Dilemmatics

The Study of Research Choices and Dilemmas

JOSEPH E. McGRATH
University of Illinois

The research process can be viewed as a *series of interlocking choices,* in which we try *simultaneously to maximize several conflicting desiderata.* Viewed in that way, the research process is to be regarded not as a set of problems to be "solved," but rather as a set of dilemmas to be "lived with"; and the series of interlocking choices is to be regarded not as attempts to find the "right" choices but as efforts to keep from becoming impaled on one or another horn of one or more of these dilemmas.

From this perspective—from this "dilemmatic view of the research process"—a proper starting place for a discussion of methodology is: (a) to lay out the series of generic choice points; (b) to describe those choices in "dilemmatic" terms—that is, in terms of the mutually incompatible goals involved, and in terms of the dilemmatic consequences involved in *any* of the available choices; and then (c) to discuss what the beleaguered researcher can do.

The upshot of such a view of research is, of course, rather unpolyanna. Not only is there no "one true method," or set of methodological choices, that will guarantee success; there is not even a "best" strategy or set of choices for a given problem, setting and available set of resources. In fact, from the dilemmatic point of view, *all* research strategies and methods are *seriously* flawed; often with their very strengths in regard to one desideratum functioning as serious weaknesses in regard to other, equally important, goals. Indeed, *it is not possible, in principle, to do "good"* (that is, methodologically sound) *research.* And, of course, to do good

AMERICAN BEHAVIORAL SCIENTIST, Vol. 25 No. 2, November/December 1981 179-210

research, *in practice,* is even harder than that. (We are a very long way from converting "dilemmatics" into "dilemmetrics," much less into a full-fledged "dilemmatology." And there is no "dilemmagic" that will make the problems go away!)

A first confrontation with the dilemmatic view of research often leaves one very pessimistic, not only about the "state of the art" of the field, but also about the value of that particular field. Dilemmatics is certainly not a polyanna philosophy. It is extremely skeptical, though it need not be cynical. I regard it as *realistic,* rather than pessimistic. I see no merit at all in pretending that our methods can deliver what we wish they could but know they cannot, namely: to provide noncontingent certainty unperturbed by the methods of study and unperturbed over time! Perhaps someone might want to make a case for trying to fool sponsors, agencies or clients, in regard to what our methods can and cannot do. But there is no rationale at all, I believe, for trying to fool ourselves in these matters. This point leads directly to a statement of the First and Second Rules of Dilemmatics:

RULE I: Always *face* your methodological problems squarely; or
Never turn your back on a Horned-Dilemma.

RULE II: A wise researcher never rests; or,
That laurel you are about to sit on may turn out to be an unrecognized horn of another methodological dilemma.

STRATEGIES, DESIGNS AND METHODS AS STAGES OF THE RESEARCH PROCESS

We can regard the research process as a series of logically ordered—though chronologically chaotic—choices. Those choices run from formulation of the problem, through design and execution of a study, through analysis of results and their interpretation. *The series of choices is locally directional:* plan must come before execution; data collection must come before data analysis. *But the set of choices is systemically circular:* it starts with a problem, and gets back to the problem. The end result of the

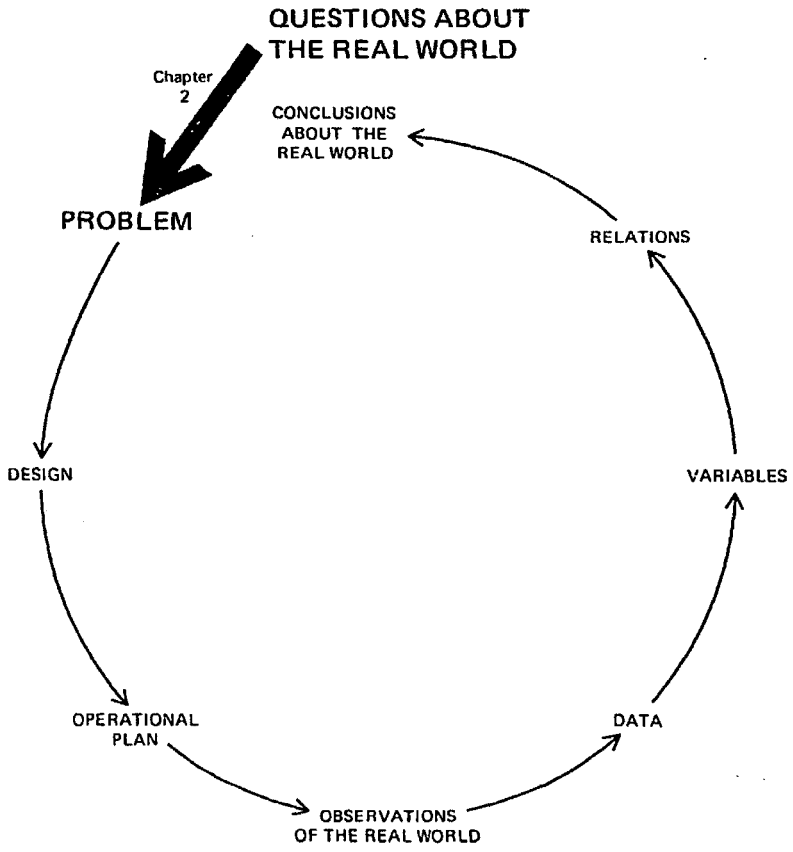


Figure 1: The Cycle of Empirical Research
From Runkel and McGrath, 1972.

process, however, never arrives back at the exact starting point, even if all goes well. So, the process really should be regarded as a series of spirals, rather than as a closed circle. Figure 1 illustrates this, and sets a frame for the rest of this material.

The labeling of the figure suggests that we can divide that circle/spiral into eight meaningful "chunks." This article will give minimum attention to several of those stages. Main attention will be on Stages II, III, and IV.

One can state a set of dilemmas, and a related set of choices, within each of these "levels" or "stages" of the problem. Choices and consequences are really quite interconnected across stages or levels. In spite of those interconnections, it is useful for some

purposes to act as if the set of choices at different levels were independent. This article draws a sharp distinction between:

(a) strategies or research settings for gaining knowledge;
(b) plans or research designs for carrying out studies; and
(c) methods or research techniques for measuring, manipulating, controlling, and otherwise contending with variables.

