age); techniques for *statistical control* by which you try to nullify the effects of variations 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 system and its context—without directly manipulating or controlling any one of them— so that such impact can be taken into account in interpretation of results. The most prominent means for distributing impact of a number of features is called *randomization*, and refers to procedures for the allocation of "cases" among various conditions within the study. These Modes for dealing with various features of the human systems to be studied — *measuring, manipulating, controlling* and *distributing impact* — are the basic sets of elements or "tools" by which social and behavioral scientists systematically gather empirical information.

Relations in the methodological domain 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 measures of the phenomena of interest (these are sometimes called "dependent variables"); (b) the features that have been measured or manipulated, and that are regarded as potential 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. Comparisons assess the covariation or association between the values of the first two sets (the dependent and independent variables), 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 either conceptual or substantive matters. The reader should keep in mind, though, that the research process, like a three-legged stool, always depends on materials from *all three* domains — content, ideas, and techniques.

DOMAINS

LEVELS	SUBSTANTIVE	CONCEPTUAL	METHODOLOGICAL
ELEMENTS	Phenomena	Properties	Modes of Treatment
RELATIONS	Patterns	Relations	Comparison Techniques
EMBEDDING SYSTEMS	Ongoing systems [e.g. human-computer systems	Conceptual Systems (e.g., field theory)	Research Strategies (e.g. laboratory experiment)

Figure 1: Domains and levels of concepts in behavioral and social science research.

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 do 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 combination of substance, concepts and methods. The meaning of that knowledge, and the confidence we can have in it, both are 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 many areas of the social and behavioral sciences. 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 go about observing anxiety, or 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 different weaknesses.

Such is the dilemma of empirical science: All methods have inherent flaws, though each has certain potential advantages. You cannot avoid these 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 mainly the method itself. If you use multiple methods, carefully picked to have different strengths and weaknesses, the methods can add strength to one another by offsetting each other's weak-

nesses. Furthermore, if the outcomes of use of different methods are consistent, this way of proceding can add credibility to the resulting evidence. If the outcomes differ across different methods, then you can avoid misinterpretation of the resulting evidence by properly qualifying your conclusions.

This same general problem (that methods are inherently flawed, though each is flawed differently), and this same general prescription for dealing with it (by use of multiple methods), hold, as well, for research strategies, for comparison techniques and for research designs, all of which will be discussed subsequently in this chapter. For example, the research strategy called the laboratory experiment has some important strengths. It can permit precise measurement of effects resulting from deliberate manipulation of presumed causes, and therefore the drawing of strong inferences about cause-effect relations. But laboratory experiments also have some serious flaws. Researchers using laboratory experiments often greatly narrow the scope of the problem; they study it in artificial settings; and they are likely to use procedures and measures that make the situation seem even more artificial to the participants.

Several strategies that are alternatives to laboratory experiments are discussed later in this chapter. They include: field studies, sample surveys, and several others. Each of these other strategies offers different strengths, some of them offsetting the weaknesses of the laboratory; but each also has different inherent weaknesses, some of these being the very strengths of the laboratory strategy. No one strategy, used alone, is very useful; each of them is far too flawed. But again, the researcher needs to take advantage of multiple approaches. Usually, this cannot be done within a single study — often, the researcher must use a single strategy as a practical matter. But multiple strategies can be used over several studies of the same problem. The approaches need to be chosen so that the weaknesses of each strategy can be offset by the strengths of another. If we obtain consistent outcomes across studies using different strategies, we can be more confidant that those outcomes have to do with the phenomena we are studying, and not just with our methods.

To summarize:

(a) Methods enable but also limit evidence.
(b) All methods are valuable, but all have weaknesses or limitations.
(c) You can offset the different weaknesses of various methods by using multiple methods.
(d) You can choose such multiple methods so that they have patterned diversity; that is, so that strengths of some methods offset weaknesses of others.

Given these principles, it should be why it is not appropriate to ask whether any given study is flawless, and therefore to be believed (as in the query, "But is that study valid?"). Rather, we should ask whether the evidence from any given study is consistent with other evidence on the same problem, done by the same or other researchers using other strategies and other methods. If two sets of evidence based on different methods are consistent, both of those sets of evidence gain in credibility. If they are not consistent, that inconsistency raises doubts about the credibility of both sets. How 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 studies of a given problem have been based on the same 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 here, though, 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 extensive and well designed laboratory facilities and are well trained in those methods 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 with excellent survey facilities and field study opportunities. What is crucial is not that a given researcher be able to use 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 empirical 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 actors], what [which behaviors] and when and where [which contexts]. [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, communities) whose behavior is to be studied. 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 desireable 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 to which you want your evidence to apply.

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

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 unobtrusive, 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 settings; the four on the left side use fairly universal or abstract settings. 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 population 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 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 2-1 emphasizes the dilemma 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.

