DOCUMENT RESUME

ED 032 759 EM 007 300

By: Roberts, Karlene H.

Understanding Research: Some Thoughts on Evaluating Completed Educational Projects. An Occasional Paper from ERIC at Stanford.

Stanford Univ., Calif. ERIC Clearinghouse on Educational Media and Technology.
Spons Agency-Office of Education (DHEW), Washington, D.C.
Pub Date Jul 69
Note-32p.
EDRS Price MF-$0.25 HC-$1.70
Descriptors-Administrative Policy, Control Groups, Data Analysis, Decision Making Skills, Educational Policy, *Educational Research, Educational Researchers, Evaluation Criteria, Evaluation Methods, Evaluation Techniques, Experimental Groups, Hypothesis Testing, Models, *Policy Formation, Reliability, *Research Criteria, Research Design, Research Methodology, Research Problems, Research Skills, *Research Utilization, Sampling, Statistical Analysis, *Validity

In order to make policy decisions, educators must evaluate educational research proposals and projects. Findings immediately related to practice are often inadequate, omitting the theoretical establishment of principles by which we can explain and predict the phenomena of our world. These theoretical linkages to the practical world of decision-making occur more frequently as researchers form more innovative hypotheses and design both innovative and rigorous investigative procedures. A research report should present a clear statement of what was studied, the method used in studying it, specification of how the data were analyzed, the results of the study, and conclusions and interpretations. The research evaluator must seek a logical presentation of the problem, since this will partially determine the researcher's ability to develop adequate methodology. The hypothesis must be defined, the sample chosen carefully for stated reasons, the design determined to be valid, and the dependent and independent variables defined and measured. Each study is only a single block in the construction of a theory applicable to educational policy making. No study is without flaws or takes into account all the variables of today's world. Knowledge of research evaluation, therefore, is essential. (MM)

A PAPER FROM

ERIC at Stanford

ERIC Clearinghouse on Educational Media and Technology
at the Institute for Communication Research, Stanford University, Stanford, Calif. 94305

UNDERSTANDING RESEARCH: SOME THOUGHTS ON EVALUATING COMPLETED EDUCATIONAL PROJECTS

By Karlene H. Roberts

Department of Psychology
University of California
Berkeley

An Occasional Paper from ERIC at Stanford

July 1969

FOREWORD

ERIC at Stanford commissioned Dr. Roberts to prepare a "methodological primer" for two kinds of readers:

1. ERIC staff members who index and abstract research reports.
2. Educators who come to us seeking research evidence to guide policy decisions.

Not all research reports are exemplars of sound method or clear presentation. Not all are relevant to any policy decision in the real world. But *some* are, and the proportion of excellent work grows every year.

Since the consumer of research reports shops in a *caveat emptor* marketplace where worthless reports have the same covers, typeface, bulk, tables, and rhetoric as worthwhile reports, we thought there would be some use for a Consumers' Report on educational research. This is what Dr. Roberts has prepared for us.

The project was not as simple as we thought it would be. At first we discussed only a verbal checklist of research quality, derived from the 12 "threats to internal and external validity" listed by Donald Campbell and Julian Stanley in their *Experimental and Quasi-Experimental Designs for Research.* It quickly became clear that research could founder on many reefs beyond these 12. Thus the project grew and the primer grew.

Even at its present length, the primer is selective rather than comprehensive. This must be the case, since a good textbook on method (for example, Kerlinger's *Foundations of Behavioral Research*) is about 20 times as long.

By asking questions about conceptualization, design, execution, analysis, and presentation, Dr. Roberts hopes to forewarn the reader against the most serious and common defects in reported research. When a point requires extended discussion, she refers the reader to methodological source texts that are widely accepted and used.

This is probably not the only version of the primer that we will issue. We hope that readers' reactions will show us which points are inadequately treated, either in length or in clarity.