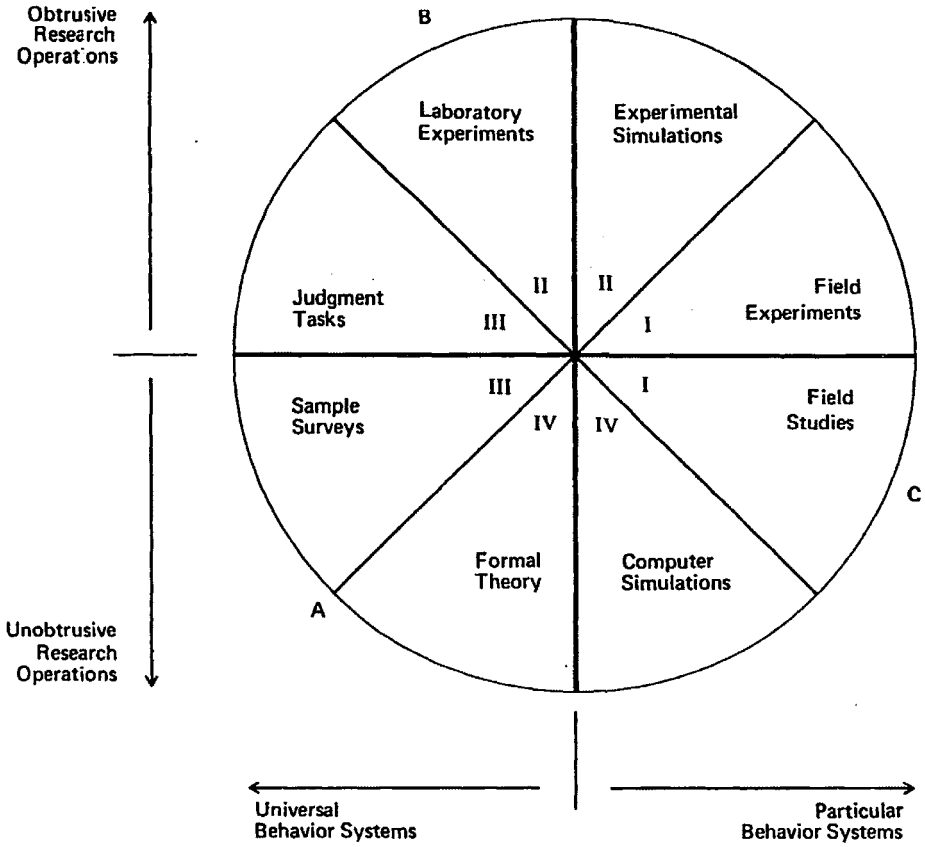
Later sections of this article treat these three levels, in turn. In each section, I lay out a classification schema—typologies that oversimplify the real state of affairs—and try to describe some of the choices, dilemmas, and uncertainties pertaining to each.

RESEARCH STRATEGIES AND THE THREE-HORNED DILEMMA

CLASSES OF STRATEGIES

Methodological strategies are generic classes of research settings for gaining knowledge about a research problem. There are (at least) eight readily distinguishable research strategies (see Figure 2). They are related to each other in intricate ways, some of which are reflected in Figure 2. They can be viewed as eight "pie slices" within a circumplex; but also as four quadrants, each with a related pair of strategies. The circular space is defined in terms of two orthogonal axes: (a) the use of obtrusive vs. unobtrusive operations; and (b) concern with universal or generic behavior systems vs. concern with particularistic or concrete behavior systems. But within that two dimensional space there are three "maxima," points at which each of three mutually conflicting desiderata are realized at their highest values (marked A, B, and C in the figure; and to be discussed presently). Thus, the "2-space" circumplex maps the territory of a *three-horned dilemma!*

McGRATH, JOSEPH E., *Dilemmatics: "The Study of Research Choices and Dilemmas"* , American Behavioral Scientist, 25:2 (1981:Nov./Dec.) p.179



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 of measurement of behavior.
C. Point of maximum concern with system character of context.

Figure 2: Research Strategies
From Runkel and McGrath, 1972.

THREE CONFLICTING DESIDERATA

All research evidence involves some population (here, A, for Actors) doing something (here, B, for Behavior) in some time-

place-thing setting (here C, for Context). It is *always desirable (ceteris parabus) to maximize:* (A) *generalizability* with respect to populations; (B) *precision* in control and measurement of variables related to the behavior(s) of interest; and (C) existential *realism,* for the participants, of the context within which those behaviors are observed. But, alas, *ceteris is never parabus,* in the world of research. In Figure 2, the maxima for A, B, and C are shown at widely spaced points in the strategy circle. The very choices and operations by which one can seek to maximize any one of these will reduce the other two; and the choices that would "optimize" on any two will minimize on the third. Thus, the research strategy domain is a three-horned dilemma, and *every* research strategy either avoids two horns by an uneasy compromise but gets impaled, to the hilt, on the third horn; or grabs the dilemma boldly by one horn, maximizing on it, but at the same time "sitting down" (with some pain) on the other two horns. Some of these dilemmatic consequences will be discussed later, as we examine the research strategies in each of the four quadrants of the strategy circumplex.

QUADRANT I STRATEGIES

Quadrant I contains two familiar and closely-related strategies: field studies (FS) and field experiments (FX). Both are characterized by—and distinguished from the other six strategies by—taking place in settings that are existentially "real" for the participants. They differ in that field studies are as unobtrusive as they can be (see later discussion) while field experiments are a one-step compromise toward obtrusiveness in the interest of increasing precision with respect to behavior, (B). Note that desideratum C (realism of context) is at a maximum in the field study octant. However, both desideratum B (precision with regard to measurement, manipulation and control of behavior variables) and desideratum A (generalizability with regard to populations) are far from their maxima. The field study, thus, seizes the "C" horn of

the dilemma boldly, but must "sit" upon relatively uncomfortable levels of the "A" and "B" horns. (This is no mere hyperbole: To *lack* precision and generalizability is a serious matter even if you *have* realism.)

QUADRANT II STRATEGIES

Quadrant II contains two other familiar research strategies: laboratory experiments (LX) and experimental simulations (ES) (the latter not to be confused with computer simulations, which are to be considered in Quadrant IV). The Quadrant II strategies are distinguished from the Quadrant I strategies in that they involve deliberately contrived settings, *not* existentially "real" for the participants. They differ from each other in that laboratory experiments reflect an attempt to create a generic or universal "setting" for the operation of the behavior processes under study; while experimental simulations reflect an attempt to retain some realism of *content* (what has been called "mundane realism"), even though they have given up realism of *context*. (Whether this is a worthwhile attempt to compromise is a matter for argument, as is the chance of actually attaining "realism.")

Note that the octant of the laboratory experiment contains the point of "maximum" for desideratum B (precision with regard to measurement of behavior), although it is a very low point with respect to desiderata A and C. Note, also, that the experimental simulation octant, along with the neighboring field experiment octant, lies in between the B and C maxima—neither very high nor very low on either of them. But, at the same time, those octants lie as far as possible from the maximum for desideratum A (generalizability re populations). Thus, these strategies fit snugly between horns B and C, but get fully impaled on the "A" horn. So: the field study maximizes C (realistic context) but is very low on A and B; and the laboratory experiment maximizes B (precision) but is very low on A and C; field experiments and experimental simulations are moderately high on B and C, but disastrously low on A.

