Sage Books

Unobtrusive Measures, Revised Edition

Pub. Date: 2012

Product: Sage Books

DOI: https://doi.org/10.4135/9781452243443

Keywords: internal validity, external validity, interviews, validity, sampling, indexing, measurement

Disciplines: Simulation & Gaming, Research Methods for the Social Sciences, Evaluation, Research Methods for Sociology, Social Research, Sociology

Access Date: May 23, 2023

Publishing Company: SAGE Publications, Inc.

City: Thousand Oaks

Online ISBN: 9781452243443

Approximations to Knowledge

This survey directs attention to social science research data not obtained by interview or questionnaire. Some may think this exclusion does not leave much. It does. Many innovations in research method are to be found scattered throughout the social science literature. Their use, however, is unsystematic, their importance understated. Our review of this material is intended to broaden the social scientist's currently narrow range of utilized methodologies and to encourage creative and opportunistic exploitation of unique measurement possibilities.

Today, the dominant mass of social science research is based upon interviews and questionnaires. We lament this overdependence upon a single, fallible method. Interviews and questionnaires intrude as a foreign element into the social setting they would describe, they create as well as measure attitudes, they elicit atypical roles and responses, they are limited to those who are accessible and will cooperate, and the responses obtained are produced in part by dimensions of individual differences irrelevant to the topic at hand.

But the principal objection is that they are used alone. No research method is without bias. Interviews and questionnaires must be supplemented by methods testing the same social science variables but having *different* methodological weaknesses.

In sampling the range of alternative approaches, we examine their weaknesses, too. The flaws are serious and give insight into why we do depend so much upon the interview. But the issue is not choosing among individual methods. Rather, it is the necessity for a multiple operationism, a collection of methods combined to avoid sharing the same weaknesses. The goal of this monograph is not to replace the interview but to supplement and cross-validate it with measures that do not require the cooperation of a respondent and that do not themselves contaminate the response.

Here are some samples of the kinds of methods we will be surveying in Chapters 2 through 6 of this monograph:

- The floor tiles around the hatching-chick exhibit at Chicago's Museum of Science and Industry must be replaced every six weeks. Tiles in other parts of the museum need not be replaced for years. The selective erosion of tiles, indexed by the replacement rate, is a measure of the relative popularity of exhibits.
- The accretion rate is another measure. One investigator wanted to learn the level of whisky consumption in a town which was officially “dry.” He did so by counting empty bottles in ashcans.

- The degree of fear induced by a ghost-story-telling session can be measured by noting the shrinking diameter of a circle of seated children.
- Chinese jade dealers have used the pupil dilation of their customers as a measure of the client's interest in particular stones, and Darwin in 1872 noted this same variable as an index of fear.
- Library withdrawals were used to demonstrate the effect of the introduction of television into a community. Fiction titles dropped, nonfiction titles were unaffected.
- The role of rate of interaction in managerial recruitment is shown by the overrepresentation of baseball managers who were infielders or catchers (high-interaction positions) during their playing days.
- Sir Francis Galton employed surveying hardware to estimate the bodily dimensions of African women whose language he did not speak.
- The child's interest in Christmas was demonstrated by distortions in the size of Santa Claus drawings.
- Racial attitudes in two colleges were compared by noting the degree of clustering of Negroes and whites in lecture halls.

These methods have been grouped into chapters by the characteristic of the data: physical traces, archives, observations.

Before making a detailed examination of such methods, it is well to present a closer argument for the use of multiple methods and to present a methodological framework within which both the traditional and the more novel methods can be evaluated.

The reader may skip directly to Sherlock Holmes and the opening of Chapter 2 if he elects, infer the criteria in a piece of detection himself, and then return for a validity check.

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 summa-

ry 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 econom-

ics, 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.

Sources of Invalidity of Measures

In this section, we review frequent threats to the valid interpretation of a difference—common plausible rival hypotheses. They are broadly divided into three groups: error that may be traced to those being studied, error that comes from the investigator, and error associated with sampling imperfections. This section is the only one in which we draw illustrations mainly from the most popular methods of current social science. For that reason, particular attention is paid to those weaknesses which create the need for multiple and alternate methods.

In addition, some other criteria such as the efficiency of the research instrument are mentioned. These are independent of validity, but important for the practical research decisions which must be made.

Reactive Measurement Effect: Error from the Respondent

The most understated risk to valid interpretation is the error produced by the respondent. Even when he is well

mentioned and cooperative, the research subject's knowledge that he is participating in a scholarly search may confound the investigator's data. Four classes of this error are discussed here: awareness of being tested, role selection, measurement as a change agent, and response sets.

1. *The Guinea Pig Effect—Awareness of Being Tested.* Selltiz, Jahoda, Deutsch, and Cook (1959) make this observation:

   > **The measurement process used in the experiment may itself affect the outcome. If people feel that they are "guinea pigs" being experimented with, or if they feel that they are being "tested" and must make a good impression, or if the method of data collection suggests responses or stimulates an interest the subject did not previously feel, the measuring process may distort the experimental results. (p. 97)**

These effects have been called "reactive effect of measurement" and "reactive arrangement" bias (Campbell, 1957; Campbell & Stanley, 1963). It is important to note early that the awareness of testing need not, by itself, contaminate responses. It is a question of probabilities, but the probability of bias is high in any study in which a respondent is aware of his subject status.