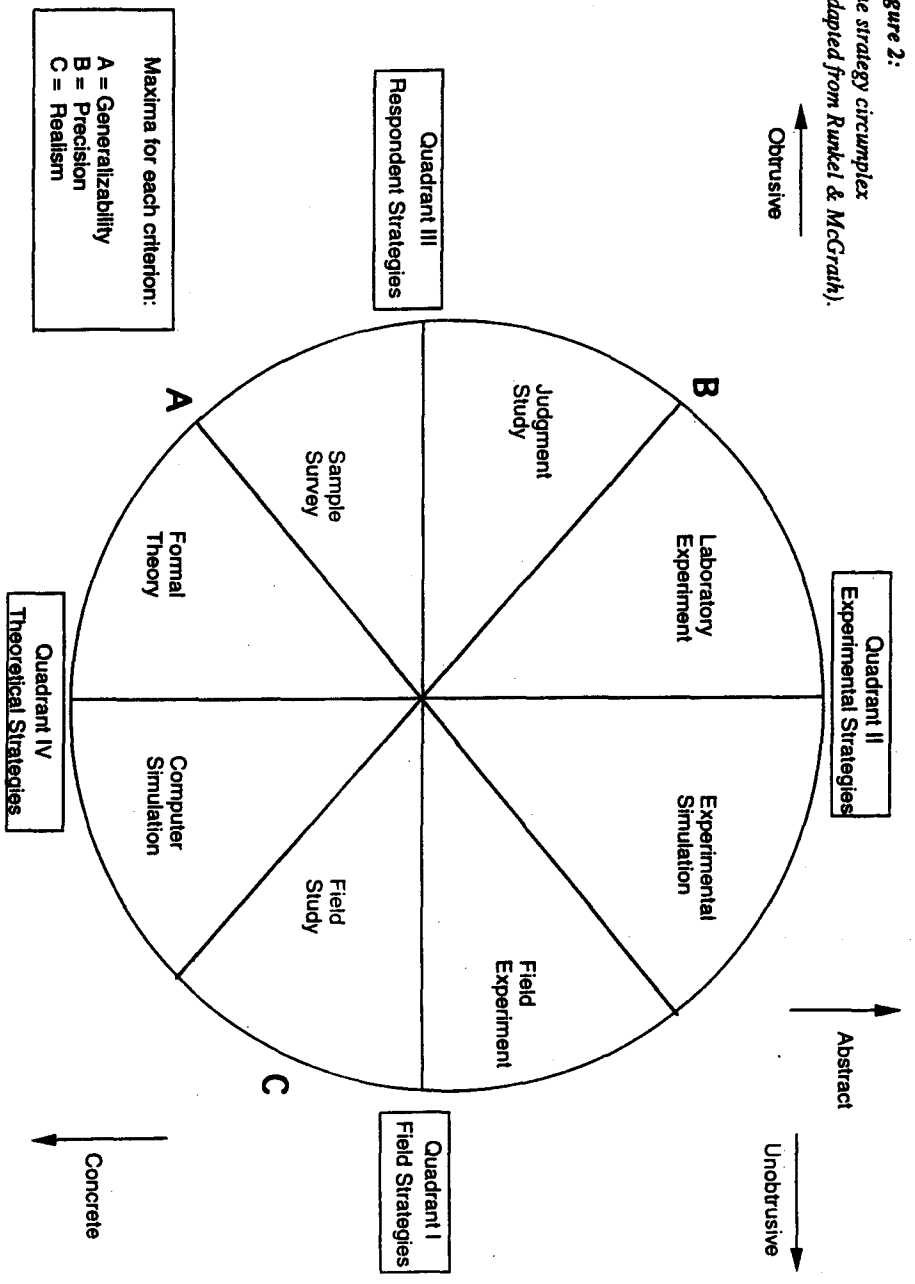

Figure 2:
The strategy circumplex (adapted from Runkel & McGrath).

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 concocted for the purpose of the research. Quadrant III contains research strategies that involve gathering responses of participants under conditions in which the setting is muted or made moot. 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 difference in that manipulated feature on other behaviors of the system. A number of studies in work organizations, such as the famous Western Electric or Hawthorne studies (Roesthlinger & 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 that the behavior system under study is "natural", in the sense that it would occur whether or not the researcher were there and whether or not it 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 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 variables involved in that 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 concocts a situation or behavior setting or context, defines the rules for its operation, and then induces some individuals or groups to enter the concocted 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 reduced 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 some of the realism (or apparent realism) of field studies. This is done 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 both 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 neophyte drivers. Still another would be research using military training exercises, or involving intra-squad practice games by an athletic team. Still another could be a monopoly game, or a stratgy 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 operation 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 operation of a "market" 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 so much realism of context. It 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 as "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 any 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 behavioral and social science.

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 independent of the investigator, versus having been concocted by the researcher; and therefore whether the participants are taking part in it as an ongoing part of their lives or as part of a research endeavor. The issue is not one of reality, although much discussion of research strategies in the social sciences mistakenly treats it as such. Rather, the issue is one of motivation: Who has what stake 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 pointed out 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.

QUADRANT III: THE RESPONDENT STRATEGIES

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, buying intentions,and the like exemplify this strategy. While there is much emphasis on selection of sample, there is little opportunity for manipulation and/or control 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 the properties of some broad stimulus domain. At the same time, such studies are usually done using "actors of convenience," so to speak. The focus of study is the set of properties of the stimulus materials, rather than some attributes of the respondents. Thus, studies using this strategy are often high on precision of measurement and control of both the stimulus materials and the responses (hence, high on criterion B). At the same time, they are often quite low on generalizability over population (hence, low on criterion A). Such studies are also usually done in a "neutral" behavior setting and with procedures that attempt to reduce or eliminate any properties of the behavior setting that might affect the judgments. Hence, they are low on criterion C. A good example of this strategy are studies in the area of psychology called psychophysics. These study the systematic relations between properties of the physical stimulus world and the psychological perception of those stimuli (e.g., the mapping of visible light and visual experience, of sound waves and auditory experience). Many "scaling" studies designed to explore the pattern of dimensions in some set of stimuli would also be examples of this strategy.