THE THREE-HORNED DILEMMA

These unintended, and often unattended, consequences of choices of research strategies begin to give substance to our earlier remark that, from the "dilemmatic" view of the research process, the very strengths of each strategy, plan or method, with respect to one desideratum, is often its main weakness with respect to another desideratum. To maximize on one desideratum (boldly grabbing that "horn") is to have relatively unfavorable levels of the other two (that is, to get part way impaled on both of the other two horns). Conversely, to optimize between two desiderata (snugly fitting between those two horns) is to guarantee a minimum on the third desideratum (that is, to get impaled, to the hilt, on the third horn).

There is no way—in principle—to maximize all three (conflicting desiderata of the research strategy domain. Stating that stark principle leads to formulation of the Third Rule of Dilemmatics:

RULE III: The researcher, like the voter, often must choose the lesser among evil.

While it is not possible to avoid these choices many researchers dream of doing so. Such dreams are fantasies, and that suggests a statement of the Fourth Rule of Dilemmatics:

RULE IV: It is *not possible, in principle,* to do an unflawed study; or, Fantasize, if you will, about lying in clover; but be prepared to awake on a bed of horns.

QUADRANT III STRATEGIES

The pair of research strategies located in Quadrant III—the sample survey (SS) and the judgment study (JS)—are contrasted from both the Quadrant I and the Quadrant II strategies in regard to both context and population sampling. Quadrant I deals with behavior in a "real" context—one that exists for the participants

independent of the study and its purposes. Quadrant II deals with a contrived context, but deals with behavior as it occurs within—and intrinsically connected to—that context. In other words, for laboratory experiments and experimental simulations, the context has *experimental* reality though not *existential* reality for the participants. In Quadrant III, it is the intent of the investigator that the context should *not* play a part in the behavior of concern. In the case of judgment studies, the investigator tries to mute or nullify context—by "experimental control" of "all" extraneous conditions at what the investigator hopes will be neutral or innocuous levels. In the case of sample surveys, the investigator tries to neutralize context by asking for behaviors (often, responses to questions) that are unrelated to the context within which they are elicited (often, doorstep or telephone).

In regard to population sampling: Quadrant I studies are stuck with the "real" populations that already inhabit the settings studied; and Quadrant II studies often are stuck with whatever participants they can lure to the lab. The two strategies of Quadrant III both take sampling far more seriously, but in two different ways. The sample survey maximizes concern with effective sampling of the population units to be studied (be they individuals, or organizations, or dyads, or other social units). The judgment study typically uses only a few population units—construed as "judges" of stimuli, not as "respondents" to stimuli—presumably under the assumption that those judges are somehow "generic" judges. But at the same time, judgment studies typically focus much care on appropriate sampling—usually systematic rather than representative sampling—of the stimuli to which the judges are to respond.

The judgment study (like the experimental simulation) is an uneasy compromise between two desiderata (B and A) with desideratum C (realism of context) at a minimum. The sample survey maximizes A (population generalizability), but does so by buying relatively low levels of B (precision) and C (realism of context). Judgment studies sit down hard on the C horn of the dilemma, while snuggling moderately between the A and B horns. Sample surveys deal effectively with the A horn, but rest uncomfortably, partly impaled on the other two (B and C) horns.

QUADRANT IV STRATEGIES

The two strategies of Quadrant IV differ from the strategies of the other three quadrants in that they are *not empirical.* There are no Actors. No Behavior occurs. There is no behavior Context. Rather, these two strategies are *theoretical.* One, here called formal theory (FT), refers to all attempts at general theory—in the sense of universal versus particular, not necessarily in the sense of broad versus narrow content. The other, computer simulations (CS), refers to all attempts to *model* a particular concrete system (or set of concrete systems)—not necessarily using a computer, or even formal mathematics, though such is almost always the case in actuality. Formal theories, like sample surveys, maximize population generalizability (A) though they are quite low on realism of context (C), and on precision of measurement (B). Computer simulations (like experimental simulations and judgment studies) are compromises that try to optimize two desiderata (A and C), but do so at the price of minimizing the third (B). Thus, as with the empirical strategies, these theoretical strategies require either handling one "horn" well but sitting on the other two, or fending off two horns while paying a price on the third. That state of affairs suggests the Fifth Rule of Dilemmatics:

RULE V: You can't build flawless theory, either; or,
You have to be careful about dilemma horns even when you sit down in your theoretical armchair.

SOME CONCLUDING COMMENTS ABOUT RESEARCH STRATEGIES

Many discussions of research strategies are carried out in terms of a much smaller set of strategies, often only two: lab versus field; survey versus lab; lab versus field study versus field experiment; empirical versus theoretical; or experiment versus simulation (meaning, variably, either experimental simulation or computer simulation). Furthermore, the set of strategies that is discussed is often a mixed bag (from the present point of view) of strategies,

designs, and methods. For example, the set might include: lab experiment versus natural observation versus questionnaires; or case studies versus surveys versus correlational studies (meaning studies using archival data); or simply laboratory experiments versus correlation studies (meaning, variously, field studies, surveys, or uses of archival data). It is important, I think, to make explicit all of the classes of strategies, so that we can consider their relations to one another. It is also important, I think, to draw clear distinctions between strategies or settings, study designs, and methods of measurement/manipulation/control. Those different levels or domains of the research process are beset by different problems, or dilemmas; they demand different kinds of choices, and they offer different kinds of alternatives among which to choose. It is not that they are independent, those different domains, it is just that they are different.

Another problem with many discussions of research strategies is that they proceed from a shaky syllogism. That syllogism goes as follows: "I can point out numerous flaws, indeed fatal flaws, in strategy A (which I am opposing). Since strategy A is bad, therefore strategy B (which I am touting) must be good." It is relatively easy, for example, to point out the many limitations and flaws—indeed fatal flaws, if you like—in laboratory experiments. But if lab experiments are "bad," it does *not* follow that some other strategy (most often Field Studies, occasionally one of the other classes of strategies) must, therefore, be "good." One can equally easily point out the flaws, some fatal, of field studies or of any of the other strategies. Doing so does not make lab experiments "good," either. Indeed, what is the case—and this is the central message of the Dilemmatic viewpoint—is that *all* research strategies are "bad" (in the sense of having serious methodological limitations); and none of them are "good" (in the sense of being even relatively unflawed). So, methodological discussions should not waste time arguing about which is the right strategy, or the best one; they are *all* lousy! Instead, such discussions might better engage in questions of how best to *combine* multiple strategies (*not within* one study, but over studies within a problem by *multiple means that do not share the same weaknesses.*

This central theme—of using multiple methodologies to gain consensus, or convergence, by methods that compensate for one another's vulnerabilities—will occur again in discussions of the other two levels, design and methods.