William Paisley and Don Coombs

TABLE OF CONTENTS

UNDERSTANDING RESEARCH:
SOME THOUGHTS ON EVALUATING COMPLETED EDUCATIONAL PROJECTS*

This paper is intended primarily for persons evaluating educational research proposals or projects for their scientific merit and application to policy formation. It underscores basic points to be noted in assessing a research report and then offers a more extended discussion of these points for the interested reader.

Rigorous research evaluation is crucial because of the urgent need to better understand educational processes and to form effective educational policy. Educational research is emerging from a pseudo-scientific past when it emulated other disciplines considered more exact. The past was characterized by poorly stated research questions forced into ill-fitting research plans with correlation coefficients reported to an inappropriately fastidious third decimal place. Today the trend is toward greater innovation in asking questions and in designing research to explore them. A less pedantic reliance on form, combined with greater understanding of the methods of science, has led to increased regard for research procedures both innovative and rigorous. This experience, if applied, can ensure greater fruitfulness in future educational research.

The Uses and Abuses of Science

You are faced with the necessity of making a decision with some long-term consequence. Research evidence, even if weak, should not be ignored in policy decisions. In our everyday conjectures about what to do, good research offers the best available evidence for understanding relationships among the complex events in an imperfect world. We must, however, clearly understand the limitations of any particular research report, and be able to assess its scientific as well as its practical merit. Most educational researchers and research users look for findings which can be *related to practice.* This approach is valuable but not entirely adequate, because research *related to theory* provides the mortar to tie together our concepts and practical findings.

(Empirical research is taken here to mean findings based on observation rather than on speculation. A concept is an idea representing a number of individual instances all having something in common. A theory is a general principle proposed to explain a group of relationships which are believed to exist in a large body of facts.)

Empirical science has two major objectives: to describe the phenomena of our world, and to establish principles by which we can explain and predict these phenomena (Hempel, 1952). We add to this two-pronged empirical and theoretical approach yet a third objective: the use of research and theory to guide policy decisions. Thus, we have a triangle. Each corner (theory, research, and policy formation) "asks questions" of the other corners. For example, research asks questions of theory, and vice versa. The questions, each appearing as a leg of the triangle, depend on the corner at which one begins, and the direction of travel. Progress is made when we can go from theory to research to policy based on that research. We can profitably begin at any of the three corners.

Research Links to Theory

If the purpose of empirical science is to describe, explain, and predict events, scientists must develop concepts suited to these purposes. Concepts must offer a means for systematizing experience, and

*The author wishes to thank Professors Donald F. Roberts and Eugene J. Webb for their helpful comments.

this is possible only when these ideas can be concretely related to observations. Ideally, all the statements of empirical science should be capable of test by reference to public evidence. We look for links connecting theory to hypotheses to data gathering. (Hypotheses are simply statements about relationships among variables.)

Although a theory-research linkage is ideal, it does not always exist in the behavioral sciences. Theories often contain speculative elements beyond our capacity to observe. And while theory may work as the most efficient means of accounting for data at any one time it is always open to revision. The validity of a theory derives from an accretion process dependent on numerous studies, each approaching a particular theoretical issue from a different perspective. Though we long for data which "prove" theories, it is impossible to prove any theory. Observations only offer support for a theoretical position through a continual process of piling brick upon brick, empirical investigation upon empirical investigation. A single study can simply add or subtract one block to support or discredit a conceptual framework. Even then the block is solid only when the study meets standards for acceptable research. These standards will be examined later.

Before coming to some conclusion regarding the usefulness of a particular theoretical formulation, then, we should review as many relevant research reports as possible. A collection of various research studies is most adequate when each individual piece asks the theoretical question in a different manner. For instance, Chu and Schramm (1967) draw inferences about the impact of instructional television on learning by pulling together the results of many studies which look at the problem in a variety of ways. While each of the studies is characterized by a number of methodological difficulties, the flaws are rarely the same in all of them. We draw realistic conclusions from summarizing the numerous researches in an area, each study with its unique assets and deficits.

