Operationism and Multiple Operations

The social sciences are just emerging from a period in which the precision of carefully specified operations was confused with operationism by definitional fiat—an effort now increasingly recognized as an unworkable model for science. We wish to retain and augment the precision without bowing to the fiat.

The mistaken belief in the operational definition of theoretical terms has permitted social scientists a complacent and self-defeating dependence upon single classes of measurement—usually the interview or questionnaire. Yet the operational implication of the inevitable theoretical complexity of every measure is exactly opposite; it calls for a multiple operationism, that is, for multiple measures which are hypothesized to share in the theoretically relevant components but have different patterns of irrelevant components (e.g., Campbell, 1960; Campbell & Fiske, 1959; Garner, 1954; Garner, Hake, & Eriksen, 1956; Humphreys, 1960).

Once a proposition has been confirmed by two or more independent measurement processes, the uncertainty of its interpretation is greatly reduced. The most persuasive evidence comes through a triangulation of measurement processes. If a proposition can survive the onslaught of a series of imperfect measures, with all their irrelevant error, confidence should be placed in it. Of course, this confidence is increased by minimizing error in each instrument and by a reasonable belief in the different and divergent effects of the sources of error.

A consideration of the laws of physics, as they are seen in that science's measuring instruments, demonstrates that no theoretical parameter is ever measured independently of other physical parameters and other physical laws. Thus, a typical galvanometer responds in its operational measurement of voltage not only according to the laws of electricity but also to the laws of gravitation, inertia, and friction. By reducing the mass of the galvanometer needle, by orienting the needle's motion at right angles to gravity, by setting the needle's axis in jeweled bearings, by counterweighting the needle point, and by other refinements, the instrument designer attempts to minimize the most important of the irrelevant physical forces for his measurement purposes. As a result, the galvanometer reading may reflect, *almost* purely, the single parameter of voltage (or amperage, etc.).

Yet from a theoretical point of view, the movement of the needle is always a complex product of many physical forces and laws. The adequacy with which the needle measures the conceptually defined variable is a matter for investigation; the operation itself is not the ultimate basis for defining the variable. Excellent illustrations of the specific imperfections of measuring instruments are provided by Wilson (1952).

Starting with this example from physics and the construction of meters, we can see that no meter ever perfectly measures a single theoretical parameter; all series of meter readings are imperfect estimates of the theoretical parameters they are intended to measure.

Truisms perhaps, yet they belie the mistaken concept of the “operational definition” of theoretical constructs which continues to be popular in the social sciences. The inappropriateness is accentuated in the social sciences because we have no measuring devices as carefully compensated to control all irrelevancies as is the galvanometer. There simply are no social science devices designed with so perfect a knowledge of all the major relevant sources of variation. In physics, the instruments we think of as “definitional” reflect magnificently successful theoretical achievements and themselves embody classical experiments in their very operation. In the social sciences, our measures lack such control. They tap multiple processes and sources of variance of which we are as yet unaware. At such a stage of development, the theoretical impurity and factorial complexity of every measure are not niceties for pedantic quibbling but are overwhelmingly and centrally relevant in all measurement applications which involve inference and generalization.

Efforts in the social sciences at multiple confirmation often yield disappointing and inconsistent results. Awkward to write up and difficult to publish, such results confirm the gravity of the problem and the risk of false confidence that comes with dependence upon single methods (Campbell, 1957; Campbell & Fiske, 1959; Campbell & McCormack, 1957; Cook & Selltiz, 1964; Kendall, 1963; Vidich & Shapiro, 1955). When multiple operations provide consistent results, the possibility of slippage between conceptual definition and operational specification is diminished greatly.

This is not to suggest that all components of a multimethod approach should be weighted equally. Prosser (1964) has observed: “… but there is still no man who would not accept dog tracks in the mud against the sworn, testimony of a hundred eye-witnesses that no dog had passed by” (p. 216). Components ideally should be weighted according to the amount of extraneous variation each is

known to have and, taken in combination, according to their independence from similar sources of bias.