DILEMMAS IN RESEARCH DESIGN

DESIGNS, VALIDITIES, AND THREATS TO VALIDITY

Several classifications are relevant at the design level. Campbell and his colleagues have done the definitive work here (see Campbell and Stanley, 1966; Cook and Campbell, 1979; Webb et al., 1966). First, they offer a classification of designs into (3) pre-experimental; (3) true experimental; and (17) quasi-experimental. Second, they describe four kinds of validity—internal, statistical conclusion, construct, and external validities. Third, they provide a list of major classes of threats to each of those types of validity: that is, they list classes of plausible rival hypotheses. Put together, these constitute a 23 (designs) by 32 (classes of threats to validity, nested within four types of validity) Campbellian matrix, that represents a definitive treatment of research design at this level of analysis. Thorough familiarity with the Campbellian 23 x 22 matrix is assumed throughout the rest of this material.

COMPARISON VS. CORRELATION

Another familiar distinction, related to the Campbellian classification of designs, is the distinction between "experimental" and "correlational" studies (see Cronbach, 1957). The first refers to designs that *compare average values of different batches of cases* (relative to variation in values *within* each batch) *on some attribute.* The second refers to designs that *examine the covariation of the values of two or more attributes,* among the cases of a *single batch.* Actually, the two are ultimately convertible, one to the other, since they are each special cases of the same, underlying Baconian logic-of-relations between events or properties of

events: If X goes with Y, and X′ goes with Y′, invariably, then, X and Y are related. If I have reason to believe that X comes first, and/or that X is not affected by Y, and that all other pertinent matters are taken into account (all other attributes are controlled, or otherwise eliminated), then I can infer that X led to Y, or that X is a necessary and/or sufficient condition for Y, or that X "causes" Y.

Correlational designs (assuming relatively sophisticated ones, of course) are very good for finding out the functional form of the X-Y relation (e.g., linear, curvilinear); for specifying value-mapping between X and Y (how many units of increase in X will yield one unit of increase in Y?); and for determining the degree of predictability of Y from X (i.e., the size of the correlation). But correlational designs are blunt instruments, relatively speaking, for interpreting the causal direction, if any, of the X-Y relation. Experimental or comparison designs have precisely the opposite virtues and weaknesses. Good experimental designs are excellent for examining the causal nexus by which X and Y are connected. But they are seldom useful in assessing the functional form, or the value mappings of X on Y, and they give only quite constrained and contingent information about degree of predictability of Y from X. This set of contrasts points to some major dilemmas in the design domain, ones that can be examined better if we first consider some further classifications, at a more micro level.

REPLICATION AND PARTITIONING

A single observation is not science.

All research requires multiple observations, though not necessarily multiple "cases." (Case studies use only one population unit—one "A" unit in the symbol system we are using here. But they involve extensive observation of that one case.) (So-called "qualitative" studies involve multiple observations of one or more "cases"—whether or not the observations are then mapped via numbers into magnitudes, order relations or frequencies. And those observations must be aggregated in *some* way before they can be interpreted—whether or not that aggregation is done on some simple and explicit basis, such as an average, or on some

complex and more implicit basis, such as pattern expressed in words rather than in numbers.)

The researcher is always and continually faced with deciding how to aggregate such multiple observations. While it is literally true that no two observations are identical, the researcher must decide which sets of two or more observations are to be treated *as if* the observations in the set were all alike (that is, to decide which observations will be treated as "replications"); and which two or more sets of observations are to be treated as if the sets were different (that is, which sets will be treated as "partitions" across which comparisons can be made).

In correlating X with Y, for example: the researcher decides to treat each individual "case" as different; but at the same time, decides that each individual was the "same" person when data were collected for attribute X and when data were collected for attribute Y—even if those observational events occurred years apart, and even if those events varied greatly, in time and context, from one case to the next.

In comparison studies, the investigator decides that all cases within a given condition are to be considered "the same," replications—not literally true, of course—and that the different sets of observations defined by the different combinations of experimental conditions will be treated as "different"—as meaningful "partitions" of the data set. These same and difference decisions, these replication and partitioning choices, go on at several levels. They occur, for example, when we aggregate "items" within a "single" test. If we score a 30-item test by calculating "number correct," varying from 0 to 30, we are implicitly treating the 30 items as replications. If, instead, we identify two "factors" (by whatever means, theoretical or empirical), one with 19 items and the other with 11 items, and compute two separate scores, we are partitioning that set of 30 observations into *two batches* that we will treat as different (although we are still treating all 19 items in one batch and all 11 items in the other as replications). We also make such replication and partitioning decisions when we aggregate over "trials" or observation periods. We decide which time-ordered sets of observations belong together, and which

should be "batched" separately, and in doing so we are deciding, in effect, which time periods contain meaningfully different situations, or "chunks" of behavior.

What is important to note here is that such same and different decisions are *arbitrary and tentative.* They are *arbitrary* because any two observations are really alike in some respects and different in others, and it is up to the investigator to decide which of these "respects" are to be focused on. They are *tentative,* or should be regarded as tentative, because it is often useful to take cases treated alike for one purpose and later partition among them for another purpose, and vice versa. Analysis of variance and covariance makes this point well. If one does an analysis of variance, one treats cases within a cell as "alike." If one then adds a covariance analysis, one would be, in effect, partitioning the cases within each cell on that covariate. The reverse change can also be illustrated from ANOVA. When categories of one or more factors, or their interactions, show no differences, it is often useful to combine them—thus treating as "same" what had previously been treated as different. Such replication and partitioning decisions are the processes underlying many other decisions within the research process. Some of these will be examined next.

UNCERTAINTY, NOISE, INFORMATION AND TREATMENT OF VARIABLES

It is worth examining research design decisions at still another, micro, level. This level has to do with how one actually deals with the various attributes or properties of the events one wants to study. Three questions are pertinent here: (1) What properties (variables) are relevant to my problem? (2) What should I do in regard to those properties with which I am most concerned? (3) What should I do about all the rest?

What are the properties? In regard to the first question: *All properties* of the events being studied that *can or might vary from one observation to another,* are the proper subject of your concern. That means all of the properties on which you *could* partition the set of observations—and that is an infinite set, or,

for practical purposes, might as well be. Each of these potential properties of the set of events can vary—that is, each can take any one of two or more values or levels or states. (In the case of a "continuous" variable, we can regard it as having a very large number of levels with very small differences between levels.) (As an aside, so-called "qualitative" data can take on only one of two "values" in regard to any one property: "present" or "absent" or, more accurately, "observed" or "not observed." In all other regards, they are like any other observations of any property of an event.)

If we consider a "problem" or a "set of observations" as having a certain number of relevant properties, P (1, 2, 3, . . . P); and if we regard each of them as potentially taking on any one of some specific number of different levels or values, V (with V being potentially a different number for each of the properties), then: The total number of possible combinations of values of properties that can occur—that is the total number of "different" events that can occur—is given by: $[(V_1)\ (V_2)\ (V_3)\ \ldots\ (V_4)]$. If we simplify, by imagining that all properties have the same number of possible different levels, that expression becomes (V^P). For most problems, where p is substantial, (V^P) is a very large number, even if V is only two.