To recapitulate, theory and data may not link because a theory is untestable, or the link may not exist because the available test fails to meet standards for adequate observation and deduction.

Research and Practice

In addition to the connection between research and theory, research and policy formation should also be linked. Here we are faced with the difficult problem of asking how pertinent the results of a single investigation are to an upcoming decision. In a world of continuing problems which must be solved, it is often impossible to await the long and arduous process of building and verifying theory.

We take what is available and recognize that we must act on the basis of imperfect evidence. For example, the conditions of one research finding may no longer apply at the time of our interest, and in accepting research results we should not ignore the historical and political context in which the study was conducted and in which the policy is to be formulated. Nor should we fail to consider the series of interlocking problems which must be faced in making any decision. Critical among these problems is the question of what exactly it is possible to change at any single point in time.

The social consequences of any influence depend on its strategic importance in a particular subgroup and whether or not it can be changed. We might, for instance, be interested in reducing delinquency in a particular geographical area. We find the available research clearly indicates a link between poverty and deliquency in that area, but knowing this is of little consequence unless we can reduce the poverty level. Such knowledge may be essential for understanding but unusable for short-term practice.

Criticism of the Scientific Approach

Quantitative research is most frequently criticized on two grounds: (1) statistics can be mustered to support any position, and (2) statistical analyses may lead to sterile conclusions which ignore the complexities of everyday life.

Figures can undoubtedly be manipulated to support a favored position, but in his explorations the qualified researcher should be more interested in truth than favor. By understanding the research process, the research user can objectively assess both a research problem and design. He can then decide for himself the appropriateness of criticism in any specific instance.

While it is true that conclusions based on empirical investigations may be sterile, outdated, and unrepresentative of the "real world," this too is a function of the researcher–of his ingenuity. We can assess the questions he selects and his approaches to them in order to determine the importance and applicability of his results. Non-empiricists argue for conclusions based largely on intuition. Yet, intuition is itself grounded in observation and experience, the synthesis of a chain of past events somehow combined in an individual's memory. Mere intuition, though, does not provide as powerful a basis for understanding phenomena or making decisions as does empirical investigation.

Non-empirical statements are apt to be little more than conjecture. For example, it may seem to me that "more third graders than ever before are having difficulty learning to read." Empiricists usually want to tie this sort of statement down. They would like to indicate, with reasonable accuracy, what percentage of children from what geographical areas are having trouble learning to read, and why, and how things have changed over time.

Understanding Research

Before we can assess the usefulness of a particular piece of research we must be able to understand what is reported. Conventional methods of reporting have been developed for empirical investigations. If they are followed by the researcher and understood by the consumer, any empirical effort can be assessed in terms of its theoretical or policy value. A research report should present, in this order:

1. A clear statement of what was studied
2. The method used in studying it
3. Specification of how the data were analyzed
4. The results of the study
5. Conclusions and interpretations.

There are numerous examples in educational research of clearly presented empirical studies, and there are also specimens of wordy, difficult, mind-boggling presentations. A research report should not read like a mystery story, leaving until the final page revelation of what it is all about. It should rather read like a cookbook. The dish to be concocted is defined at the beginning. A list of ingredients is followed by the method for combining them, and then the outcome. The direction "add one half cup of butter" is significantly clearer than "add enough butter." Writing clarity is important and should not be overlooked in evaluating a study. If the researcher is unable to clearly state his work, he may not clearly understand it. (Relevant points are offered in the *Publication Manual* of the American Psychological Association (1967), and by Kerlinger (1964).

The Research Report

Hindsight being what it is, it is relatively easy to admonish some other investigator for his pedestrian approach, oversights, and myopic vision. It is decidedly more difficult to achieve ingenuity, insight, and sensitivity oneself. A research report should be approached with objectivity and vigilence, but also with sympathy.