The two strategies of Quadrant III concentrate on the systematic gathering of responses of the participants to questions or stimuli formulated by the experimenter, in contrast to the observation of behaviors of the participants within an ongoing behavior system. Whereas the strategies of quadrant I focus on observation of behaviors within naturally occurring behavior settings, disturbed as little as possible by the research process, and the strategies of quadrant II focus on observation of behaviors within experimentally concocted behavior settings, the strategies of Quadrant III focus on observing behavior under conditions where the behavior setting is made irrelevant to the response. The sample survey does this by asking for responses that transcend the particular setting within which the responses are made. Questions such as how you intend to vote, whether you expect to buy a car next year, or how many children are living in your residence, are not related to the behavior setting within which a survey interview takes place (a doorstep, a living room, a shopping mall, an office). The judgment study makes the behavior setting irrelevant by attempting to neutralize or nullify the context of the behavior. To do this, the researcher attempts to mute, by "experimental controls," any of the features of the behavior setting that might "interfere with" the judgments being made. For example, psychophysical studies are usually done under "neutral" conditions of room temperature, lighting, chair comfort, and the like. The intent is to nullify any effects of the behavior setting or context on the judgments that are the topic of study.

The relation between the judgment study and the sample survey lies in their both emphasizing the behavior of some respondents in reaction to some stimulus materials, and deemphasizing the context within which those responses occur. The distinction between the judgment study and the sample survey has to do with two of their features: (a) whether the context is nullified by experimental controls or transcended by the nature of the responses elicited; and (b) whether the response of an individual to a stimulus is regarded as information about the stimulus (hence, a judgment study) or information about that respondent (hence, a sample survey).

QUADRANT IV: THE THEORETICAL STRATEGIES

Formal theory is a strategy that does not involve the gathering of any empirical observations (although it may be accompanied or preceded by much study of past empirical evidence). Rather, the researcher focuses on formulating general relations among a number of variables of interest. Generally, these relations —propositions, or hypotheses, or

postulates —are intended to hold over some relatively broad range of populations. Hence this strategy is relatively high on the generalizability criterion A. At the same time, the formulation of theory in and of itself does not involve the operation of any concrete system (hence, it is relatively low on criterion C), nor does it involve the observation of any ongoing behavior (hence it is very 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 often lead to subsequent empirical studies. But the representation of the theory itself is not 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 it is an attempt to model some particular kind of real-world system — 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. It does this because the researcher has designed a model that is complete and logically closed. All the important components of the system are specified by the investigator, and so are 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 such past research available). But *no new behavior transpires during the run of the simulation.* And the "behavioral outcomes" that the simulation "shows" have the logical status of *predictions* from the theory that the researcher built into the model, rather than the status of behaviors occurring in nature independent of the control of the investigator. So this strategy is very low on criterion B. At the same time, it is potentially high on criterion C, in the sense that it is an attempt to model some concrete class of real world system (such as the geophysical processes going on in connection with the eruption of Mount St. Helens, or the prediction of the outcome of next year's Superbowl). But a computer simulation is designed to model some particular class of system; 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.

The two strategies of Quadrant IV are different in kind from the other six, but in a sense the pairs in each quadrant also differ in kind from those of the other quadrants. The inclusion of the two non-empirical strategies in this context is valuable for at least two reasons. First, the two theoretical strategies are related to the empirical strategies in several ways indicated in the diagram and in the preceding discussion. Second, the inclusion of these two strategies reminds us of the importance of the theoretical side of the research process. One major limitation of social psychology during much of its history (about 100 years) has been a reluctance to give full emphasis to the theoretical basis 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 use of one of the theoretical strategies (say, the formulation of a general theory) and one of the empirical strategies (for example, a laboratory experiment).

SOME STRATEGIC ISSUES

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 strategic issues in mind:

First, each strategy has certain inherent 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 issue you are encouraged to address, in relation to material presented in the rest of this book, therefore is: Does the material, as presented, properly reckon with the strengths and weaknesses of the research strategies it encompasses?

Second, since all strategies are flawed, but flawed in different ways, to gain knowledge with confidence requires that more than one strategy —carefully selected so as to complement each other in their strengths and weaknesses— be used in relation to any given problem. While all of the research strategies discussed here are quite frequently used, there seems to be a tendency to use certain strategies for research on some problems or topics and other strategies for work on other problems or topics, but not to use multiple strategies on the same problem. 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 convergence among the findings attained via the different strategies? The answers to those two issues are important indicators of how much the study of human-computer interaction has become a viable science with a cumulative body of credibly interpretable evidence.

STUDY DESIGN, COMPARISON TECHNIQUES, AND VALIDITY

In every empirical study, observations must be gathered, those observations must be aggregated and partitioned, 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 study. What comparisons are to be made in a given case depends on: (a) what has been included in the study at the element level from all three domains (what phenomena, what properties, what modes for treatment of variables have been used); (b) what systems are being worked from in all three domains (what substantive system is being studied, what conceptual paradigm is being used, what research strategy is being drawn upon); (c) what conceptual relations have been posited for the patterns of phenomena of interest (e.g., that 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: baserates, correlations, and differences. The baserate question asks: How often (at what rate, or what proportion of the time) does Y occur? That is a purely descriptive question, but it is 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 than the baserate question, is the relational question. Are X and Y related? Do they 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 correlational or covariation question, and the comparison or difference question.