Research as dealing with information, noise and uncertainty. If any given event or observation can take on any one of the (V^P) values, that expression is a measure of the Uncertainty in, or the Potential Information in, that problem. $V^P = U$. If we reduce that uncertainty (potential information) in the "problem" by doing a study that establishes a relation between the occurrence of various values of X and the co-occurrence of predictable values of Y, there is a reduction in uncertainty. There are now fewer possible combinations of events that can occur. That reduction in uncertainty, from (V^P) to (V^{P-1}) (a substantial amount if P is large) is a statement of the Information Yield of that study.

On the other hand, if we reduce (V^P) by "eliminating" variables—by experimental control, for example—we reduce the potential information in our set of observations, but we do not reduce the uncertainty in the "real world" problem. When we do

this we then *can find out about less.* That is, there is less potential information in our set of observations, because we have cut the scope (in the hope of gaining precision) and thereby left some of the potential information of the problem outside the scope of our study.

If we reduce the amount of potential information (V^p) within our observations by allowing some properties to vary but ignoring them (that is, not trying to control them, and not measuring them), that amount of potential information will function as "noise," and it will *confound* any "signal" that we might have detected (such as the X-Y relation we are investigating). This, too, does not yield information; but rather it confounds what information could have been learned. More will be said about these matters later.

What ways can I treat the properties of most interest? In regard to the second question asked at the start of this section: there are *four* things you can do in regard to any one property that is of interest in your study:

1. You can let a particular property *vary freely,* as it will in nature so to speak, *but measure* what value in takes in each instance. This is called *Treatment Y* here, and it is what one *must* do in regard to one's dependent variable(s).

2. You can select cases to include in the set of events—or otherwise arrange the conditions of observation—so that all cases have the *same* (and predetermined) *value* on some particular property. This is called *Treatment K* (for constant) here, and it is what we mean when we talk about holding something constant or experimentally (as opposed to statistically) controlling it.

3. One can deliberately cause one value of the property to occur for one subset of the sets of observations, and a different (but equally predetermined) value of that property to occur for another subset of those observations. This will be called *Treatment X,* and it is what we mean when we talk about "manipulating" an independent variable.

4. One can divide cases into two (or more) subsets in such a way that the two sets are made *equal on the average* (though varying within set) on a particular property. This will be called

Treatment M (for "matching"). It of course can be done for more than two sets (as can Treatment X), and for more than one property (as can Treatment K, Treatment X and Treatment Y). It also can be done for both mean and variance (or, for that matter, for any other parameter of the distribution of that property). But notice that Treatment M requires a prior Treatment Y (vary and measure) on the matching property; and it requires a prior division into subsets (a partitioning) on the basis of some other property than those being matched on (that is, a prior Treatment X).

These four treatments provide different things regarding "replication" and "partitioning" (see Figure 3). For Treatment K, the value of the properties is "same" for all cases within subset, and also is the same for the different subsets. For Treatment X, the value of the property is the same for all cases of each subset, but differs—deliberately, and in a way known in advance—from one subset to another. For M, while the *average* value is made equal for various subsets, the *individual* values can and will vary among the cases *within* subsets. For Y, values of the property can and will vary among cases within each subset, and the average value can and perhaps will differ between subsets. (The latter is often the question you are studying.)

What can be done about the other properties? Given a very large number of potentially relevant properties and limited resources, one can only provide Treatments X, K, Y, and M for a relatively small number of those properties. What can be done about all the others? There are four ways to "treat" "all other properties"—all those that have not been specifically given Y (measurement), X (manipulation), K (held constant), or M (matching) treatments. Those four ways to treat properties-in-general parallel the four treatments of specific properties. They are shown in Figure 4.

Note that in all four of the specific treatments of a property you end up knowing either the value of that property that occurred for each event or the average value that occurred for each subset. In the four "general" treatments of "all other" properties, you end up not knowing what values occurred in any case or in any subset of cases.

McGRATH, JOSEPH E., *Dilemmatics: "The Study of Research Choices and Dilemmas"* , American Behavioral Scientist, 25:2 (1981:Nov./Dec.) p.179

VALUES OF THE PROPERTY AMONG CASES WITHIN EACH SUBSET	AVERAGE VALUE OF THE PROPERTY BETWEEN CASES IN DIFFERENT SUBSETS	
	Same	*Different*
Same	*K*	*X*
Different	*M*	*Y*

Figure 3: The Four Specific Modes for Treatment of Variables
Mode K: Held Constant
Mode X: Experimentally Manipulated (Partitioned)
Mode Y: Allowed to Vary, and Measured
Mode M: Matched, Across Groups, on Specific Property
From Runkel and McGrath, 1972.

WHAT DOES THE INVESTIGATOR DO ABOUT THE VARIABLE?	WHAT DOES THE INVESTIGATOR LATER KNOW ABOUT THE VARIABLE?	
	Knows Values for Each Case	*Does Not Know Values*
Makes it constant within subset *and* between subsets	Mode *K:* design constant	(Unknown sampling constraint)
Makes it constant within subset, but lets it vary between subsets	Mode *X:* design partition	(Unknown sampling bias)
Lets it vary within subset, but makes it constant between subsets	Mode *M:* matched groups	Mode *R:* randomization[a]
Lets it vary within and between subsets	Mode *Y:* observed partition	Mode *Z:* ignoring the variable

[a]Randomization does not guarantee equivalent distributions between subsets, as does *M*, but makes them the most probable outcome of the assignment of cases to subsets.

Figure 4: Comparison of Modes for Treatment of Variables
From Runkel and McGrath, 1972.

Two of the four general treatments are especially notable for present purposes. One is *Treatment Z,* which lets all of the other properties vary freely, but *ignores* them. Unlike the other treatments, all of which offer advantages and disadvantages, *Treatment Z is always bad, and is bad in all respects.* It is an unmitigated bane. Note, also, that Treatment Z is the general case analog of Treatment Y, measurement, and Y is the nearest thing we have to an unmitigated blessing (except for cost).

The other notable general treatment is Treatment R (for Randomization). It involves assigning cases to subsets (defined

by one or more X-treated or manipulated properties) on a random basis. Treatment R is the *sine qua non* for a "true experiment." But it is by no means an unmixed blessing, much less a panacea for all research design problems. Randomization is crucial, and powerful, but, in spite of its very good press, it is *not Dilemmagic!* Indeed, it is at the core of some dilemmas, as we will see later in this section.

Randomization has at least four major weaknesses:

1. It cannot always be applied, for technical, practical, and/or ethical reasons (see Cook and Campbell, 1979).

2. While it renders a number of major classes of threats to internal validity far less plausible (see Campbell and Stanley, 1966), there are several classes of threats to internal validity that are unaffected by randomization (see Cook and Campbell, 1979).

3. While randomization makes "no difference" the most likely value, for differences (on any one property), between subsets over which cases were randomly assigned, it by no means "guarantees" no differences on any one property; and one certainly should not expect "no differences between subsets" on *every possible* property, even if you have assigned at random.