Our purpose will, in part, determine our evaluation of the quality of a report. Certain conceptual and methodological points should not be omitted from any research, but the importance given to any one point may depend on the purpose of the review. Investigations intended to further scientific understanding are not always those best suited to help make a practical decision, but as we become increasingly comfortable with the methods of science it will be easy to determine the kinds of studies applicable to our particular needs.

A researcher's attention to trivia does not compensate for his failing to consider important issues. We should beware of investigators who carefully report percentages to the third decimal place when the figures derive from a poorly designed investigation. How to assess the quality of an investigation will become clearer as we continue.

Producing a piece of research is like running a maze. In both cases, a decision made at any one point hinges on the decision immediately preceding it and will affect those following it. Any decision, at any point, can result in a dead end, or in a few more steps in the direction of problem solution. The investigator and the research reviewer must consider many of the same decision points. The major decision points of research are indicated in Appendix 1. Both the researcher and reviewer should also answer the series of linked questions presented in Appendix 2. If researchers more frequently engaged in this drill we would have fewer examples of expensive and time-consuming research projects with meaningless results. The same exercise is valuable to the reviewer in assessing the quality of a piece of research before accepting or rejecting the conclusions based on it. And because he uses the results of research, the reviewer may be in a position to effectively pressure for better research. If he understands the difference between good and bad research he can point out useless studies and ask that resources be committed to better work.

The following sections review the basic questions we should ask when reading a research report. In each section these questions are followed by a further explanation of the concepts and terms involved. Some readers may find it useful to read these extended explanations, while others will want to skip them.

PROBLEM PRESENTATION

"Scientific inquiry is an undertaking geared to the solution of problems. The first step in the formulation of research is to make the problem concrete and explicit" (from Selltiz *et al.*, p. 31).

"The ability to perceive in some brute experience the occasion for a problem and especially a problem whose solution has a bearing on the solution of other problems, is not a common talent among men. . . ." (cited by Selltiz *et al* from Cohen and Nagel, 1934).

We begin our assessment of a project be asking such questions as these:

1. Does the author present his area of interest in a way that is understandable to me?
2. Is the question he asks an important one to me?
   a. Does the research merely look at some small and relatively unimportant aspects of a problem, or does the research ask an important question in an appropriate manner?
   b. Will the answer to the problem help further scientific understanding and/or will it provide guidelines to decision-making?
3. Is the general research area logically drawn from a reasonable body of thought--from past work or from the author's own thinking?
   a. Is this a relatively virgin area, so that we expect the author to have some difficulty formulating his problem? (In exploratory studies* the discovery of ideas and relationships is emphasized.)
   b. Is it a well-studied path, so that problems along it can and should be fairly well explicated?
   c. Does the author seem to know what was done before, so that he is drawing upon, but not re-tracing the steps of others?
   d. If the author does not draw on previously developed theory, does he tell you this or does he merely ignore any mention of the conceptual framework in which his problem lies?
4. Is the specific area of research well explained, given its limitations?
   a. Does the author draw a manageable problem from his over-all area of interest?
   b. Is the problem presented concretely?
   c. Does the author carefully explain the terms he uses?
5. From this more or less concrete problem is the researcher able to designate concepts which logically follow? Does he specify how to measure variables? Are they relevant to the over-all concepts and purposes of the research?
6. Does the researcher frame reasonable, testable, propositions in light of his concepts and variables?

Whether or not the author presents his problem logically will partially determine his ability to develop adequate methodology. He may earn plus points in one area only to slip up in the next, and oscillation between conceptually adequate and operationally feasible inquiry is characteristic of all empirical investigation. Given knowledge of a completed study, we might ask ourselves how we would have approached the problem, conceptualized it, narrowed it, developed propositions from it, and operationalized (or specified how to measure) the terms. This drill helps us develop the skill of recognizing adequate research. If we can, we might indicate to ourselves what appear to be appropriate and workable methodological approaches. We then may want to see how we disagree with the author. At any rate, we should compare all studies under review on questions like those above. For readers who are still in doubt as to whether a particular research problem has been adequately presented, the following remarks may be helpful. Other readers should skip to the method section.