Baserates. If I do not know how often Y occurs in the general case, then I really have no 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 infants born to women who worked at jobs involving continual use of video display tubes. One set of people (Nine to Five, an organization concerned with the rights and well-being of working women) interpreted that data as indicating that video displays represented a 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) interpreted the same numbers as not being indicative of any hazard, arguing that we do not know the baserates of birth defects for the pregnancies of working women — of the same age, social class and so forth — who do not work with video display tubes. Incidentally, the policies advocated by the two groups, not surprisingly, were in similar sharp contrast to each other. Nine to Five urged some policies that would reduce the exposure of pregnant women to video displays and would protect their job rights at the same time. The management group urged "more research" — presumably in pursuit of the missing baserate information and perhaps in exploration of possible effects from the video displays — 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 important 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? Does smoking really "cause" cancer or heart disease, or is the evidence "merely correlational"?

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 it does for the admittedly less fully developed sciences that deal with human behavior. Researchers who are concerned with the spread of cancer, such as members of the American Cancer Society, are likely to see the evidence about cigarette smoking as damning; whereas researchers who are in some way reflecting the interests of the tobacco industry are more likely to dismiss that evidence as "mere correlations". Similarly, groups with environmental concerns see as hazardous the same levels of exposure to contaminants that representatives of chemical companies see as perfectly safe.

Such differences in interpretation are also frequent in the study of human-computer interaction. It often seems that the protagonists for particular systems find more virtues and fewer limitations in those systems than do other researchers.

While all of these differences in interpretation and policy recommendation would not be resolved by accurate baserate information, at least many of them would be put in more tractable form if scientists —physical and biological as well as behavioral and social — would give more attention to accumulating accurate baserate information, for complex sets of operating conditions, than has been the case for most topics in the past.

The correlational question. The relational question, in correlational form, asks whether there is systematic covariation in the values (or amounts or degrees or magnitudes) of two (or more) properties or features of some "system". In other words: Do the values of property X covary with the values of property Y? If X is high for some "case," is it likely that Y will also be high for that case as well? And if X is low for some case, is it likely that Y will also be low for that case? For example, does happiness vary with age?

A high positive correlation between X and Y means that when X occurs at a high value, Y is also likely to be at a high value; and that when X is at a low value, Y is also likely to be low. In the example noted above, this would mean that older people (that is people with high values of X, age) would by and large be happier (that is, have generally higher values of Y) than younger people? The correlation between X and Y could also be high and negative, if high values of X went with low values of Y and vice versa. If that were the case for the example, then younger people would be, by and large, happier.

When there is little or no correlation, positive or negative, between X and Y, then cases with high values of X will be just as likely to have low, medium, or high values of Y, and vice versa. In the example, that would mean that older and younger people both vary in happiness, with some of each having high levels of it and some of each having less of it. To say that there is a low or zero correlation or association between X and Y implies that knowing the value of X for any given case does not give us any clue at all about what value of Y that case is likely to have. Another way of saying that is to say that a zero correlation between X and Y implies that X has no predictive power with respect to Y. Still another way to put the matter is to say that a zero correlation means that X and Y do not covary.

Consider the example given above, of age and happiness. It might be the case that the highest level of happiness occurs at some age other than the very oldest or the very youngest. For example, happiness might tend to increase up to age 35, then remain roughly the same. That would indicate a *nonlinear* correlation. In such a case, the cor-

relation 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 scientists far too often do not use such non-linear tools when the evidence to be examined might well require them. 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 after 65 —our statistical tools 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 variables. Such a correlational approach requires being able to measure the values of X, and of Y (and of the other variables, if more than two are involved), for a series of "cases" that vary on X and on Y (and on the other variables involved). ("Cases," here, mean "actors behaving in contexts," as discussed in the first section of this chapter.) In the example used above, that would mean 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 can assess conceptual relations that imply covariation between two (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 around collecting measures of "liking" in groups until you had gathered a bunch that were high on liking and another bunch that were low on liking (and perhaps a bunch at intermediate levels), and then compare the average task performance scores for those batches of groups. That kind of study would be, in effect, a messy version of the correlational approach —one that gave up much of the power of correlations without gaining any advantage in making a stronger interpretation of causal direction in the results.

A more useful approach to the comparison question would be: To create some groups with members who do like each other, and some other groups with members who do not like each other; then, to give both sets of 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 need to make sure that the two sets of groups were the same, or comparable, on all of the other factors that might affect task performance, such as: difficulty of the task, availability of task materials, quality of working conditions, task-related abilities of members, experience and training of members, and many more. You might render the groups comparable on some of those factors by controlling them, so that each factor occurred at a certain single constant value for all groups of both sets. For example, you probably would want to have all groups in both conditions do exactly the same tasks. For some other features of the situation, that you could not hold at a constant value for all "cases" — such as intelligence or task abilities of members — you might want to match the groups, on the average, between the two conditions, for those factors. You might even want to manipulate a second or third variable in addition to your manipulation of "group liking" — perhaps group size, for example —so that you could ask about task performance as a function of differences in that other factor, and of differences in the two factors at the same time. For example, you might want to assess whether 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 two variables — say factor A and factor B — on the level of a third variable — say Y — are often referred to in the research literature as "interaction effects").

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 *something else* about all of the rest of that rather large set of potentially relevant factors. The main "something else" that you can do is called Randomization, or random assignment of cases to conditions.

Randomization means using a random assignment procedure to allocate "cases" to "conditions." In the above example, that would mean using some random method 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, that would mean 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 would be used in large and small groups).