Although the methods to be reviewed here do not involve "respondents," comparable reactive effects on the population may often occur. Consider, for example, a potentially nonreactive instrument such as the movie camera. If it is conspicuously placed, its lack of ability to talk to the subjects doesn't help us much. The visible presence of the camera undoubtedly changes behavior, and does so differentially depending upon the labeling involved. The response is likely to vary if the camera has printed on its side "Los Angeles Police Department" or "NBC" or "Foundation Project on Crowd Behavior." Similarly, an Englishman's presence at a wedding in Africa exerts a much more reactive effect on the proceedings than it would on the Sussex Downs.

A specific illustration may be of value. In the summer of 1952, some graduate students in the social sciences at the University of Chicago were employed to observe the numbers of Negroes and whites in stores, restaurants, bars, theaters, and so on on a south side Chicago street intersecting the Negro-white boundary (East 63rd). This, presumably, should have been a nonreactive process, particularly at the predominantly white end of the street. No questions were asked, no persons stopped. Yet, in spite of this hopefully inconspicuous activity, two merchants were agitated and persistent enough to place calls to the university which somehow got through to the investigators; how many others tried and failed cannot be known. The two calls were from a store operator and the manager of a currency exchange, both of whom wanted assurance that this was some

university nosiness and not a professional casing for subsequent robbery (Campbell & Mack, 1966). An intrusion conspicuous enough to arouse such an energetic reaction may also have been conspicuous enough to change behavior; for observations other than simple enumerations the bias would have been great. But even with the simple act of nose-counting, there is the risk that the area would be differentially avoided. The research mistake was in providing observers with clipboards and log sheets, but their appearance might have been still more sinister had they operated Veeder counters with hands jammed in pockets.

The present monograph argues strongly for the use of archival records. Thinking, perhaps, of musty files of bound annual reports of some prior century, one might regard such a method as totally immune to reactive effects. However, were one to make use of precinct police blotters, going around to copy off data once each month, the quality and nature of the records would almost certainly change. In actual fact, archives are kept indifferently, as a low-priority task, by understaffed bureaucracies. Conscientiousness is often low because of the lack of utilization of the records. The presence of a user can revitalize the process—as well as create anxieties over potentially damaging data (Campbell, 1963). When records are seen as sources of vulnerability, they may be altered systematically. Accounts thought likely to enter into tax audits are an obvious case (Schwartz, 1961), but administrative records (Blau, 1955) and criminal statistics (Kadish, 1964) are equally amenable to this source of distortion. The selective and wholesale rifling of records by ousted political administrations sets an example of potential reactive effects, self-consciousness, and dissembling on the part of archivists.

These reactive effects may threaten both internal and external validity, depending upon the conditions. If it seems plausible that the reactivity was equal in both measures of a comparison, then the threat is to external validity or generalizability, not to internal validity. If the reactive effect is plausibly differential, then it may generate a pseudo-difference. Thus, in a study (Campbell & McCormack, 1957) showing a reduction in authoritarian attitudes over the course of one year's military training, the initial testing was done in conjunction with an official testing program, while the subsequent testing was clearly under external university research auspices. As French (1955) pointed out in another connection, this difference provides a plausible reactive threat jeopardizing the conclusion that any reduction has taken place even for this one group, quite apart from the external validity problems of explanation and generalization. In many interview and questionnaire studies, increased or decreased rapport and increased awareness of the researcher's goals or decreased fear provide plausible alternative explanations of the apparent change recorded.

The common device of guaranteeing anonymity demonstrates concern for the reactive bias, but this concern may lead to validity threats. For example, some test constructors have collected normative data under conditions of anonymity, while the test is likely to be used with the respondent's name signed. Making a response public, or guaranteeing to hide one, will influence the nature of the response. This has been seen for persuasive communications, in the validity of reports of brands purchased, and for the level of antisocial responses. There is a clear link between awareness of being tested and the biases associated with a tendency to answer with socially desirable responses.

The considerations outlined above suggest that reactivity may be selectively troublesome within trials or tests of the experiment. Training trials may accommodate the subject to the task, but a practice effect may exist that either enhances or inhibits the reactive bias. Early responses may be contaminated, later ones not, or vice versa (Underwood, 1957).

Ultimately, the determination of reactive effect depends on validating studies—few examples of which are currently available. Behavior observed under nonreactive conditions must be compared with corresponding behavior in which various potentially reactivity conditions are introduced. Where no difference in direction of relationship occurs, the reactivity factor can be discounted.

In the absence of systematic data of this kind, we have little basis for determining what is and what is not reactive. Existing techniques consist of asking subjects in a posttest interview whether they were affected by the test, were aware of the deception in the experiment, and so forth. While these may sometimes demonstrate a method to be reactive, they may fail to detect many instances in which reactivity is a serious contaminant. Subjects who consciously dissemble during an experiment may do so afterward for the same reasons. And those who are unaware of the effects on them at the time of the research may hardly be counted on for valid reports afterward.

The types of measures surveyed in this monograph have a double importance in overcoming reactivity. In the absence of validation for verbal measures, nonreactive techniques of the kind surveyed here provide ways of avoiding the serious problems faced by more conventional techniques. Given the limiting properties of these “other measures,” however, their greatest utility may inhere in their capacity to provide validation for the more conventional measures.