Interpretable Comparisons and Plausible Rival Hypotheses

In this monograph, we deal with methods of measurement appropriate to a wide range of social science studies. Some of these studies are comparisons of a single group or unit at two or more points in time; others compare several groups or units at one time; others purport to measure but a single unit at a single point in time; and, to close the circle, some compare several groups at two or more points in time. In this discussion, we assume that the goal of the social scientist is always to achieve interpretable comparisons, and that the goal of methodology is to rule out those plausible rival hypotheses which make comparisons ambiguous and tentative.

Often it seems that absolute measurement *is* involved, and that a social instance is being described in its splendid isolation, not for comparative purposes. But a closer look shows that absolute, isolated measurement is meaningless. In all useful measurement, an implicit comparison exists when an explicit one is not visible. “Absolute” measurement is a convenient fiction and usually is nothing more than a shorthand summary in settings where plausible rival hypotheses are either unimportant or so few, specific, and well known as to be taken into account habitually. Thus, when we report a length “absolutely” in meters or feet, we immediately imply comparisons with numerous familiar objects of known length, as well as comparisons with a standard preserved in some Paris or Washington sanctuary.

If measurement is regarded always as a comparison, there are three classes of approaches which have come to be used in achieving interpretable comparisons. First, and most satisfactory, is experimental design. Through deliberate randomization, the *ceteris* of the pious *ceteris paribus* prayer can be made *paribus.* This may require randomization of respondents, occasions, or stimulus objects. In any event, the randomization strips of plausibility many of the otherwise available explanations of the difference in question. It is a sad truth that randomized experimental design is possible for only a portion of the settings in which social scientists make measurements and seek interpretable comparisons. The number of opportunities for its use may not be staggering, but, where possible, experimental design should by all means be exploited. Many more opportunities exist than are used.

Second, a quite different and historically isolated tradition of comparison is that of index numbers. Here, sources of variance known to be irrelevant are controlled by transformations of raw data and weighted aggregates. This is analogous to the compensated and counterbalanced meters of physical science which also control irrelevant sources of variance. The goal of this old and currently neglected social science tradition is to provide measures for meaningful comparisons across wide spans of time and social space. Real wages, intelligence quotients, and net reproductive rates are examples, but an effort in this direction is made even when a percentage, a per capita, or an annual rate is computed. Index numbers cannot be used uncritically because the imperfect knowledge of the laws invoked in any such measurement situation precludes computing any effective all-purpose measures.

Furthermore, the use of complex compensated indices in the assurance that they measure what they are devised for has in many instances proved quite misleading. A notable example is found in the definitional confusion surrounding the labor force concept (Jaffe & Stewart, 1951; Moore, 1953). Often a relationship established between an over-all index and external variables is found due to only one component of the index. Cronbach (1958) has described this problem well in his discussion of dyadic scores of interpersonal perception. In the older methodological literature, the problem is raised under the term *index correlations* (e.g., Campbell, 1955; Guilford, 1954; Stouffer, 1934).

Despite these limitations, the problem of index numbers, which once loomed large in sociology and economics, deserves to be reactivated and integrated into modern social science methodology. The tradition is relevant in two ways for the problems of this monograph. Many of the sources of data suggested here, particularly secondary records, require a transformation of the raw data if they are to be interpretable in any but truly experimental situations. Such transformations should be performed with the wisdom accumulated within the older tradition, as well as with a regard for the precautionary literature just cited. Properly done, such transformations often improve interpretability even if they fall far short of some ideal (cf. Bernstein, 1935).

A second value of the literature on index numbers lies in an examination of the types of irrelevant variation which the index computation sought to exclude. The construction of index numbers is usually a response to criticisms of less sophisticated indices. They thus embody a summary of the often unrecorded criticisms of prior measures. In the criticisms and the corrections are clues to implicit or explicit plausible rival interpretations of differences, the viable threats to valid interpretation.