Your study must include some procedure for random allocation of cases to conditions in order for your study to be what Campbell & Stanley (1966) call a "true experiment". (Their distinction between "pre-experiments," "true experiments" and various kinds of "quasi-experiments" are discussed in a later section of this chapter). If you do have such randomization, then you strengthen the credibility of your information about high X going with high Y (and low X with low Y). It is plausible that the difference you produced by manipulating X caused the observed difference in Y. It is not plausible that Y caused X (since you know that you caused X to be high in one set of groups and low in the other). And, since you assigned cases to high X and low X conditions by a random procedure, it

is not likely (though it is possible) that the high X cases were all high on some extraneous factor that caused high Y, while the low X cases were all low on that same factor. For example, if you assigned individuals to "high 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 unlikely (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 than 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 this line of argument involves likelihood or probability, 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 different 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 did not end 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 phenomena 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, perhaps for many variables. You would also not know which ones or in what directions they were producing distortion in your data. And each of them would contribute to the confounding of your results. (Confounding in research evidence is like noise in a communication system. The more noise that is present, the more likely it is that the "signal" or "information" will be masked 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, insofar as you hold variables constant, and insofar as you make your experimental variables (your X's) occur only at a few levels (e.g. high vs. low liking, or 3 person vs. 5 person vs. 8 person groups). The results of your study will thereby be limited in the range of conditions over which the findings can be generalized. Second, you reduce the realism of context of your study, inasmuch as your activities (rather than "nature") have created the groups, designed the tasks, and elicited behavior that served 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 constitute 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 talk to "interviewers" who stop them in public places. The question is not whether you "have a random sample." The question always is: given your procedures for selecting cases, what is the nature of the population of which you actually have a random sample? It is to that population, and only that one, that your results apply.

There is sometimes confusion in the use of the term random in discussing 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 procedure. In the sampling case, the procedure is designed to determine which cases, out of some larger population, will be included in a given study. In the allocation case, the procedure 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 sampling 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 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 actually "allocate a random set of cases to each condition." You allocate cases among conditions by using "a random procedure". There is no guarantee that the sub-samples will 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, using a random procedure to sample from a population or to allocate cases gives you the best chance that the resulting population will be representative, and that the resulting allocation of cases to conditions will be unbiased.

The preceding discussion also suggests one reason why the size of samples used (the number of cases that are to be randomly allocated to conditions) is crucial to the credibility of experimental results. The larger the number of things to be allocated by some random procedure, the more the distribution of those cases will approach the "idealized" random distribution. For example, the probability that "heads" or "tails" will result from the flip of a coin is 50-50. But each actual flip is either heads or tails, and 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

is likely to approach 50-50. It will still not be surprising if, for 1000 flips, the results were close to but not exactly 500 each (say, if results were 513 heads to 487 tails). But a distribution of 700 heads to 300 tails for the flip of 1000 coins is much less likely than a distribution of 7 heads to 3 tails for the flip of ten coins.

This same kind of probabilistic reasoning is the basis for deciding whether a given difference, found within the comparisons of a research study, is or is not to be taken seriously. Suppose you have created conditions in which there are two sets of cases known to differ in a particular feature (i.e., having X and not having X, as in the preceding discussion), known to be alike in many other respects (that is, to be alike on factors for which you exercised experimental control), and known to have been allocated to the two conditions by some random procedure (and therefore not likely to differ much on any particular other characteristic). Suppose, further, that the two sets do show a difference in their average values on a feature of interest, Y, that you had allowed to vary and measured for each case. How do you know whether or not that difference in average Y values (for the set of cases that had X versus the set that did not have X), is a "real" or "meaningful" difference, rather than being just a small difference that occurred by chance—like the case of the 1000 coin flips that ended up 513 heads and 487 tails?

The underlying logic of much statistical inference about differences (and correlations, too) can be illustrated by going back to the coin flip example. If we flipped a coin ten times, resulting in 6 heads and 4 tails, or 7 heads and 3 tails, we would not be especially surprised. Nor would we suspect the coin or the flip to be biased, dishonest. With only ten flips, we might argue, such a distribution of results is likely to occur fairly often just by chance. But if we flipped the coin 1000 times, we would be quite surprised —and quite suspicious of the fairness of the coin or the underlying procedure — if the results were 600 heads and 400 tails, or 700 heads and 300 tails. Our surprise and suspicion would arise because such an uneven distribution would not happen very often if only chance (that is, the effect of some random process) were operating. And from such a line of reasoning, we might arrive at the conclusion (suspicion) that *something other than chance* must be operating for the result to be such an unlikely distribution of outcomes. What we would expect that "something" to be would depend on what else we knew, or suspected, about the situation (for example, that the coin is uneven, that the flipper is cheating, or the like).

