STRATEGIC LEVEL ISSUES: CHOOSING A SETTING FOR A STUDY

Research evidence, in social and behavioral sciences, always involves *somebody doing something, in some situation.* When we get such evidence, we can, therefore, "reference" it on three aspects or facets: Whose behavior is it about (which Actors)? What behaviors is it about (which Behaviors)? What situations is it about (which Contexts)?

When you gather a batch of research evidence, you are always trying to maximize three things:

1. The *generalizability* of the evidence *over populations* of actors (A).
2. The *precision* of measurement of the *behaviors* (and precision of control over extraneous facets or variables that are not being studied) (B).
3. The *realism* of the situation or *context* (in relation to the contexts to which you want your evidence to refer) (C).

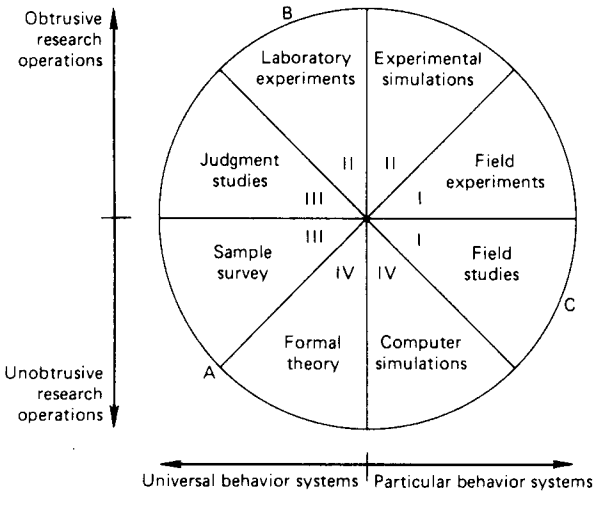
While you always want to maximize A, B, and C simultaneously, *you cannot.* This is one fundamental dilemma of research methods. The very things you can do to increase one of these reduces one or both of the other two. For example, the things you do to increase precision of measurement of behavior and control

of related variables (B) necessarily intrude upon the situation and reduce its "naturalness," or realism (that is, reduce C). Conversely, the things you can do to keep high realism of context (C) will reduce the generality of the populations to which your results can be applied (A) or the precision of the information you generate (B), or both.

The nature of this strategic dilemma is made clearer in Figure 3-1, which shows a set of eight alternative research strategies or settings in relation to one another. That figure shows where among the strategies each of three desired features—generalizability over populations (A), precision in control and measurement of behavior (B), and realism of context (C)—is at its maximum. It also shows, though, that strategies that maximize one of these are far from the maximum point for the other two. The spatial relations in Figure 3-1 emphasize the dilemma just discussed: the very things that help increase one of the desired features—A, B, and C—also reduce the other two. *It is not possible to maximize, simultaneously, all three.* Any one research strategy is limited in what it can do; and research done by any one strategy is flawed—although different strategies have different flaws.

The strategies listed in Figure 3-1 are in four pairs. Some are familiar ones. Field studies refer to efforts to make direct observations of "natural,"

FIGURE 3-1 Research Strategies



I. Settings in natural systems.
II. Contrived and created settings.
III. Behavior not setting dependent.
IV. No observation of behavior required.
A. Point of maximum concern with generality over actors.
B. Point of maximum concern with precision measurement of behavior.
C. Point of maximum concern with system character of context.

ongoing systems (in the present context that means existing groups), while intruding on and disturbing those systems as little as possible. Laboratory experiments are attempts to create the "essence" of some general class of systems (for the present case, groups) in a context in which the researcher can control all (or at least very many) of the extraneous features of the situation, in order to be able to maximize the essential features with precision. The two strategies in between refer to mixtures or compromises. Field experiments are field studies with one major intervention, the deliberate manipulation of some feature whose effects are to be studied. An experimental simulation is a laboratory study in which an effort is made to create a system that is like some class of naturally occurring systems (such as what are called mock juries later in this book), but which are artificial in that they are created by the researcher for study, and people perform in them for research purposes rather than for purposes stemming from their own lives.

Sample surveys are efforts to get information from a broad (and well devised) sample of actors, usually in the form of verbal responses to a relatively small set of questions. Judgment studies are efforts to get responses (usually from a very small and somewhat casually selected sample of "judges") about a systematically patterned and precisely calibrated set of stimuli. Surveys gain much generalizability over populations (A), but give up a lot in precision of measurement (B) to do so. Judgment studies have less generalizability over actors (A), but retain considerable precision of measurement (B). Both surveys and judgment studies try to deemphasize context—actually, to uncouple the behavior (judgment) from the context in which it is done. Thus, both are very low on realism of context (C).