*Terms such as this are discussed in the extended sections, which can be identified by wider margins so that readers may skip them if they desire.

Problem presentation--the logical sequence

Whitehead (1911) points out that in creative thought common sense is a bad master. He states that the single common sense criterion for judgment is that new ideas shall look like the old. Science and common sense differ in a number of respects. Most important, the scientist uses conceptual schemes differently than "the man in the street." He tests his theories using controls, and in this way seeks to rule out metaphysical explanations of phenomena as well as hypotheses rivaling those he tests.

The general problem statement in a piece of research is usually quite broad. It sets the stage for what is to come and allows the author to zero in as he proceeds. The problem may be based on a theoretical framework, derived from hunch, past experience, or some body of thought not itself related to theory.

Theory. "A theory is a set of interrelated constructs (concepts), definitions, and propositions that present a systematic view of the phenomena by specifying relations among variables, with the purpose of explaining and predicting phenomena" (Kerlinger, 1964, p. 11). Problems based on theory are more likely than others to contribute to the advancement of scientific knowledge, because their exploration fits into some broad scheme of things. Problems not so derived are often explored with limited but sometimes more readily applicable results. The results of studies not linked to theory, then, are more often tied to the situation at hand. This difference is important to think about in evaluating a report, for the purpose of the researcher will partially determine his study's value to us.

Problem formation. It may take a researcher years to even formulate a problem well, but one can generally discern from his publication the degree of success he has had in defining concepts, constructs, and variables.

(We already know that a concept is an idea representing or summarizing a number of individual instances. A variable is a symbol to which values are assigned. A construct is a property ascribed to at least two objects as a result of scientific observation. It is a concept more formally stated, complete with definitions and limitations which are explicitly related to a body of empirical data.)

Operationalism. An operational definition assigns meaning to a construct or variable by specifying the activities necessary to measure it.

Underwood (1952, p. 53) states that operational thinking produces better scientists because it forces them to remove fuzz from their minds. It facilitates communication among scientists because the meanings assigned to concepts are less easily misunderstood. The adequacy of any particular operational definition is difficult to assess. No clear cut rules exist, and we can only accept what seems consistent and logical.

Operationalism should not be emphasized to the neglect of importance. Explicit operationalism may be mandatory in the examination of certain theoretical questions, and should be demanded when this is the case. However, such precision sometimes detracts from the use of research results in practical settings. For instance, one operational definition of anxiety is an individual's score on a specific test, say the Taylor Manifest Anxiety Scale. Other less precise but more easily obtained indicators may be better for assessing anxiety in the classroom (e.g., percent of time student is observed fidgeting during a one hour period). We must be careful in deciding how stringent a researcher should have been in his operationalisms, but recognize too that there are numerous studies in which better operational definitions would have been helpful.

Many investigations, for example, compare the effects of television versus conventional instruction on learning, and fail to adequately define what is mean by "television instruction," "conventional instruction," or "learning."

Hypotheses. Hypotheses should be logically derived from the problem statement. A hypothesis may assert that something is the case in a given instance, it may have to do with the frequency of occurrence of association among variables, or it may

claim that some characteristic determines another. In short, a hypothesis is an "if . . . then" statement relating two concepts. A study's hypotheses act (for the reviewer) as guides to the kind of data which should have been collected and imply the manner in which the relationships among variables might be tested.

In summary, the introductory section of a report previews what is to come:

> The scientist needs viable and plastic form with which to express his scientific aims. Without content—without good theory, good hypotheses, good problems—the design of research is empty. But without form without structure adequately conceived and created for the research purpose, little of value can be accomplished. Indeed, it is no exaggeration to say that many of the failures of behavioral research have been failures of disciplined and imaginative form (Kerlinger, 1964, p. 290).