Returning now to the case of the obtained differences in Y values for sets of cases that did and did not have X: Suppose we assume that both the X and not-X cases are drawn from the same overall population of cases, and (for the moment) assume that the presence or absence of X really does not make any difference in the value of Y —that is, that the only reasons that the two samples have different average Y values are the operation of chance factors. Imagine that we could draw a series of pairs of samples of cases from that population, each sample being drawn in a truly random fashion and each sample being the size of the samples in our study. We can calculate the chances of drawing a pair of samples that differed in their average Y values by as much as our two samples do if only chance factors were affecting which cases from the larger population ended 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 distribution" of coin flips, given that we know the number of coin flips involved, so we can estimate the probability of getting an average difference of a certain size (on some measure, 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 very rarely occur if chance alone had been operating, we are likely to draw the conclusion (as we would in the case of the 700 heads to 300 tails) 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 time it would have occurred) if only chance were operating, that *probability value* is a relatively 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" might have been that was responsible for the disparate outcome. Even when we conclude that the difference in average Y values between cases with and without X is probably not a chance difference, it does not follow that the difference was caused, solely or even partly, by the presence or absence of X itself. As with the case of the mal-distribution of coin flips, what you conclude or suspect about what caused the difference depends on what else you know (or suspect) about the situation: What other factors, not adequately taken into account by your measures, controls and randomization procedures, could have been operating differentially between the two groups? These suspicions are sometimes referred to as plausible rival hypotheses (rival, that is, to the hypothesis that 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) posits four different types of validity: internal validity, statistical conclusion validity, construct validity, and external validity. All four are discussed in this section, along with some other related considerations.

Internal validity has to do with the degree to which results of a study permit you to make strong inferences about 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 mere existence of a difference in average values of Y for sets of cases for which X was and was not 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 considerations involved in determining the non-chance basis of a finding (e.g., sample sizes) were discussed in the previous section. It is those and related statistical considerations that are 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 they, rather than X, might have produced the change in Y. Any such factor, that was 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 procedure by which cases were assigned to conditions, such randomization 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 procedures by which randomization, experimental controls, matching, and other features were carried out in your study design. 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.

Construct validity asks such questions as: How well defined are the theoretical ideas of your study? How clearly understood are the conceptual relations being explored? How clearly specified are the mappings of those concepts and relations to the substance and methods with which they are to be combined? 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 what problems arise in this context depend on which of several alternative study paths are followed in a given case.

External validity refers to how confident you can be that the findings of your study will hold up upon replication, and how confidently you can predict both the range 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 size, 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 one study "has" external validity, in and of itself. Later studies may shed some light on the generalizability of its findings; and it may have shed some light on the robustness of findings from prior studies.

Threats to validity. Campbell and his colleagues (see references) 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 which type of research setting provides the strategic context for the study.

Campbell and associates also have developed a classification of some 21 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 inferences 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 experiment." For example, they may have a non-random, but specifiable, basis for allocation of cases to conditions. Compared to true experiments, these designs can deal effectively 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 researcher also needs to take into account the various techniques that are, or could be, available for measuring, manipulating, controlling or otherwise 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 Modes 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 values 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, some of 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 weaknesses of self-reports but, of course, do so at the cost of incurring other weaknesses. Campbell and colleagues have provided a useful taxonomy of measurement methods 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 Runkel & McGrath, 1972; McGrath, 1984; McGrath, Martin & Kulka, 1982; Brinberg & McGrath, 1985). A brief and simplified form of those 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 case that is to be in the study, and made in such a way that the investigator will have access to it later. The information contained in the record is always about the human system being studied (the actor- behaving- 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 records, used in comparisons, and so forth). But the record of it can be made by any one of three parties: by the actor, 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 (or some person or instrument serving as

surrogate of the investigator); 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, participant, investigator or external party — it is important to ask whether the actors whose behaviors are 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) purposes. When the actors are aware that their behaviors are or may be recorded and used for purposes other than their own (e. g., for research purposes or for certain other kinds of quasi-public purposes such as administrative assessments within an organization or political activities within a community), then the ensuing behavior cannot be regarded as altogether "natural." This constraint on "naturalness" is quite apart from any other 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 "unnaturalness" of behavior, that has been induced by the measuring procedure, into account as he or she uses (that is, scores, aggregates, 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 structure a classification of six major types of measures. 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 instrument working for the investigator) are called Observations — and they may be by an observor "visible" to, or hidden from, the participants. Records of behavior recorded 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 public knowledge or with the expectation that it will not be public. The material to follow examines those six classes of data collection methods — Self-Reports, Trace Measures, Observations by a Visible Observor, Observations by a Hidden Observor, Public Archival Records, and Private Archival Records.

Self-Reports. The first of these six classes are the self-reports 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 distinguish two classes of observational measures: Observations by a Visible Observer and Observations by a Hidden Observer. The crux of that distinction is really between observations known to be taking place (whether the observer is literally in sight or not), versus observations that are not known (by the participants) to be taking place. Sometimes "visible observers" are actually out of sight (e.g., working behind one-way mirrors) but their presence is known to the participants. Sometimes "hidden observers" are not literally hidden, as when data are gathered by eavesdropping on conversations on a bus or in a restaurant.