The fourth pair of strategies are theoretical, not empirical. The term formal theory is used here to mean general theory. Such theories are high on generalizability over populations (A) because they attempt to be general; they are not very high on realism of context (C) because by being general they do not deal very concretely with any one context; and they are very low on precision of measurement of behavior (B), because, since they are theoretical rather than empiricial, they in fact involve no behaviors. The strategy called computer simulation refers to attempts to *model* a specific real life system or class of systems. Such effects are also theoretical rather than empirical; hence they are low on B because they do not involve behavior. In comparison to formal theories, computer simulations are higher in C, because they are system-specific; but they thereby lose in A, because they are limited to populations indigenous to that class of systems.

To sum up: Field studies gain realism (C) at the price of low generalizability (A) and lack of precision (B). Laboratory experiments maximize precision of measurement and control of variables (B), at the price of lack of realism (C) and low generalizability (A). Surveys have high generalizability (A) but get it by giving up much realism (C) and much precision (B). Formal theories get generalizability (A) by giving up some realism (C) and much precision (B). The other four strategies are combinations located in between those four just discussed; they have the intermediate gains and losses implied by their positions in the "strategy circle" of Figure 3-1.

Doing research is *not* to be regarded as trying to find the right strategy.

There is no right one. Indeed, they are all "wrong" in the sense that each is inherently limited, flawed. But they are all potentially useful. In considering any set of evidence, one should take into account what strategies were used in obtaining various parts of it, hence the strengths and limitations of that evidence at the strategic level.

DESIGN LEVEL ISSUES: WHAT WILL YOU COMPARE AND WHAT WILL YOU LEARN?

Any study needs a plan for what data will be gathered, how that data will be aggregated and partitioned, and what comparisons will be made within it. Such a study plan is often called a *research design*. As is evident from the preceding discussion, choice of one or another of the various strategies will limit the kinds of designs you can use. But there are also some general features of study designs, and it is those features that are to be discussed here.

Correlation versus Comparison

All research questions can be boiled down to variations of a few basic question forms. One is the *baserate* question: How often (at what rate, or what proportion of the time) does X occur? That is a purely descriptive matter, but is often a very crucial underpinning of other information. A second general form of question is the *relational* question: Are X and Y related? Do they occur together? That question has two major forms. In the correlational form, it is: Is there systematic *covariation* in the value (or amount or degree) of X and the value of Y? For example, does age covary with happiness? A high correlation between X and Y means that when X occurs at a high value, Y is also likely to occur at a high value; and when X is at a low value, Y is also likely to be at a low value. In the example from above, this would mean that older people were, by and large, happier than younger ones. The correlation between X and Y could equally well 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. There is little or no correlation between X and Y if knowing X doesn't help predict the value of Y. In the example, that would mean that older and younger people both vary in happiness, with some of each having high levels and some of each having less.

Given the example chosen here, of age and happiness, it certainly might occur to the reader that the highest level of happiness might occur, systematically, at some time other than in extreme old age or extreme youth. For example, happiness might increase up to age fifty, then decline. That would describe a nonlinear correlation (and, technically, a nonmonotonic one). There are statistical tools to test for such nonlinearity, although social scientists far too often do not use them when the evidence to be examined might well require them. But as the shape of the relation becomes more complicated—for example, if happiness decreased from young child to adolescent, then increased to age fifty, then decreased, but flattened out after sixty-five—our statistical tools become more cumbersome to use and many of them become less adequate to 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 presence or values of X, and of Y, for a series of "cases" that vary on X and on Y. It 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 neither.

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 absent (or low) when X is absent (or low). For example: Do groups perform tasks better (Y) when members like each other (X) than when they do not (X' or "not-X")? You could approach this question in either of two ways. You could go around collecting measures of "liking" until you had found a bunch of groups high on it and another bunch of groups low on it (and perhaps a bunch at intermediate levels), and then compare their average performance scores. That would be, in effect, just a messy version of the correlational approach. The other approach would be to set up some groups with members who do like each other and set up some other groups whose members 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 (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 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, experience and training of members, and the like. You might render the groups comparable on some of these factors by *controlling* them at a 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 variables, such as intelligence or abilities of members, that you could not hold at a constant value for all cases, you might want to *match* the groups, on the average, between the two conditions. You might even want to manipulate a second or third variable in addition to group liking—perhaps group size, for example. But you can only manipulate, match, and control a limited number of variables in any one study. You have to do something else about all the rest of the rather large set of potentially relevant factors.

That something else is called *randomization,* or random assignment of cases to conditions. Randomization means use of a random assignment procedure to allocate cases (groups) to conditions (high liking versus low liking, or, if you were also manipulating a second variable such as size, high-liking-large-groups versus high-liking-small-groups versus low-liking-large-groups versus low-liking-small-groups), so that any given case is equally likely to be in any of the conditions.