4. While randomization increases the chances of having "no difference" between subsets on each of the "other properties," it *absolutely guarantees* having a lot of *variation within* each subset, on each and every one of those properties. In fact, the most likely outcome is that each property will vary as widely within each subset as it does in the whole set (taking different sample sizes into account). Note that cases within a subset are to be treated *as if alike* for analysis purposes, and that variation within subset functions as random error or "noise." So, if *any* of these properties have *any* effects on the X-Y relation being examined, then Treatment R will act to increase noise (relative to the X-Y "signal") and thereby to reduce the chances of detecting the X-Y signal even if it is truly "there."

SOME DILEMMAS IN THE DESIGN DOMAIN

These points bring to the fore some of the dilemmas of research design. First, there is the R dilemma. Randomization is both a

cure and a curse, a bane and a boon. On the one hand, Randomization is costly, does not help reduce certain major threats to internal validity, often poses practical and ethical problems, does not guarantee comparability, and does guarantee high within subset variability (i.e., noise). On the other hand, Randomization is *essential,* because without it one cannot disentangle causal connections for theX-Y relation.

The other treatment operations are also dilemmatic in their effects. The treatment operations (X and K) that give you the most logical leverage regarding the X-Y relation, by cutting down "noise," are the very operations that limit the *scope* of the question, so that resulting information is very much constrained. (This has to do with one aspect of external validity.) So, there is a trade-off between scope (the amount of potential information in the problem) and precision (the amount of reduction of noise). On the other hand, the treatment operations (Y and R) that allow broader generalization from results are the very ones that incorporate much "noise" into the information that is contained in the set of observations, making it hard to detect "signal" if it is there. This is another trade-off between scope (amount of information in the problem) and precision (amount of noise in the information). Together, these pose the researcher with the following choices: You can reduce noise, by cutting scope; so you can learn more about less. Or, you can leave scope broad, by accepting noise along with signal; in which case you can learn less about more. At the limit, if you constrain scope by manipulating and controlling more and more variables, there will be no potential information left, and you will then be able to learn everything about nothing. That is the case, in fact, for the research strategy called Computer Simulations, where there are lots of X and K treatments, and some R treatments, but no Y treatments—thus *no* potential information *at all.* At the other limit, you can constrain less and less—eskewing X and K and M treatments—to maximize scope (and noise), using Y for some variables of interest and Z for all the rest. Here, at the limit, you can learn little or nothing about everything. This is more or less what happens in that research called Field Studies.

The latter dilemma—information versus noise, or scope versus precision—is one instance of a very pervasive set of dillemmas within the research process. The general class of dilemmas can be characterized as *Standardization versus Generalizability* (which is really replication versus partitioning in disguise!). There is a direct conflict between maximizing two desiderata. (a) On the one hand, it is desirable to maximize *standardization*—of "irrelevant" conditions (time of day, color of walls, and so on), of methods of measurement, and so forth—because from such standardization we hope to *gain precision by reducing noise* (i.e., reducing variations within cell, among cases treated alike). (b) On the other hand, it is equally desirable to maximize the range of conditions over which a relation has been tested—by varying "irrelevant conditions," varying methods of measurement, and the like—because from such heterogeneity we hope to gain increased *generalizability* with regard to those varying properties, and thereby gain heightened confidence in the breadth and robustness of the X-Y relation we are assessing.

There is another dilemma that was suggested but not emphasized earlier in this section on design. It has to do with deciding how many different combinations of conditions are to be compared, and how many cases are to be obtained for each combination of conditions. Again, there is a direct conflict between two desiderata. On the one hand, it is *always desirable to increase the number of levels* of an independent variable whose effect is to be studied; and, indeed, it is desirable to test multiple levels of multiple independent variables (that is, it is desirable to increase the total number of partitions). On the other hand, it is *always desirable to increase the number of cases within each subset*, the number of cases to be treated as replications. The latter gives more stability to our estimate of the average value of each subset, each combination of conditions. The former gives more stability to our estimate of the functional relations between variables, X and Y.

But there is always an upper limit to the total number of cases or observations—an upper limit in principle, as well as in practice, because at some point "new" observations must be

viewed as "different from" earlier ones. If we take N as the total number of observations, k as the total number of combinations of conditions, or number of "cells," and m as the number of observations in each "cell," then: It is inexorably the case that N = km. If there is a fixed N, then any operation that increases either m or k—both of which are desiderata to be maximized—will inevitably decrease the other.

I am sure the reader will see that the four major dilemmas described here—the dilemma of R, the precision versus scope or information versus noise dilemma, the standardization/generalizability dilemma, and the N/k/m dilemma—are really all related to one another, and to the dilemmas of the strategy level (and, it will turn out, to those of the methods level, our next topic). I am sure the reader will also see that, at the heart of all these matters is the Campbellian matrix of designs, forms of validity, and classes of threats to validity. Equally at the heart of these matters is the Cronbachian treatment of the "two disciplines of psychology," experimental and correlational. What is offered here is a Dilemmatic view of these matters, one that points to the *inherent limitations of all choices* in the design domain—as well as in the strategy domain and, yet to come, in the method domain. I would hope, with such a view, to discourage the reader from seeking "the right design", either in general or for a particular problem; and encourage him or her, instead, to *accept the inevitable limitations and dilemmas of our methods* as constraints within which we must work, and therefore to set out to do the best we can with what we've got!

DILEMMAS AT THE METHODS LEVEL

The third domain of methods deals with how we can measure, manipulate, control, and otherwise contend with variables. Here we find some ties with what has gone before. For one thing, these map to the treatments of properties discussed under Design. Y Treatment is measurement; X is manipulation; K is experimental control; and M, R, and alas, Z are ways of "contending with"

other variables. So, we have already said much about these methods. Moreover, we again find a strong Campbellian influence. That is particularly so in regard to the ideas of convergent and discriminant validity, multiple operationalism, and unobtrusive measurement (see Campbell and Fiske, 1959; Webb et al., 1966). That work has had a strong influence on the material to be presented here.

RELIABILITY, VALIDITY, AND GENERALIZABILITY

Psychology has had a long history of concern with the reliability and validity of measures, and that work has led to the unfolding of many complexities in these concepts. While much progress has been made, it is clear that these matters are far from settled. But in spite of all the complications and unsettled issues, some matters are now clear. One of these is the need for multiple operations—in the interest of assessing both reliability and validity. Another is the importance of seeking convergence among measures that differ in their methodological weaknesses. A third is that the same problems, and requirements, exist with regard to reliability and validity for manipulations of independent variables as for measurement of dependent variables. A fourth is that we cannot determine the validity and reliability of constructs without reference to a network of relations involving those constructs; but we cannot establish the reliability and validity of such relations between constructs without reference to the validity and reliability of the individual constructs. These points suggest that there are some method-level dilemmas to be explored. They also suggest that we might adopt Commoner's second law of ecology as our next rule of Dilemmatics:

RULE VI: You can't do *one* thing.

