, trace measures, and archival materials). They take relatively little time to construct and to apply. But self-reports have a serious Achilles' heel: They are potentially reactive, since respondents may well be aware that their behavior is being done for the sake of the researcher's purposes. Such knowledge may influence how they respond. Participants may try to make a good impression, to give socially desirable answers, to help the researcher get the results he or she wants (or, alternately, to make the researcher's problem as unclear as possible). Such influences may or may not be deliberate or unwillingly. All self-reports are potentially fraught with potential, though self-reports are nevertheless a very useful form of evidence.

Concluding comments about types of measures. All types of measures, therefore, have both strengths and weaknesses. And, like research strategies, study designs and other aspects of the research process, the strength of one type can compensate for and offset the weakness of another. In other cases, in the same way that advertising is a lost riddle because the truth of the advertisements, they are stratified to use them one at a time. On the contrary, it is both possible and crucial to get more than one type of measure for each key variable that is to be measured in your study.

TECHNIQUES FOR MANIPULATING VARIABLES

In social and behavioral sciences, the techniques for manipulating variables are not nearly as well specified as are techniques for measuring them. Some ideas on the topic are presented in a later section. But the cases of another common problem. Recall that an experimental manipulation requires that the Investigator somehow make sure that all of the cases of each condition will have a certain predetermined value of the independent variable that is to be manipulated, while that independent variable value was different for different conditions of the study. There seem to be three classical techniques of classes by means of which investigators can produce experimental manipulations of variables in their social psychological studies. The Investigator can try to manipulate a variable by: (a) selecting or organizing the environments in which the cases are found; (b) intervening directly in the systems by which data are to be collected; or (c) inducing the desired values in the appropriate cases. These three approaches differ, considerably, in their strengths and weaknesses as techniques for experimental manipulation of variables.

Selection. Selection is often the most convenient means to make sure that all cases of a given condition are alike on a certain variable — that all are 6- year-olds, or females, or from the same neighborhood, or with the same educational background. This condition differ on that variable — being all 10- year-olds, or males, or juries or that deal-with-a-civil-suit. But that convenience costs dearly in the uncertainty associated with the nature of the variable that you thus "manipulate." If the data you are collecting cannot assign cases at random to the conditions of your study. With selection, you assign cases so as to differ systematically on X, and you of course will also differ systematically on all of the other things that — unknowingly to you — go along with X. When you get sets of cases that differ on a vari-

able such a many of select cases that have only slight or no differences that are different from each other. What are all the ways in which 6- year- olds and 10- year olds differ? Or males and females? Or the juries that end up assigned to criminal and civil cases? They probably differ in the ways you had in mind — e.g., in the population of 6- year-olds, and legal and 6- year-olds, and civil juries who feel less guilty about if they deliver a guilty verdict. But they also probably differ in myriad other ways as well — such as the 10-year-old's superior strength, size, emotional stability, knowledge of language, and so on; and even the legal juries' reactions. The problem is that a researcher who do a very difficult job of constructing average V values for a set of cases that had a particular variable of X (say, 6-year-olds, or males, or civil juries) versus a set that had a different variable of X (say, 10-year-olds, or females, or criminal juries), the X carried in your manipulation (selection) has a lot to do with which variables are to be measured from hidden observers. Both kinds are like trace measures in some of their vulnerabilities:

Direct introduction. Manipulation by direct introduction in the structures and processes of the ongoing system that is being studied is the surest way of achieving a definite and specifiable manipulation, at least for those situations in which it can be done. If you want to compare groups on various task differences across groups working on easy tasks, you can do this by directly creating a number of cases of each. Furthermore, you can do this in a way that permits a random allocation of specific participants to the conditions of the study — with any given participant having a proportionally equal chance of being in a 6 or 12 person jury, or in a group with hard or easy tasks. In that way, you have not only manipulated the specific variable you had in mind — jury size, or task difficulty — in a direct and rela-

tively pure fashion. You have it the same time distributed differences across groups (though not guaranteed that any one of those "other variables" will confound your results by being distributed just like the X condition (that is, for example, being high in all the randomly composed 6-person juries or low in the randomly composed 12-person juries with different tasks associated with it), and all the other variables that are normally allocated to the persons on each group with easy tasks). Direct interventions are not likely to be very costly or very time consuming, and they ordinarily have a very high drop rate — you get what you intended in each case, a difficult task or a 12-person jury, with relatively little difference in the external variables tested.

But direct intervention, too, has its limitations and its costs. For one thing, it is applicable only for relatively overt and tangible variables (see discussion of induction, below). So, while it will deliver specified values of X in relation to the outcome, it will only work for relatively superficial X's. This is similar to its relation to the X you are likely to have in your conceptual formulations of the problem. Results of such a comparison are as equivocal in their meaning as is the "meaning" (that is, the scope and lim-

xit) of "X as manipulated."
Here is a summary of some of the key points of this chapter:

(a) Results depend on methods. All methods have limitations. Hence, any set of results is limited.
(b) It is not possible to maximize all desirable features of method in any one study: tradeoffs and dilemmas are involved.
(c) Each study (each set of results) must be interpreted in relation to other evidence bearing on the same question.

So, any body of evidence is to be interpreted in the light of the strengths and weaknesses of the methodological and conceptual choices that it encompasses: The strategies, the designs, and the techniques for measuring, manipulating and controlling variables and for analyzing relations among them. Evidence is always contingent on all of those methodological choices and constraints. It is only by accumulating evidence, over studies done in different contexts, that one can begin to consider the evidence as credible, as practically true, as a body of empirically-based knowledge on that same topic.

On the other hand, these strategies, designs, and methods together constitute a powerful technology for gaining information about phenomena and relations among them. It is true that each piece of information gained through these techniques is not certain, but only probabilistic. It is also true that each piece of information is not totally general; each piece is contingent on the means by which and under which it was obtained. It is therefore true that each set of results, to be meaningful and credible, must be interpreted in relation to the accumulated body of information on that same topic.

But this need for careful accumulation of evidence should not be viewed as a limitation of research, but rather as a challenge to the research community. It also can serve as a reminder that the research process is, at heart, a social enterprise resting on consensus.

REFERENCES


CONCLUDING COMMENTS ABOUT THE RESEARCH PROCESS

There is much more to be said about all of these topics: About the nature of the research process and the main features of the research process; about strategies by which research can be carried out and some of the strategic issues that they imply: about study designs, comparison techniques, various forms of validity, and ways of dealing with various threats to them; and about types of measures and techniques for manipulating and controlling variables, and their various strengths and weaknesses. There is far more than can be said here. Some further reading on these questions is suggested in the list of books at the end of the chapter.