2. *Role Selection.* Another way in which the respondent's awareness of the research process produces

> differential reaction involves not so much inaccuracy, defense, or dishonesty, but rather a specialized selection from among the many “true” selves or “proper” behaviors available in any respondent.

By singling out an individual to be tested (assuming that being tested is not a normal condition), the experimenter forces upon the subject a role-defining decision—What kind of a person should I be as I answer these questions or do these tasks? In many of the “natural” situations to which the findings are generalized, the subject may not be forced to define his role relative to the behavior. For other situations, he may. Validity decreases as the role assumed in the research setting varies from the usual role present in comparable behavior beyond the research setting. Orne and his colleagues have provided compelling demonstrations of the magnitude of this variable's effect (Orne, 1959, 1962; Orne & Evans, 1965; Orne & Scheibe, 1964). Orne (1962) has noted:

> **The experimental situation is one which takes place within context of an explicit agreement of the subject to participate in a special form of social interaction known as “taking part in an experiment.” Within the context of our culture the roles of subject and experimenter are well understood and carry with them well-defined mutual role expectations. (p. 777)**

Looking at all the cues available to the respondent attempting to puzzle out an appropriate set of roles or behavior, Orne labeled the total of all such cues the “demand characteristics of the experimental situation.” The recent study by Orne and Evans (1965) showed that the alleged antisocial effects induced by hypnosis can be accounted for by the demand characteristics of the research setting. Subjects who were not hypnotized engaged in “antisocial” activities as well as did those who were hypnotized. The behavior of those not hypnotized is traced to social cues that attend the experimental situation and are unrelated to the experimental variable.

The probability of this confounding role assumption varies from one research study to another, of course. The novelty of a test-taking role may be selectively biasing for subjects of different educational levels. Less familiar and comfortable with testing, those with little formal schooling are more likely to produce nonrepresentative behavior. The act of being tested is “more different.” The same sort of distortion risk occurs when subject matter is unusual or novel. Subject matter with which the respondent is unfamiliar may produce uncertainty of which role to select. A role-playing choice is more likely with such new or unexpected material.

Lack of familiarity with tests or with testing materials can influence response in different ways. Responses

may be depressed because of a lack of training with the materials. Or the response level may be distorted as the subject perceives himself in the rare role of expert.

Both unfamiliarity and “expertness” can influence the character as well as the level of response. It is common to find experimental procedures which augment the experting bias. The instruction which reads, “You have been selected as part of a scientifically selected sample … it is important that you answer the questions …” underlines in what a special situation and what a special person the respondent is. The empirical test of the experting hypothesis in field research is the extent of “don't know” replies. One should predict that a set of instructions stressing the importance of the respondent as a member of a “scientifically selected sample” will produce significantly fewer “don't knows” than an instruction set that does not stress the individual's importance.

Although the “special person” set of instructions may increase participation in the project, and thus reduce some concern on the sampling level, it concurrently increases the risk of reactive bias. In science as everywhere else, one seldom gets something for nothing. The critical question for the researcher must be whether or not the resultant sampling gain offsets the risk of deviation from “true” responses produced by the experting role.

Not only does interviewing result in role selection, but the problem or its analogues may exist for any measure. Thus, in a study utilizing conversation sampling with totally hidden microphones, each social setting elicits a different role selection. Conversation samples might thus differ between two cities, not because of any true differences, but rather because of subtle differences in role elicitation of the differing settings employed.

3. *Measurement as Change Agent.* With all the respondent candor possible, and with complete role representativeness, there can still be an important class of reactive effects—those in which the initial measurement activity introduces real changes in what is being measured. The change may be real enough in these instances, but be invalidly attributed to any of the intervening events, and be invalidly generalized to other settings not involving a pretest. This process has been deliberately demonstrated by Schanck and Goodman (1939) in a classic study involving information-test taking as a disguised persuasive process. Research by Roper (cited by Crespi, 1948) shows that the well-established “preamble effect” (Cantril, 1944) is not merely a technical flaw in determining the response to the question at hand, but that it also creates attitudes which persist and which are measurable on subsequent unbiased questions. Crespi (1948) reports additional research of his own

confirming that even for those who initially say “don't know,” processes leading to opinion development are initiated.

The effect has been long established in the social sciences. In psychology, early research in transfer of training encountered the threat to internal validity called “practice effects”: the exercise provided by the pretest accounted for the gain shown on the posttest. Such research led to the introduction of control groups in studies that had earlier neglected to include them. Similarly, research in intelligence testing showed that dependable gains in test-passing ability could be traced to experience with previous tests even where no knowledge of results had been provided. (See Anastasi, 1958, pp. 190–191, and Cane & Heim, 1950, for reviews of this literature.) Similar gains have been shown in personal “adjustment” scores (Windle, 1954).

While such effects are obviously limited to intrusive measurement methods such as this review seeks to avoid, the possibility of analogous artifacts must be considered. Suppose one were interested in measuring the weight of women in a secretarial pool, and their weights were to be the dependent variable in a study on the effects of a change from an all-female staff to one including men. One might for this purpose put free weight scales in the women's restroom, with an automatic recording device inside. However, the recurrent availability of knowledge of one's own weight in a semisocial situation would probably act as a greater change agent for weight than would any experimental treatment that might be under investigation. A floor-panel treadle would be better, recording weights without providing feedback to the participant, possibly disguised as an automatic door-opener.