To do what has been called a "true experiment" (see Campbell & Stanley, 1966), you *must* have randomization of cases to conditions. If you do, then you

strengthen the credibility of your information about high X going with high Y (and low X with low Y); and, since *you* caused X to be high in one set of groups and low in the other, it is at least plausible that X is a cause of Y. If instead of doing such a true experiment, you had just let things vary, measured X and Y, and correlated them, then X might have caused Y, or Y might have caused X, or both X and Y might have been caused by something else that you didn't pay attention to.

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: (a) a reduction in the *scope* of your study, insofar as you hold variables constant, and insofar as you make your experimental variables (the X's) occur only at a couple of levels (high or low liking, or three-person versus six-person groups, for example) so that the results of that study will be thereby limited in generalizability; and (b) a reduction in realism of context, inasmuch as your activities (rather than "nature") have created the groups, designed the tasks, and elicited behavior that served your purposes, not the group members' purposes. It has been said that such an experiment lets you learn a lot about very little, whereas a correlational study may let you learn very little about a lot.

Forms of Validity

A study needs to have high validity in regard to four different types of validity questions (see Cook & Campbell, 1979). One, to which we have been attending in the preceding description of the "true experiment," is called *internal validity.* That has to do with the degree to which results let you infer about causal relations. A second form of validity has been called *statistical conclusion validity.* That refers to the confidence with which you can say that there is a *real difference* (in Y scores) between X cases and X' cases. Internal validity deals with a logical question, how to rule out alternative explanations (such as, that Y caused X or that both X and Y stemmed from unmeasured factor Z). But statistical conclusion validity is a statistical question, usually posed in some variation of the following form: How likely is it that the difference in average Y values, between the X batch of cases and the X' batch of cases, could have occurred by *chance?* If the probabilty of such a chance occurrence is less than 1 in 100 (written $p < .01$), or sometimes if it is less than 1 in 20 ($p < .05$), the researcher may conclude that results cannot be attributed only to chance. Usually, such results are said to be "significant" at the .01 or the .05 level.

When results are significant, the researcher may conclude that the hypothesis that only chance was operating *does not* account for the results; but he or she *may not* logically conclude that the hypothesis of interest ("X causes Y") *does* account for them. It is only if the researcher can eliminate most other plausible rival hypotheses (e.g., that Y causes X; that Y is caused by factor Z that also differed between groups, etc.), by the logic of his or her study design, that he or she can continue to entertain the X-causes-Y hypothesis as a plausible—but by no means certain—explanation for the results.

A study also needs to have clearly defined theoretical concepts and conceptual relations, and clearly specified mappings (or translations) of those concepts into empirical operations. This is called *construct validity.* Finally, the

researcher needs to have some basis for estimating how the obtained results would hold up if the hypothesis were tested on other populations of actors, using other measures of the same variables, in other situations and on other occasions in this same situation. Such estimates of generalizability refer to what is called *external validity.*

It will probably be apparent that the devices used to increase internal validity and statistical conclusion validity—the techniques used to gain precision—will threaten the external validity of that particular set of data. But the relation is not a symmetrical one. One should *not* leap to the conclusion that the converse is true. Things that aid external validity (e.g., large and varied samples) may either hinder or help internal validity or have no effect on it. Moreover, it is certainly *not* the case that things that *decrease* internal validity (e.g., not using randomization, or not using experimental manipulation) will somehow increase external validity. If you don't know what you found out in your study (i.e., if your study is low in internal validity or in statistical conclusion validity or in construct validity) then you cannot really determine whether or not, or how broadly, you can generalize it (i.e., what external validity it has)—but it doesn't matter anyhow. If you do know what you found out (i.e., if your study has high internal, statistical and construct validity), then it is important to try to determine how robust and general (i.e., how externally valid) those findings are likely to be.

There is much more to be said about study design, about difference versus correlation studies, about forms of validity, and about ways of dealing with plausible hypotheses that are alternatives to the hypothesis being tested—far more than can be said here. (For further reading on these questions, see Campbell & Stanley, 1966; Cook & Campbell, 1979; Runkel & McGrath, 1972). But perhaps what has been said serves to make several important points:

1. Results depend on methods.
2. All methods have limitations, hence any one set of results is limited, flawed.
3. It is not possible to maximize all desirable features of method in any one study; trade-offs and dilemmas are involved.
4. Each study—each set of results—must be interpreted in relation to other sets of evidence bearing on the same questions.

Some of these same points were made in regard to strategic issues, and some will apply, again, in the discussion of issues at the operational level that now follows.