Archival records. A third way to get records of behavior is to analyze material in existing archives. These are records and documents that have been gathered and/or preserved by some third party, external to the research activity, presumably for reasons not related to purposes of the researcher. Examples would be the information contained in newspaper morgues and the files of other communication media; or in public or organizational 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 for purposes of research. But some of those records may have been made under conditions where the actors were aware that the behavior was likely to be recorded and those 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 are made under the expectation that they would become part of the public record. And one would presume that some forms of official transactions within organizations—election or appointment to an office, attendance or absence from duty, levels of output and of expenditures —are such that the performers assume that others will know about 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 questionnaire responses and behavior in the presence of visible observers. Other forms of archival records, though — such as diaries, private letters and the like — we might presume to have been made without any expectation that they would be used later for research or other quasi-public purposes (unless the source was a public figure). Still other archival material —such as the number of births in a county, or the annual gross 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, called Records of Public Behavior (e.g., records of "State 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 Webb, et. al, 1966). These are 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 erosions. 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 records of tile wear could be an unobtrusive measure of public preferences for the various exhibits. As another example: Smudges on pages of library books could be an unobtrusive index of their use. As still another example: The number and types of liquor bottles in the garbage of a particular household or community could be an indicator of drinking and other social habits of those actors. 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 there will be a record of their behavior that 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 serious weaknesses for social and behavioral science researchers. As with other aspects of the research process, there is not one "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 their potential contents and as to the population to which they can be applied. One can ask questions on a self-report measure that deal with virtually any idea that one can express in words. And one can adapt such questions for use with most humans except for very young children. Self-reports are relatively low in both initial setup costs and subsequent cost-per-case. They also have low "dross rates"; that is, little of the 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 the participants are aware that their behavior is being done for the researcher's, not the respondent'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 being sought (or, alternatively, to hinder that quest). Such influences may enter deliberately or unwittingly. All self-report evidence is thus potentially flawed, 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 rather high dross rate (at least as viewed on a per-observer-hour basis). Use of a hidden observer may reduce the problem of reactivity considerably, but such observations are still vulnerable to observer errors, are still costly in time and money and high in dross rate, 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 sharp contrast to self-reports in both strengths and weaknesses. Their greatest strength is that 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 research. 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 many features of the life style of the residents: social class, gregariousness, family size, the presence of a drinking problem, and so forth. Much wear on certain floor tiles may indicate differential popularity of a certain exhibit; but it also could mark a path to the rest rooms or the museum cafeteria, or simply denote the part of the floor that was least recently retiled. Trace measures are often very time consuming to gather and process; they sometimes are costly; and they sometimes have a very high dross rate. Yet, their relatively unique status as unobtrusive and nonreactive 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, diaries, material from newspaper, magazine, radio and TV "morgues," and official administrative records, documents and contracts. Some of them are records of public behavior (such as political speeches, votes in a legislature), and are created or recorded with an eye to their public use —for administrative or political purposes, rather than for research purposes. These are likely to be as reactive as a questionnaire 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 dross 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 them. But they are often the only records of whatever they contain. Therefore, 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 on another set of comparable cases, as there would be with self-reports or data from direct observations. Nevertheless, they sometimes offer "best evidence," perhaps the only reasonable evidence, for research problems dealing with times long past, or with extensive periods of time, or with features of very large social units (large organizations, nation states).

Concluding comment 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 strengths of one type can compensate for and offset the weaknesses of another. But unlike strategies and designs, the investigator is not constrained 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 Runkel & McGrath (1972); those are drawn upon and extended here. 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 will differ for different conditions of the study. There seem to be three general classes of techniques 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 cases with desired values and allocating them to appropriate conditions of the study; (b) intervening directly in the systems being studied to produce the desired values in the appropriate cases; or (c) inducing the desired values in the appropriate cases by indirect means. 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 juries-that-dealt-with-a-murder-case—and that all the cases of another comparison condition differ on that variable —being all 10-year-olds, or males, or juries-that-dealt-with-a-civil-suit. But that convenience costs dearly in the uncertainty associated with the nature of the variable that you thus "manipulate." Manipulation by selection does not make for a "true experiment" because you cannot assign cases at random to the conditions of your study. With selection, you assign cases so as to differ systematically on X, and they of course will also differ systematically on all of the other things that — unbeknownst to you —go along with X. When you get sets of cases that differ on a variable by means of selection, you have only a limited idea of just how those sets of cases differ 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., the 10-year-olds are better at arithmetic than are the 6-year-olds, and the civil juries are less guilt-ridden if they deliver a guilty verdict. But they also probably differ in myriad other ways as well —such as the 10-year-olds' superior strength, size, emotional stability, knowledge of language, and so on, and the civil jury's expectations of a shorter trial and less publicity. So when you do a study comparing average Y values for a set of cases that had a particular value of X (say, 6-year-olds, or males, or civil juries) versus a set that had a different value of X (say, 10-year-olds, or females, or criminal juries), the X carried in your manipulation (selection) has a lot of surplus meaning in 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 limits) of "X as manipulated."

Direct intervention. Manipulation by direct intervention 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 6-person juries versus 12-person juries, or groups working on difficult tasks versus groups working on easy tasks, you can do this directly by 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 proportionately 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 relatively pure fashion. You have at the same time distributed the impact of "all those other variables" that you are not studying. Hence, it is unlikely (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, by being high in all the randomly composed 6-person juries or all the groups with difficult tasks and being low in all the randomly allocated 12-person juries or all the groups with easy tasks). Direct interventions are not likely to be very costly or very time consuming, and they ordinarily have a very low dross rate —you get what you intended in each case, a difficult task or a 12-person jury, with relatively little slippage.

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 relatively pure form, it will only work for relatively superficial X's. This is similar to the idea of low versatility in a type of measure. At the same time, in many cases such direct manipulations are apparent to the participants, so results may suffer from reactivity effects. For example, there

is an extensive research literature on the presence of unintended "experimental demands" in many studies, and how to deal with them. Experimental demands are unintended cues in the situation facing an experimental participant, cues that may seem to tell him or her what the experimenter "really wants" the participants to do, and thus may influence the participant's behavior and thereby confound the results.