4. *Response Sets.* The critical literature on questionnaire methodology has demonstrated the presence of several irrelevant but lawful sources of variance. Most of these are probably applicable to interviews also, although this has been less elaborately demonstrated to date. Cronbach (1946) has summarized this literature, and evidence continues to show its importance (e.g., Chapman & Bock, 1958; Jackson & Messick, 1957).

Respondents will more frequently endorse a statement than disagree with its opposite (Sletto, 1937). This tendency differs widely and consistently among individuals, generating the reliable source of variance known as acquiescence response set. Rorer (1965) has recently entered a dissent from this point of view. He validly notes the evidence indicating that acquiescence or yea-saying is not a totally general personality trait elicitable by items of any content. He fails to note that, even so, the evidence clearly indicates the methodological

problem that direction of wording lawfully enhances the correlation between two measures when shared, and depresses the correlation when running counter to the direction of the correlation of the content (Campbell, 1965b). Another idiosyncrasy, dependably demonstrated over varied multiple-choice content, is the preference for strong statements versus moderate or indecisive ones. Sequences of questions asked in very similar format produce stereotyped responses, such as a tendency to endorse the righthand or the lefthand response, or to alternate in some simple fashion. Furthermore, decreasing attention produces reliable biases from the order of item presentation.

Response biases can occur not only for questionnaires or public opinion polls, but also for archival records such as votes (Bain & Hecock, 1957). Still more esoteric observational or erosion measures face similar problems. Take the example of a traffic study.

Suppose one wanted to obtain a nonreactive measure of the relative attractiveness of paintings in an art museum. He might employ an erosion method such as the relative degree of carpet or floor-tile wear in front of each painting. Or, more elaborately, he might install invisible photoelectric timers and counters. Such an approach must also take into account irrelevant habits which affect traffic flow. There is, for example, a general right-turn bias upon entering a building or room. When this is combined with time deadlines and fatigue (Do people drag their feet more by the time they get to the paintings on the left side of the building?), there probably is a predictably biased response tendency. The design of museums tends to be systematic, and this, too, can bias the measures. The placement of an exit door will consistently bias the traffic flow and thus confound any erosion measure unless it is controlled. (For imaginative and provocative observational studies on museum behavior, see Melton, 1933a, 1933b, 1935, 1936; Melton, Feldman, & Mason, 1936; Robinson, 1928.)

Each of these four types of reactive error can be reduced by employing research measures which do not require the cooperation of the respondent and which are “blind” to him. Although we urge more methodological research to make known the degree of error that may be traced to reactivity, our inclination now is to urge the use of compensating measures which do not contain the reactive risk.

Error from the Investigator

To some degree, error from the investigator was implicit in the reactive error effects. After all, the investigator is an important source of cues to the respondent, and he helps to structure the demand characteristics of the

interview. However, in these previous points, interviewer character was unspecified. Here we deal with effects that vary systematically with interviewer characteristics, and with instrument errors totally independent of respondents.

5. *Interviewer Effects.* It is old news that the characteristics of the interviewer can contribute a substantial amount of variance to a set of findings. Interviewees respond differentially to visible cues provided by the interviewer. Within any single study, this variance can produce a spurious difference. The work of Katz (1942) and Cantril (1944) early demonstrated the differential effect of the race of the interviewer, and that bias has been more recently shown by Athey, Coleman, Reitman, and Tang (1960). Riesman and Ehrlich (1961) reported that the age of the interviewer produced a bias, with the number of "unacceptable" (to the experimenter) answers higher when questions were posed by younger interviewers. Religion of the interviewer is a possible contaminant (Hyman, Cobb, Feldman, Hart, & Stember, 1954; Robinson & Rohde, 1946), as is his social class (Lenski & Leggett, 1960; Riesman, 1956). Benney, Riesman, and Star (1956) showed that one should consider not only main effects, but also interactions. In their study of age and sex variables they report: "Male interviewers obtain fewer responses than female, and fewest of all from males, while female interviewers obtain their highest responses from men, except for young women talking to young men" (p. 143).

The evidence is overwhelming that a substantial number of biases are introduced by the interviewer (see Hyman et al., 1954; Kahn & Cannell, 1957). Some of the major biases, such as race, are easily controllable; other biases, such as the interaction of age and sex, are less easily handled. If we heeded all the known biases, without considering our ignorance of major interactions, there could no longer be a simple survey. The understandable action by most researchers has been to ignore these biases and to assume them away. The biases are lawful and consistent, and all research employing face-to-face interviewing or questionnaire administration is subject to them. Rather than flee by assumptions, the experimenter may use alternative methodologies that let him flee by circumvention.

6. *Change in the Research Instrument.* The measuring (data-gathering) instrument is frequently an interviewer, whose characteristics, we have just shown, may alter responses. In panel studies, or those using the same interviewer at two or more points in time, it is essential to ask: To what degree is the interviewer or experimenter the same research instrument at all points of the research?