THE CONSTRUCT VALIDITY DILEMMA

When we want to test the relation between constructs X and Y, we develop an operational definition (x) for construct X, and an

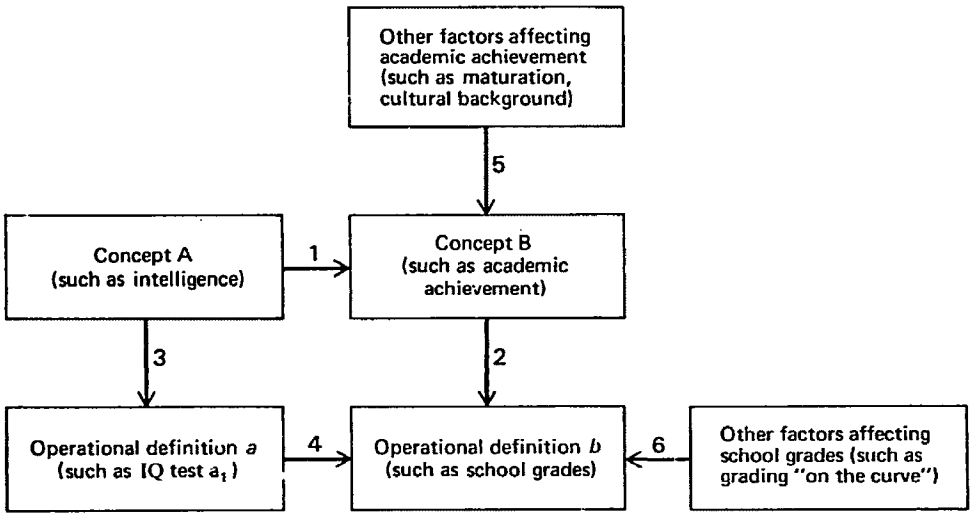


Figure 5: The Rationale for Predictive Validity
From Runkel and McGrath, 1972

operational definition (y) for construct Y; then find some setting in which to test the empirical relation between x and y, so that we can draw an inference about X in relation to Y. There are four relations involved here (see Figure 5). One, X-Y, is conceptual and cannot be tested empirically. Two are definitional, X-x and Y-y, and they can only be tested indirectly. The fourth is an empirical relation, x-y; and it is used as the empirical lever by which all three of the others can be assessed.

If we are hypothesis testing, then what we want to know is about the conceptual relation, as in the example above, (relation 1 in Figure 7), X-Y. If so, then we must assume that relations 2 (Y-y) and 3 (X-x) hold, so that we can use relation 4 (x-y) to test relation 1 (X-Y). If x-y is strong, then (subject to all the constraints already discussed about design and inference) we might interpret it as evidence for X-Y. But what if x-y does not hold? This could occur because x is a poor mapping of X (relation 3), or y is a poor mapping of Y (relation 2) or because the x-y data at hand is a poor set of evidence, as well as because the X-Y conceptual relation does not hold.

If what one had in mind, instead, was to develop a good measure of X, then one might *do* the same things (build an x, build or

choose a y, and test x-y empirically), but interpret the x-y outcome as having to do with the X-x relation—this time assuming X-Y, and Y-y. Suppose X were "intelligence," x were some particular test of intelligence, Y were "academic performance" and y were some operational definition of that, such as GPA. If we correlated scores on an IQ test with GPA and found a strong, positive association, we might conclude favorably about X-x (this is a good IQ test), Y-y (GPA is a good index of academic performance), X-Y (intelligence forecasts academic performance). But we only have evidence for any one of them if we assume that the other two are true. And, since the empirical operations are the same in all these cases, we can equally reasonably assume any two of them and test the third. Which of these we think we are testing, and which we are assuming, is entirely arbitrary and depends on the purposes of the investigator.

One point to make here is that all knowledge is contingent on assumptions. To test something, you must assume some other things; and results of your test hold only to the extent that those assumptions hold. Another point to be made is that construct validation (X-x and Y-y), predictive validity (x-y), and testing theoretical hypotheses (X-Y) are all part of the same enterprise. You are either doing all of them, or none of them. Again, you can't do *one* thing.

CONVERGENT AND DISCRIMINANT VALIDITY

Campbell and Fiske (1959) made it clear more than 20 years ago that we gain validity by means of *convergence* of different measures. It is necessary to have more than one measure of a construct so that the "method variance" associated with any one measure can be separated out. And it is necessary that the multiple measures be of different types, so that the weaknesses of any one type of method can be counter-vailed by coupling it with other methods that have different weaknesses (Webb et al., 1966). So while multiple operations, all of the same type of method, may help establish reliability, they are not as useful in establishing construct validity. Furthermore, to be confident about our

constructs we not only have to establish reliability by correlation of multiple operational definitions of the construct, and establish convergent validity by correlation of different methods of measuring the construct; we also have to establish discriminant validity, showing the boundaries of the construct, by showing *lack* of correlation of measures of the construct with methodologically similar measures of substantively distinct constructs. To know what the construct is (convergent validity) we have to have some knowledge of what it is not (discriminant validity).

This poses another dilemma for the beleaguered researcher, one that is hard to present clearly. It hinges on simultaneously considering convergent and discriminant validity of a construct and hypothesis testing of the relations of that construct with some other construct. When two measures are very similar in form and substance, we tend to think of them as alternative forms of the same measures and, if they correlate highly, regard that as evidence of reliability of that construct. If two measures differ in form but are similar in substance, we might well regard them as alternative measures of the same construct and regard their correlation as evidence of convergent validity. But if two measures differ in substance, we are not altogether sure how to regard them. If they fail to correlate, we might regard that lack of correlation as evidence of discriminant validity of one of the constructs. But if they correlate highly, or even moderately, we might regard that correlation as evidence for a relation between two different constructs, as in substantive hypothesis testing. This set of considerations reminds us that "same" and "different" decisions are made at several levels within the research process, and that they are arbitrary. If two measures are too similar, their high correlation is not remarkable, and is regarded as "merely" a reliability. If two measures are too dissimilar, and they don't correlate, that too is not remarkable, and may be regarded as "merely" evidence of discrimination. But if two measures are different, and they correlate, that is regarded as remarkable, and we often take it to be evidence supporting some substantive hypothesis. This state of affairs is not so much a dilemma as it is a paradox or a quandry. It suggests Dilemmatics Rule 7.

(1) WHO PERFORMS THE BEHAVIOR UNDER STUDY? ALWAYS THE ACTOR.

(2) WHO OBSERVES AND RECORDS THE BEHAVIOR?	(3) IS THE ACTOR AWARE THAT HIS BEHAVIOR IS BEING RECORDED FOR RESEARCH?	
	Yes: Observation May Be Reactive	*No: Observation is Nonreactive*
Actor	Subjective reports	Traces
Researcher	Visible observer	Hidden observer
Recorder in the past	Records of public behavior	Archival records

(4) WHO TRANSLATES RECORDS INTO DATA? ALWAYS THE RESEARCHER.