METHOD

The propositions or hypotheses stated in the initial section of the research report determine the approach to design. In behavioral science research the null hypothesis traditionally is the one tested. It represents a position of skepticism which says that regardless of how I predict two groups may differ along some dimension, I test that difference using a statistical test that assumes they are the same. The null hypothesis is *not* the stated hypothesis of interest to the researcher. This matter is further discussed in the extended section on method. Investigators should state their hypotheses in terms of hoped for or expected differences, and if possible they should state the direction of the expected difference. A researcher might hypothesize:

Hy 1 – Mathematical learning ability is differentially related to intelligence in children.

better yet:

Hy 2 – Children with high intelligence test scores will learn mathematical principles more quickly than will children with low intelligence test scores.

but more refined still:

Hy 3 – Ten year old children scoring over 110 on the WISC will receive better grades at the end of the first six weeks of introductory mathematics than will ten year old children scoring below 110 on the WISC

We see that while more refined hypotheses suggest more refined measures, they may reduce the generalizability of results. Remembering that the nature of the research influences the degree of specificity with which hypotheses can be stated, we can proceed with caution, and again ask a series of linked questions:

1. Does the author state his hypotheses in terms of expected differences? If they were possible to state did he include the directions of the expected differences?
2. Does the author say enough about his research so that it can be repeated by someone else?
   a. Is the design orderly and understandable?
   b. Can it be blocked out? That is, can you draw a chart showing, in a time sequence, what the author did?
3. Does the author state the population he is interested in, and does it seem a relevant population in terms of the nature of the problem and the hypotheses?
4. How was the sample drawn from the population?
   a. Was it randomly selected? (Explanations of terms such as this are included in the following extended section on methodology.)
   b. If not randomly selected, was the sampling method the best that can be expected under the circumstances?
   c. Did the author avoid the regression effect?
   d. If groups were matched, was matching done within randomized groups, and were groups matched on all possible relevant variables?
   e. Is the total sample size adequate in terms of the size of the smallest subgroup the author subsequently identifies? That is, are there enough cases in his smallest subgroup to supply adequate data for analysis?
   f. If the research involved a survey, did it draw on respondents who had the knowledge to

answer the questions asked?

5. Does the design meet the criteria for good design, considering the limitations imposed by the nature of the author's problem? (Some acceptable designs are detailed in the following section.)
   a. Does the experiment work? Or can apparent differences in results among groups tested be explained by some measurement artifact rather than the variable of interest to the researcher?
   b. Are the results of the study generalizable, at least to groups of interest to the author?
   c. Is the matter of causation considered, or is the investigator merely interested in showing an association among variables?
   d. Does the author consider the conditions under which relationships among his variables hold (specification), and those under which relationships disappear?
6. Are measures of the independent variables clearly specified? Are these measures appropriate to the problem at hand?
   a. Are experimental manipulations carried out in a logical, unbiased manner?
   b. Do questions used in interviews or questionnaires meet the criteria for question writing? (Again, see the foliowing section for some of these criteria.)
   c. What reliability and validity data are offered for any of the tests used?
   d. Can the independent variable measures be improved?
7. Are the dependent variable measures satisfactory?
   a. Are the observations independent of one another?
   b. Do questionnaires, interviews, and tests used meet acceptable standards? (The extended section may be helpful.)
   c. Could the dependent variable measures be improved?
8. Are there multiple indicators of both independent and dependent variables? (This would be a desirable arrangement.)
9. Were appropriate ancillary data collected?
   a. Are demographic variables included which will help specify the conditions of a relationship?
   b. Are normative data included for comparative purposes?
   c. Are unnecessary data included?

The implication of the answers to any of these questions will depend on the nature and purposes of the study. No single study will score perfectly in all areas. We may have to accept trade-offs. For example, whether we accept the results of a study with less than rigorous measurement of a carefully chosen sample from a known population, or a study using tightly specified measures but perhaps employing a less well-defined sample, will depend on our purposes. If we must act tomorrow to alter the school system in "Typical Town," it is better to give greater weight to research on a sample of carefully selected typical town children than be overly concerned with the reliability of the author's measures. If, on the other hand, we are selecting a severely retarded child for lifetime institutionalization, measurement reliability and error are of prime importance.