Just as a spring scale becomes fatigued with use, reading "heavier" a second time, an interviewer may also measure differently at different times. His skill may increase. He may be better able to establish rapport. He

may have learned necessary vocabulary. He may loaf or become bored. He may have increasingly strong expectations of what a respondent “means” and code differently with practice. Some errors relate to recording accuracy, while others are linked to the nature of the interviewer's interpretation of what transpired. Either way, there is always the risk that the interviewer will be a variable filter over time and experience.

Even when the interviewer becomes more competent, there is potential trouble. Although we usually think of difficulty only when the instrument weakens, a difference in competence between two waves of interviewing, *either increasing or decreasing*, can yield spurious effects. The source of error is not limited to interviewers, and every class of measurement is vulnerable to wavering calibration. Suicides in Prussia jumped 20% between 1882 and 1883. This clearly reflected a change in record-keeping, not a massive increase in depression. Until 1883, the records were kept by the police, but in that year the job was transferred to the civil service (Halbwachs, 1930, cited in Selltiz et al., 1959). Archivists undoubtedly drift in recording standards, with occasional administrative reforms in conscientiousness altering the output of the “instrument” (Kitsuse & Cicourel, 1963).

Where human observers are used, they have fluctuating adaptation levels and response thresholds (Campbell, 1961; Holmes, 1958). Rosenthal, in an impressive series of commentary and research, has focused on errors traceable to the experimenter himself. Of particular interest is his work on the influence of early data returns upon analysis of subsequent data (Rosenthal, Persinger, Vikan-Kline, & Fode, 1963; see also Kintz, Delprato, Mettee, Persons, & Schappe, 1965; Rosenthal, 1963, 1964; Rosenthal & Fode, 1963; Rosenthal & Lawson, 1963).

Varieties of Sampling Error

Historically, social science has examined sampling errors as a problem in the selection of respondents. The person or group has been the critical unit, and our thinking has been focused on a universe of people. Often a sample of time or space can provide a practical substitute for a sample of persons. Novel methods should be examined for their potential in this regard. For example, a study of the viewing of bus advertisements used a time-stratified, random triggering of an automatic camera pointed out a window over the bus ad (Politz Media Studies, 1959). One could similarly take a photographic sample of bus passengers modulated by door entries as counted by a photo cell. A photo could be taken one minute after the entry of every twentieth passenger.

For some methods, such as the erosion methods, total population records are no more costly than partial ones. For some archives, temporal samples or agency samples are possible. For voting records, precincts may be sampled. But for any one method, the possibilities should be examined.

We look at sampling in this section from the point of view of restrictions on reaching people associated with various methods and the stability of populations over time and areas.

7. *Population Restrictions.* In the public-opinion-polling tradition, one conceptualizes a "universe" from which a representative sample is drawn. This model gives little or no formal attention to the fact that only certain universes are possible for any given method. A method-respondent interaction exists—one that gives each method a different set of defining boundaries for its universe. One reason so little attention is given to this fact is that, as methods go, public opinion polling is relatively unrestricted. Yet even here there is definite universe rigidity, with definite restrictions on the size and character of the population able to be sampled.

In the earliest days of polling, people were questioned in public places, probably excluding some 80% of the total population. Shifting to in-home interviewing with quota controls and no callbacks still excluded some 60%—perhaps 5% inaccessible in homes under any conditions, 25% not at home, 25% refusals, and 5% through interviewers' reluctance to approach homes of extreme wealth or poverty and a tendency to avoid fourth-floor walkups.

Under modern probability sampling with callbacks and household designation, perhaps only 15% of the population is excluded: 5% are totally inaccessible in private residences (e.g., those institutionalized, hospitalized, homeless, transient, in the military, mentally incompetent, and so forth); another 10% refuse to answer, are unavailable after three callbacks, or have moved to no known address. A 20% figure was found in the model Elmira study in its first wave (Williams, 1950), although other studies have reported much lower figures. Ross (1963) has written a general statement on the problem of inaccessibility, and Stephan and McCarthy (1958), in their literature survey, show from 3% to 14% of sample populations of residences inaccessible.

Also to be considered in population restriction is the degree to which the accessible universe deviates in important parameters from the excluded population. This bias is probably minimal in probability sampling with adequate callbacks, but great with catch-as-catch-can and quota samples. Much survey research has centered on household behavior, and the great mass of probability approaches employ a prelisted household as

the terminal sampling unit. This frequently requires the enlistment of a household member as a reporter on the behavior of others. Since those who answer doorbells overrepresent the old, the young, and women, this can be a confounding error.

When we come to more demanding verbal techniques, the universe rigidity is much greater. What proportion of the population is available for self-administered questionnaires? Payment for filling out the questionnaire reduces the limitations a bit, but a money reward is selectively attractive—at least at the rates most researchers pay. A considerable proportion of the populace is functionally illiterate for personality and attitude tests developed on college populations.

Not only does task-demandingness create population restrictions, differential volunteering provides similar effects, interacting in a particularly biasing way when knowledge of the nature of the task is involved (Capra & Dittes, 1962). Baumrind (1964) writes of the motivation of volunteers and notes, “The dependent attitude of most subjects toward the experimenter is an artifact of the experimental situation as well as an expression of some subjects' personal need systems at the time they volunteer” (p. 421).