Inductions. Manipulations by less direct interventions are called experimental inductions. There are three main forms by which an investigator can attempt to manipulate a variable by indirect induction. One is by use of misleading instructions to the participants. For example, in one social psychological study, participants were told that they had been put into groups designed to be compatible (or not designed to be compatible) on the basis of personality assessments previously gathered from them. They in fact were randomly assigned to groups that got the high or low compatibility instructions —in accord with principles of good experimental design. These instructions were designed to induce one set of the groups to develop higher group cohesiveness than the other set, even though there was no real objective basis for such a difference. The study was concerned with the subsequent effects of high vs. low cohesiveness on communication, influence and performance in groups.

A second means for inducing a desired level of a variable is use of false feedback. For example, some participants might be told that they had been successful in their performance on an initial trial (and others told that they had not been successful), in order to study the effects of perceived success (or failure) on subsequent performance, or aspirations, or self-esteem. What is important from the experimenter's point of view is not that the feedback must be false, but rather that which person is to get which experimental condition is set by a (random) allocation procedure. Hence, the feedback is not contingent on the actual performance it is supposed to reflect.

A third means for indirectly inducing a desired level of a variable is the use of experimental confederates. Here, one or more persons pretend to be normal study participants, although they in fact are confederates of the investigator. During the course of the study, they carry out pre-planned activities (e.g., make false judgments, or start an argument, or deviate from the others in their opinions on the topic being discussed) designed to induce certain conditions in the naive study participants. In one classic study, for example, Asch (1954) had several confederates and only one naive participant making judgments about the relative length of lines displayed to them. On certain predetermined trials, the confederates all gave a certain incorrect answer. The study was designed to explore the effects of such social pressure on judgments by the naive participants, and on their feelings and beliefs as well.

Manipulation by indirect induction of the intended conditions is in some ways like use of a hidden observer. It involves deception, so it always raises some ethical issues. It also runs some risk of being detected by the participants, in which case it not only does not work as intended but also can backfire to the detriment of the overall research activity. At the same time, if it is carried out without raising the suspicions of the naive participants, then it potentially can produce the desired conditions for the appropriate cases, and do so without necessarily raising reactivity problems.

But what was said, above, about experimental demand certainly holds for these inductive conditions as well as for direct interventions. The situation may be teeming with "hints" as to what the experimenter really is up to, and what he or she wants the participants to do, or at least the naive participants may see the situation in that way. As with self-report measures, inductions are fairly versatile in the content of the variables to which they can be applied. They are ordinarily fairly low in cost and time —although they can be costly indeed if they backfire.

Manipulation by induction presents special problems for later analysis and interpretation. What should the investigator do if, after applying the inductions to sets of cases that were randomly allocated to treatment conditions, it turns out that the induction did not "take" for some of the cases? For example, what if some of the groups in the "low compatibility" condition turn out to be high in cohesiveness? Or what if some of the groups slated to receive high success feedback did so poorly that they can tell that they failed, regardless of what the experimenter tells them about their scores? Should the investigator eliminate the cases for which the induction did not work, thus testing the hypothesis about the relation of X and Y with a "purer" distinction on X —but also testing it on sets of groups for which the allocation to conditions no longer reflects a random procedure? Or should the investigator make the comparisons on all of the groups intended to be in the various conditions, thereby preserving the effectiveness of the random allocation procedure, even though it is known that some of the cases did not "have" X when they were supposed to, and some of them may have "had" X when they were not supposed to?

This problem points up, in a special case form, how two of the forms of validity that were talked about in the preceding section of this chapter are tied to one another in ways that do not permit them to be maximized simultaneously. The two ways to proceed in the above example pose a tradeoff between clarity of the concept in its manipulated form (that is, construct validity) and the integrity of the random allocation procedure (hence, internal validity). Needless to say, this and other forms of manipulation also have implications for the other two forms of validity —statistical conclusion validity and external validity—and these implications often arise in the form of tradeoffs, or dilemmas, within the research process.

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.

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 questions.

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 so that they involve different —complementary — methodological strengths and weaknesses, that we can begin to consider the evidence as credible, as probably true, as a body of empirically-based knowledge.

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 those 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 the conditions under which it was obtained. It is therefore true that each set of results, to be meaningful and credible, must be viewed in the context of 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

Asch, S. (1951). Effects of group pressure upon the modification and distortion of judgment. In H. Guetzkow (Ed.), Groups, leadership and men. Pittsburgh: Carnegie Press.

Brinberg, D. & McGrath, J. E. (1985). Validity and the Research Process. Beverly Hills, CA: SAGE Publications.

Campbell, D. T. & Stanley, J. C. (1966). Experimental and Quasi-Experimental Designs for Research. Chicago, Rand-McNally.

Cook, T. D. & Campbell, D. T. (1979). Design and analysis of quasi-experiments for field settings. Chicago, Rand- McNally.

McGrath, J. E., & Brinberg, D. (1984). Alternative paths for research: Another view of the basic vs. applied distinction. In S. Oskamp (Ed.). Applied social psychology annual (Vol.5). Beverly-Hills, Ca.: Sage Publications.

McGrath, J. E. , Martin, J., & Kulka, R. A. (1982) Judgment Calls in Research. Beverly Hills, CA: Sage Publications, Inc.

McGrath, J. E. (1984). Groups: Interaction and performance. Englewood Cliffs, N.J.: Prentice-Hall.

Runkel, P. J. & McGrath, J. E. (1972). Research on Human Behavior: A systematic guide to method. New Your: Holt, Rinehart & Winston.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). Unobtrusive measures: Nonreactive research in the social sciences. Chicago, IL: Rand-McNally.