Take so simple a measure as an index on unemployment or of retail sales. The gross number of the unemployed or the gross total dollar level of sales is useless if one wants to make comparisons within a single year. Some of the objections to the gross figures are reflected in the seasonal corrections applied to time-series data. If we look at only the last quarter of the year, we can see that the effect of weather must be considered. Systematically, winter depresses the number of employed construction workers, for example, and increases the unemployment level. Less systematically, spells of bad weather keep people in their homes and reduce the amount of retail shopping. Both periodic and aperiodic elements of the weather should be considered if one wants a more stable and interpretable measure of unemployment or sales. So, too, our custom of giving gifts at Christmas spurs December sales, as does the coinciding custom of Christmas bonuses to employees. All of these are accounted for, crudely, by a correction applied to the gross levels for either December or the final quarter of the year.

Some of these sources of invalidity are too specific to a single setting to be generalized usefully; others are too obvious to be catalogued. But some contribute to a general enumeration of recurrent threats to valid interpretation in social science measures.

The technical problems of index-number construction are heroic. “The index number should give *consistent* results for different base periods and also with its counterpart price or quantity index. No reasonably simple formula satisfies both of these consistency requirements” (Ekelblad, 1962, p. 726). The consistency problem is usually met by substituting a geometric mean for an arithmetic one, but then other problems arise. With complex indices of many components, there is the issue of getting an index that will yield consistent scores across all the different levels and times of the components.

In his important work on economic cycles, Hansen (1921) wrote, “Here is a heterogeneous group of statistical series all of which are related in a causal way, somehow or another, to the cycle of prosperity and depression” (p. 21). The search for a metric to relate these different components consistently, to be able to reverse factors without chaos, makes index construction a difficult task. But the payoff is great, and the best approximation to solving both the base-reversal and factor-reversal issues is a weighted aggregate with time-averaged weights. For good introductory statements of these and other index-number issues, see Ekelblad (1962), Yule and Kendall (1950), and Zeisel (1957). More detailed treatments can be found in Fisher (1923), Mills (1927), Mitchell (1921), and Mudgett (1951).

The third general approach to comparison may be called that of “plausible rival hypotheses.” It is the most general and least formal of the three and is applicable to the other two. Given a comparison which a social scientist wishes to interpret, this approach asks what other plausible interpretations are allowed by the research setting and the measurement processes. The more of these, and the more plausible each is, the less validly interpretable is the comparison. Platt (1964) and Hafner and Presswood (1965) have discussed this approach with a focus in the physical sciences.

A social scientist may reduce the number of plausible rival hypotheses in many ways. Experimental methods and adequate indices serve as useful devices for eliminating some rival interpretations. A checklist of commonly relevant threats to validity may point to other ways of limiting the number of viable alternative hypotheses. For some major threats, it is often possible to provide supplementary analyses or to assemble additional data which can rule out a source of possible invalidity.

Backstopping the individual scientist is the critical reaction of his fellow scientists. Where he misses plausible rival hypothesis, he can expect his colleagues to propose alternative interpretations. This resource is available even in disciplines which are not avowedly scientific. J. H. Wigmore (1937), a distinguished legal scholar, showed an awareness of the criteria of other plausible explanations of data:

> If the potential defect of Inductive Evidence is that the fact offered as the basis of the conclusion may be open to one or more other explanations or inferences, the failure to exclude a single other rational inference would be, from the standpoint of *Proof*, a fatal defect; and yet, if only that single other inference were open, there might still be an extremely high degree of probability for the Inference desired … The provisional test, then, from the point of view valuing the Inference, would be something like this: *Does the evidentiary fact point to the desired conclusion … as the inference … most plausible or most natural out of the various ones that are conceivable?* (p. 25)

The culture of science seeks, however, to systematize the production of rival plausible hypotheses and to extend them to every generalization proposed. While this may be implicit in a field such as law, scientific epistemology requires that the original and competing hypotheses be explicitly and generally stated.

Such a commitment could lead to rampant uncertainty unless some criterion of plausibility was adopted before the rival hypothesis was taken as a serious alternative. Accordingly, each rival hypothesis is a threat only if we can give it the status of a law at least as creditable as the law we seek to demonstrate. If it falls short of that credibility, it is not thereby “plausible” and can be ignored.

In some logical sense, even in a “true” experimental comparison, an infinite number of potential laws could predict this result. We do not let this logical state of affairs prevent us from interpreting the results. Instead, uncertainty comes only from unexcluded hypotheses to which we, in the current state of our science, are willing to give the status of established laws: these are the plausible rival hypotheses. While the north-south orientation of planaria may have something to do with conditioning, no interview studies report on the directional orientation of interviewer and interviewee. And they should not.