The curious, the exhibitionistic, and the succorant are likely to overpopulate any sample of volunteers. How secure a base can volunteers be with such groups overrepresented and the shy, suspicious, and inhibited underrepresented? The only defensible position is a probability sample of the units to which the findings will be generalized. Even conscripting sophomores may be better than relying on volunteers.

Returning to the rigidity of sampling, what proportion of the total population is available for the studio test audiences used in advertising and television program evaluation? Perhaps 2%. For mailed questionnaires, the population available for addressing might be 95% of the total in the United States, but low-cost, convenient mailing lists probably cover no more than 70% of the families through automobile registration and telephone directories. The exclusion is, again, highly selective. If, however, we consider the volunteering feature, where 10% returns are typical, the effective population is a biased 7% selection of the total. The nature of this selective-return bias, according to a recent study (Vincent, 1964), includes a skewing of the sample in favor of lower-middle-class individuals drawn from unusually stable, “happy” families.

There are more households with television in the United States than there are households with telephones (or baths). In any given city, one is likely to find more than 15% of the households excluded in a telephone subscription list—and most of these are at the bottom of the socioeconomic scale. Among subscribers, as

many as 15% in some areas do not list their number, and an estimate of 5% over all is conservative. Cooper (1964) found an over-all level of 6% deliberately not listed and an additional 12% not in the directory because of recent installations. The unlisted problem can be defeated by a system of random-digit dialing, but this increases the cost at least tenfold and requires a prior study of the distribution of exchanges. Among a sample of known numbers, some 50% of dialings are met with busy signals and “not-at-homes.” Thus, for a survey without callbacks, the accessible population of 80% (listed-phone households) reduces to 40%. If individuals are the unit of analysis, the effective sampling rate, without callbacks, may drop to 20%. Random-digit dialing will help; so, too, will at least three callbacks, but precision can be achieved only at a high price. The telephone is not so cheap a research instrument as it first looks.

Sampling problems of this sort are even more acute for the research methods considered in the present monograph. Although a few have the full population access of public opinion surveys, most have much more restricted populations. Consider, for example, the sampling of natural conversations. What are the proportions of men and women whose conversations are accessible in public places and on public transport? What is the representativeness of social class or role?

8. *Population Stability Over Time.* Just as internal validity is more important than external validity, so, too, is the stability of a population restriction more important than the magnitude of the restriction. Examine conversation sampling on a bus or streetcar. The population represented differs on dry days and snowy days, in winter and spring, as well as by day of the week. These shifts would in many instances provide plausible rival explanations of shifts in topics of conversation. Sampling from a much narrower universe would be preferable if the population were more stable over time, as, say, conversation samples from an employees' restroom in an office building. Comparisons of interview survey results over time periods are more troubled by population instability than is generally realized, because of seasonal layoffs in many fields of employment, plus status-differentiated patterns of summer and winter vacations. An extended discussion of time sampling has been provided by Brookover and Back (1965).
9. *Population Stability Over Areas.* Similarly, research populations available to a given method may vary from region to region, providing a more serious problem than a population restriction common to both. Thus, for a comparison of attitudes between New York and Los Angeles, conversation sampling in buses and commuter trains would tap such different segments of the communities as to be scarcely worth doing. Again, a comparison of employees' washrooms in comparable office buildings would provide a more interpretable comparison. Through the advantage of background data to

check on some dimensions of representativeness, public opinion surveys again have an advantage in this regard.

Any enumeration of sources of invalidity is bound to be incomplete. Some threats are too highly specific to a given setting and method to be generalized, as are some opportunities for ingenious measurement and control. This list contains a long series of general threats that apply to a broad body of research method and content. It does not say that additional problems cannot be found.

An Interlude: The Measurement of Outcroppings

The population restrictions discussed here are apt to seem so severe as to traumatize the researcher and to lead to the abandonment of the method. This is particularly so for one approaching social science with the goal of complete description. Such trauma is, of course, far from our intention. While discussion of these restrictions is a necessary background to their intelligent use and correction, there is need here for a parenthesis forestalling excessive pessimism.

First, it can be noted that a theory predicting a change in civic opinion, due to an event and occurring between two time periods, might be such that this opinion shift could be predicted for many partially overlapping populations. One might predict changes on public opinion polls within that universe, changes in sampled conversation on commuter trains for a much smaller segment, changes in letters mailed to editors and the still more limited letters published by editors, changes in purchase rates of books on relevant subjects by that minute universe, and so on. In such an instance, the occurrence of the predicted shift on any one of these meters is confirmatory and its absence discouraging. If the effect is found on only one measure, it probably reflects more on the method than on the theory (e.g., Burwen & Campbell, 1957; Campbell & Fiske, 1959). A more complicated theory might well predict differential shifts for different meters, and, again, the evidence of each is relevant to the validity of the theory. The joint confirmation between pollings of high-income populations and commuter-train conversations is much more validating than either taken alone, just because of the difference between the methods in irrelevant components.

The “outcropping” model from geology may be used more generally. Any given theory has innumerable implications and makes innumerable predictions which are unaccessible to available measures at any given time.

The testing of the theory can only be done at the available outcroppings, those points where theoretical predictions and available instrumentation meet. Any one such outcropping is equivocal, and all types available should be checked. The more remote or independent such checks, the more confirmatory their agreement.