Figure 6: Sources of Empirical Evidence
From Runkel and McGrath, 1972

RULE VII: One person's substantive finding is another's method variance.

CLASSES OF MEASURES

Building upon ideas in Webb et al. (1966), Runkel and McGrath (1972) discuss six classes of measures, arrayed on two factors: (a) whether or not the measurement is occurring beknownst to the actor whose behavior is being recorded, and therefore the extent to which the measure is obtrusive, and liable to have reactive effects on the behavior being studied; and, (b) whether the behavior record is being made by the actor, by the investigator (or a person or instrument serving as the investigator's surrogate), or by a third party (see Figure 6).

Each of the six classes offers strengths and weaknesses. Subjective-reports or self-reports—the most popular of the six classes—tempts the researcher to swallow one deadly flaw (reactivity) by wrapping that flaw in several tantalizing strong points (content versatility, low cost, low dross rate, and so on). Trace measures, so attractive in principle because of their unobtrusiveness, are difficult and costly to devise, and often have only a loose coupling to the intended constructs. Observations have a strong appeal as measures of "actual behavior." But observation by a visible observer combines all the reactivity problems of self-

reports with some additional observer-based sources of error; while use of hidden observers trades some of those reactivity problems for some practical and ethical—and perhaps legal—problems involved in deceptive strategies. Public records and documents may have a high or low dross rate and cost; and may or may not have problems of reactivity even though they were not recorded as part of a research effort. They resemble trace measures in sometimes being only loosely coupled to the constructs they are used to measure. They can be very valuable, though, when and if appropriate ones are available, because they often offer the best available evidence about very large systems and about the past. Figure 7 lists some of the strengths and weaknesses of these six classes of measures.

The central point here, of course, is that one *must use multiple methods, selected from different classes of methods with different vulnerabilities*—not only because this is needed to establish convergent validity, but also because it is needed in order to "finesse" the various threats to validity to which the various classes of methods are vulnerable. The researcher is not posed with a problem of choosing *which one* class of measures to use; the problem is, rather, to choose—and often to devise—a set of measures of each construct that, together, transcend one another's methodological vulnerabilities. This is a problem of creativity and cost, but it is not a true dilemma, since except for cost the desiderata involved (convergence, multiple-methods) are not in conflict. After all that has been said in previous pages, it is nice to meet a set of concerns that is "merely" a problem, one that can be solved by increased resources rather than a dilemma that can't be solved at all!

CONCLUDING COMMENTS

This article has pointed out some dilemmas—conflicting desiderata—in the strategy, design, and methods domains. It has not said anything about data collection (that is, the actual conduct of the data-getting operations), or data analysis. Some dilemmas are

Sources of Invalidity of Methods; That Is, Plausible Rival Hypotheses	METHODS OF OBSERVING AND RECORDING					
	Actor's Records		*Researcher's Records*		*Previous Records*	
	Subjective Reports	*Trace Measures*	*Visible Observer*	*Hidden Observer*	*Public Behavior*	*Archival Records*
Biases associated with the actor						
1. Guinea pig effect	H		H		M	
2. Role selection	H		H		M	
3. Measurement as change agent	H		H		M	
4. Response sets	H	M	H	M	M	M
Biases associated with the investigator						
5. Effects of interviewer or observer			H	H	M	M
6. Instrument change	H	M	H	H	M	M
Biases associated with the population						
7. Population restrictions	H	H	M	M	M	M
8. Population instability over time	M		M		M	M
9. Population instability among areas	H				M	M
Biases associated with content						
10. Content restrictions		H			M	M
11. Content instability over time		M			H	H
12. Content instability over areas		M			H	H
Other characteristics of methods						
13. Dross rate		H	M	M	M	M
14. Difficulty of access to secondary data		H				
15. Difficulty of replication		H				

KEY: H: High vulnerability to the bias or rival hypothesis.
M: Moderate vulnerability.
Absence of a symbol indicates low vulnerability.

Adapted from Webb and others, 1966.

Figure 7: Vulnerabilities of Methods of Observing and Recording

hidden in those domains, too. It also has said little about the interpretation stage, where there are some colossal dilemmas waiting. But enough has been said to make clear the central points of a "Dilemmatic" view of research:

1. The research process teems with dilemmas involving the need to maximize, simultaneously, two or, in some cases, three conflicting desiderata.

2. The researcher cannot avoid choosing, nor can he or she find a no lose strategy, nor a "compromise" that doesn't minimize some other desideratum.

3. One cannot plan, or execute, flawless research. All strategies, all designs, all methods, are seriously—even fatally—flawed.

4. No strategy, design or method *used alone* is worth a damn. Multiple approaches are *required*—at the method level, within study for every construct; at the design and strategy levels, between studies.

5. Multiple methods not only serve the purposes of replication and convergence; they serve the further, crucial purpose of compensating for inherent limitations that any one method, strategy, or design would have if used alone.

The researcher has a hard and thankless lot! He or she faces a stark set of no-win choices. But the researcher who is fully aware of the dilemmas, who is fully armed with possibilities, can handle the dilemmas. The dilemmas can be handled, *not* by trying to avoid the choices; not by trying to pretend the dilemmas don't exist; certainly not by seeking "the right choices," of strategy, design, and method. They can be handled by bowling them over with multiple methods for all constructs, embedded in multiple designs, using multiple strategies, to gain information about the research problems of concern. We should end this discussion by stating several Final Rules of Dilemmatics, which themselves present one final dilemmatic puzzle:

FINAL RULES: There is no such thing as too much research!
There is no such thing as flawless research!
But: Poor research is *much worse* than none at all.

The key to that final paradox lies in the dual meanings of good and poor research. We must distinguish between the inherent flaws of any method, when used as well as it can be used, and the quite different matter of using a method badly. The former, the inherent flaws of any method, even when used well, are neither to be decried nor to be overlooked, but rather to be made explicit. The latter—using a method badly—is never acceptable. It is that

that is referred to as "poor research" in the final rules. So, while flawless research is not possible, poor research—using potentially valuable-though-flawed methods badly—makes matters worse than they need be. But "good research"—using flawed methods well, and in effective combinations—can help us accrue "knowledge" about behavioral and social science problems that are of both theoretical and practical concern.

REFERENCES

CAMPBELL, D. T. and J. L. STANLEY (1966) Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally.

COOK, D. T. and D. T. CAMPBELL (1979) Quasi-experimental Design: Design and Analysis Issues for Field Settings. Chicago: Rand McNally.

CRONBACH, L. J. (1957) "The two disciplines of scientific psychology." Amer. Psychologist 12: 671-684.

RUNKEL, P. J. and J. E. McGRATH (1972) Research on Human Behavior: A Systematic Guide to Method. New York: Holt.

WEBB, E. J., D. T. CAMPBELL, R. D. SCHWARTZ, and L. SECHREST (1966) Unobtrusive Measures: Non Reactive Research in the Social Sciences. Chicago: Rand McNally.