For those plausible rival hypotheses to which we give the status of laws, the conditions under which they would explain our obtained result also imply specific outcomes for other sets of data. Tests in other settings, attempting to verify these laws, may enable us to rule them out. In a similar fashion, the theory we seek to test has many implications other than that involved in the specific comparison, and the exploration of these is likewise demanded. The more numerous and complex the manifestations of the law, the fewer singular plausible rival hypotheses are available, and the more parsimony favors the law under study.

Our longing is for data that prove and certify theory, but such is not to be our lot. Some comfort may come from the observation that this is not an existential predicament unique to social science. The replacement of Newtonian theory by relativity and quantum mechanics shows us that even the best of physical science experimentation probes theory rather than proves it. Modern philosophies of science as presented by Popper (1935, 1959, 1962), Quine (1953), Hanson (1958), Kuhn (1962), and Campbell (1965a, 1965b), make this point clear.

Internal and External Validity

Before discussing a list of some common sources of invalidity, a distinction must be drawn between internal and external validity. *Internal validity* asks whether a difference exists at all in any given comparison. It asks whether or not an apparent difference can be explained away as some

measurement artifact. For true experiments, this question is usually not salient, but even there, the happy vagaries of random sample selection occasionally delude one and spuriously produce the appearance of a difference where in fact none exists. For the rival hypothesis of chance, we fortunately have an elaborated theoretical model which evaluates its plausibility. A p-value describes the darkness of the ever present shadow of doubt. But for index-number comparisons not embedded in a formal experiment, and for the plausible-rival-hypothesis strategy more generally, the threats to internal validity—the argument that even the appearance of a difference is spurious—is a serious problem and the one that has first priority.

External validity is the problem of interpreting the difference, the problem of generalization. To what other populations, occasions, stimulus objects, and measures may the obtained results be applied? The distinction between internal and external validity can be illustrated in two uses of randomization. When the experimentalist in psychology randomly assigns a sample of persons into two or more experimental groups, he is concerned entirely with internal validity—with making it implausible that the luck of the draw produced the resulting differences. When a sociologist carefully randomizes the selection of respondents so that his sample represents a larger population, representativeness or external validity is involved.

The psychologist may be extremely confident that a difference is traceable to an experimental treatment, but whether it would hold up with another set of subjects or in a different setting may be quite equivocal. He has achieved internal validity by his random assignment but not addressed the external validity issue by the chance allocation of subjects.

The sociologist, similarly, has not met all the validity concerns by simply drawing a random sample. Conceding that he has taken a necessary step toward achieving external validity and generalization of his differences, the internal validity problem remains.

Random assignment is only one method of reaching toward internal validity. Experimental-design control, exclusive of randomization, is another. Consider the case of a pretest-posttest field experiment on the effect of a persuasive communication. Randomly choosing those who participate, the social scientist properly wards off some major threats to external validity. But we also know of other validity threats. The first interview in a two-stage study may set into motion attitude change and clarification processes which would otherwise not have occurred (e.g., Crespi, 1948). If such

processes did occur, the comparison of a first and second measure on the same person is internally invalid, for the shift is a measurement-produced artifact.

Even when a measured control group is used, and a persuasive communication produces a greater change in an experimental group, the persuasive effect may be internally valid but externally invalid. There is the substantial risk that the effect occurs only with pretested populations and might be absent in populations lacking the pretest (cf. Hovland, Lumsdaine, & Sheffield, 1949; Schanck & Goodman, 1939; Solomon, 1949). For more extensive discussions of internal and external validity, see Campbell (1957) and Campbell and Stanley (1963).

The distinction between internal and external validity is often murky. In this work, we have considered the two classes of threat jointly, although occasionally detailing the risks separately. The reason for this is that the factors which are a risk for internal validity are often the same as those threatening external validity. While for one scientist the representative sampling of cities is a method to achieve generalization to the United States population, for another it may be an effort to give an internally valid comparison across cities.