Within this model, science opportunistically exploits the available points of observation. As long as nature abhorred a vacuum up to 33 feet of water, little research was feasible. When manufacturing skills made it possible to represent the same abhorrence by 76 centimeters of mercury in a glass tube, a whole new outcropping for the checking of theory was made available. The telescope in Galileo's hands, the microscope, the induction coil, the photographic emulsion of silver nitrate, and the cloud chamber all represent partial new outcroppings available for the verification of theory. Even where several of these are relevant to the same theory, their mode of relevance is quite different and short of a complete overlap. Analogously, social science methods with individually restricted and nonidentical universes can provide collectively valuable outcroppings for the testing of theory.

The goal of complete description in science is particularly misleading when it is assumed that raw data provide complete description. Theory is necessarily abstract, for any given event is so complex that its complete description may demand many more theories than are actually brought to bear on it—or than are even known at any given stage of development. But theories are more complete descriptions than obtained data, since they describe processes and entities in their unobserved as well as in their observed states. The scintillation counter notes but a small and nonrepresentative segment of a meson's course. The visual data of an ordinary object are literally superficial. Perceiving an object as solid or vaporous, persistent or transient, involves theory going far beyond the data given. The raw data, observations, field notes, tape recordings, and sound movies of a social event are but transient superficial outcroppings of events and objects much more continuously and completely (even if abstractly) described in the social scientist's theory. Tycho Brahe and Kepler's observations provided Kepler with only small fragments of the orbit of Mars, for a biased and narrow sampling of times of day, days, and years. From these he constructed a complete description through theory. The fragments provided outcroppings sufficiently stubborn to force Kepler to reject his preferred theory. The data were even sufficient to cause the rejection of Newton's later theory had Einstein's better-fitting theory then been available.

So if the restraints on validity sometimes seem demoralizing, they remain so only as long as one set of data, one type of method, is considered separately. Viewed in consort with other methods, matched against the available outcroppings for theory testing, there can be strength in converging weakness.

The Access to Content

Often a choice among methods is delimited by the relative ability of different classes of measurement to penetrate into content areas of research interest. In the simplest instance, this is not so much a question of validity as it is a limitation on the utility of the measure. Each class of research method, be it the questionnaire or hidden observation, has rigidities on the content it can cover. These rigidities can be divided, as were population restrictions, into those linked to an interaction between method and materials, those associated with time, and those with physical area.

10. *Restrictions on Content.* If we adopt the research strategy of combining different classes of measurement, it becomes important to understand what content is and is not feasible or practical for each overlapping approach.

Observational methods can be used to yield an index of Negro-white amicability by computing the degree of "aggregation" or nonrandom clustering among mixed groups of Negroes and whites. This method could also be used to study male-female relations, or army-navy relations in wartime when uniforms are worn on liberty. But these indices of aggregation would be largely unavailable for Catholic-Protestant relations or for Jewish-Christian relations. Door-to-door solicitation of funds for causes relevant to attitudes is obviously plausible, but available for only a limited range of topics. For public opinion surveys, there are perhaps tabooed topics (although research on birth control and venereal disease has shown these to be fewer than might have been expected). More importantly, there are topics on which people are unable to report but which a social scientist can reliably observe.

Examples of this can be seen in the literature on verbal reinforcers in speech and in interviews. (For a review of this literature, see Krasner, 1958, as well as Hildum & Brown, 1956; Matarazzo, 1962a.) A graphic display of opportunistic exploitation of an "outcropping" was displayed recently by Matarazzo, Wiens, Saslow, Dunham, and Voas (1964). They took tapes of the speech of astronauts and ground-communicators for two space flights and studied the duration of the ground-communicator's unit of speech to the astronauts. The data supported their expectations and confirmed findings from the laboratory. We are not sure if an orbital flight should be considered a "natural setting" or not, but certainly the astronaut and his colleagues were not overly sensitive to the duration of individual speech units. The observational method has consistently produced findings on the effect of verbal reinforcers unattainable by direct questioning.

It is obvious that secondary records and physical evidence are high in their content rigidity. The researcher

cannot go out and generate a new set of historical records. He may discover a new set, but he is always restrained by what is available. We cite examples later which demonstrate that this weakness is not so great as is frequently thought, but it would be naive to suggest that it is not present.

11. *Stability of Content Over Time.* The restrictions on content just mentioned are often questions of convenience. The instability of content, however, is a serious concern for validity. Consider conversation sampling again: if one is attending to the amount of comment on race relations, for example, the occurrence of extremely bad weather may so completely dominate all conversations as to cause a meaningless drop in racial comments. This is a typical problem for index-making. In such an instance, one would probably prefer some index such as the proportion of all race comments that were favorable. In specific studies of content variability over time, personnel-evaluation studies have employed time sampling with considerable success. Observation during a random sample of a worker's laboring minutes efficiently does much to describe both the job and the worker (Ghiselli & Brown, 1955; Thorndike, 1949; Whisler & Harper, 1962).

Public opinion surveys have obvious limitations in this regard which have led to the utilization of telephone interviews and built-in-dialing recorders for television and radio audience surveys (Lucas & Britt, 1950, 1963). By what means other than a recorder could one get a reasonable estimate of the number of people who watch *The Late Show?*

12. *Stability of Content Over Area.* Where regional comparisons are being made, cross-sectional stability in the kinds of contents elicited by a given method is desirable.

Take the measurement of interservice rivalry as a research question. As suggested earlier, one could study the degree of mingling among men in uniform, or study the number of barroom fights among men dressed in different uniforms. To have a valid regional comparison, one must assume the same incidence of men wearing uniforms in public places when at liberty. Such an assumption is probably not justified, partly because of past experience in a given area, partly because of proximity to urban centers. If a cluster of military bases are close to a large city, only a selective group wear uniforms off duty, and they are more likely to be the belligerent ones. Another comparison region may have the same level of behavior, but be less visible.

The effect of peace is to reduce the influence of the total level of the observed response, since mufti is more common. But if all the comparisons are made in peacetime, it is not an issue. The problem occurs only if one elected to study the problem by a time-series design which cut across war and peace. To the foot-on-rail researcher, the number of outcroppings may vary because of war, but this is no necessary threat to internal

validity.

Sampling of locations, such as bus routes, waiting rooms, shop windows, and so forth, needs to be developed to expand access to both content and populations. Obviously, different methods present different opportunities and problems in this regard. Among the few studies which have seriously attempted this type of sampling, the problem of enumerating the universe of such locations has proved extremely difficult (James, 1951). Location sampling has, of course, been practiced more systematically with pre-established enumerated units such as blocks, census tracts, and incorporated areas.

Operating Ease and Validity Checks

There are differences among methods which have nothing to do with the interpretation of a single piece of research. These are familiar issues to working researchers, and are important ones for the selection of procedures. Choosing between two different methods which promise to yield equally valid data, the researcher is likely to reject the more time-consuming or costly method. Also, there is an inclination toward those methods which have sufficient flexibility to allow repetition if something unforeseen goes wrong, and which further hold potential for producing internal checks on validity or sampling errors.

13. *Dross Rate.* In any given interview, a part of the conversation is irrelevant to the topic at hand. This proportion is the dross rate. It is greater in open-ended, general, free-response interviewing than it is in structured interviews with fixed-answer categories; by the same token, the latter are potentially the more reactive. But in all such procedures, the great advantage is the interviewer's power to introduce and reintroduce certain topics. This ability allows a greater density of relevant data. At the other extreme is unobserved conversation sampling, which is low-grade ore. If one elected to measure attitudes toward Russia by sampling conversations on public transportation, a major share of experimental effort could be spent in listening to comparisons of hairdressers or discussions of the Yankees' one-time dominance of the American League. For a specific problem, conversation sampling provides low-grade ore. The price one must pay for this ore, in order to get a naturally occurring response, may be too high for the experimenter's resources.
14. *Access to Descriptive Cues.* In evaluating methods, one should consider their potential for generating associated validity checks, as well as the differences in the universes they tap. Looking at alternative measures, what other data can they produce that give descriptive cues on the specific nature of the method's population? Internal evidence from early opinion polls showed their population bias-

es when answers about prior voting and education did not match known election results and census data.

On this criterion, survey research methods have great advantages, for they permit the researcher to build in controls with ease. Observational procedures can check restrictions only for such gross and visible variables as sex, approximate age, and conspicuous ethnicity. Trace methods such as the relative wear of floor tiles offer no such intrinsic possibility. However, it is possible in many instances to introduce interview methods in conjunction with other methods for the purpose of ascertaining population characteristics. Thus, commuter-train passengers, window shoppers, and waiting-room conversationalists can, on a sample of times of day, days of the week, and so on, be interviewed on background data, probably without creating any serious reactive effects for measures taken on other occasions.

15. *Ability to Replicate.* The questionnaire and the interview are particularly good methods because they permit the investigator to replicate his own or someone else's research. There is a tolerance for error when one is producing new data that does not exist when working with old. If a confounding event occurs or materials are spoiled, one can start another survey repeating the procedure. Archives and physical evidence are more restricted, with only a fixed amount of data available. This may be a large amount—allowing split-sample replication—but it may also be a one-shot occurrence that permits only a single analysis. In the latter case, there is no second chance, and the materials may be completely consumed methodologically.

The one-sample problem is not an issue if data are used in a clear-cut test of theory. If the physical evidence or secondary records are an outcropping where the theory can be probed, the inability to produce another equivalent body of information is secondary. The greater latitude of the questionnaire and interview, however, permit the same statement and provide in addition a margin for error.

So long as we maintain, as social scientists, an approach to comparisons that considers compensating error and converging corroboration from individually contaminated outcroppings, there is no cause for concern. It is only when we naively place faith in a single measure that the massive problems of social research vitiate the validity of our comparisons. We have argued strongly in this chapter for a conceptualization of method that demands multiple measurement of the same phenomenon or comparison. Overreliance on questionnaires and interviews is dangerous because it does not give us enough points in conceptual space to triangulate. We are urging the employment of novel, sometimes “oddball” methods to give those points in space. The chapters that follow illustrate some of these methods, their strengths and weaknesses, and their promise for

imaginative research.

- internal validity
- external validity
- interviews
- validity
- sampling
- indexing
- measurement

https://doi.org/10.4135/9781452243443