The nature of the problem also determines the appropriateness of a particular approach to data collection. Field investigations have different strengths and weaknesses than laboratory experiments· they are generally used for different purposes and necessitate different kinds of observation. A field study is simply an investigation carried on outside the laboratory where satisfactory controls over data are more difficult to achieve. In laboratory experiments the effects on results of the numerous possible extraneous variables are minimized. The virtue of the laboratory experiment rests in the possibility of complete control and replicability; its vice is the possibility of irrelevance to the real world.

Finally, we should consider the translation of observation to measurement and categorization:

10. Do the observational categories for each variable seem relevant in terms of the author's operational definitions of his independent variables?

a. For any *one* variable, were the subcategories exhaustive? That is, were they extensive enough to include the coding of all subjects?
b. For any *one* variable, were the subcategories mutually exclusive? (That is, a subject should fall into one and only one category on a variable.)
c. Were the categories for *different* variables by nature independent of one another? A single subject coded independently for both intelligence level and eye color should not necessarily fall into one cell because the nature of the world is that all low intelligence people have brown hair. If he is intelligent, it should be possible from the nature of things for him to be placed in the cell for brown, blonde or any other color hair. That is, each variable should be capable of measurement which is independent of the measurement of other variables.
d. Were all categories for any one variable on the same level of discourse?

After examining a report's general research approach, the experienced reader can query himself as to what he would have done about analyzing the data. A useful exercise is to question the value of both an idealized approach and the one actually used by the researcher. Before considering the problems of data analysis, some readers may want to brush up on methodological points. The next section is for their benefit.

Method–Traversing the Maze

"Methodology . . . is not everlasting truth. It is a living body of ideas that changes with time" (Hirschi and Selvin, 1967, p. 6). New methodological approaches are continually developed, new measurements and critera considered. It is not surprising, then, that any two investigations may use different approaches, each with unique strengths and limitations. We attempt here to indicate current major methodological concerns, so that we are better prepared to assess a research product. All of these concerns will not be addressed in any single piece of research. However, the more points a researcher considers, the better his research.

The null hypothesis. A hypothesis is stated such that the author may test whether any difference he observes could come about as a result of chance. We state hypotheses in terms of some difference expected between two or more groups exposed to different conditions, while the chance expectation usually is one of no difference between conditions. This is the null hypothesis, or the hypothesis of no difference. The proposed hypothesis is tested against that of chance expectation by a statistical analysis. We can see now why research propositions should be stated in terms of expected differences. If the empirical data for the two or more groups fit the chance model, then they are not significant.

The researcher who can specify the direction of his prediction has the advantage. Looking back to our example of hypothesis formation in the last section, the researcher who could hypothesize the relationship of high intelligence test scores and mathematical proficiency was better off than the one who could only specify that mathematical ability is somehow related to intelligence. Reference to a statistics textbook will help us understand why hypotheses stating the direction of the expected difference can be more efficiently tested than those which do not.

Sampling-selection. Researchers usually want to generalize their findings beyond the group of people they originally study. If this original group is relevant in terms of the problems examined in the research, we want to know how the investigator selected that portion of the population to be measured. It would be ideal (if inefficient) to measure the entire population of interest, and if the total population is small or if there is some doubt about the ability to sample it appropriately, this is what the researcher should have done. Most researchers, of course, cannot study complete populations and are forced to select

some part as representative of the whole. If at all possible, sampling should be random (Anderson, 1968; Blalock, 1964, p. 24; Hays, 1963). **Randomization** (or **probability sampling**) ensures an unbiased sample because every member of the population has known probability of selection. Randomization of the total sample helps to ensure the equivalence of all groups studied, and thereby reduces the possible sources of unknown influences producing results.

Kerlinger states (1964, p. 60) that "when working with samples that have not been selected at random, generalization to the characteristics or relations between characteristics in the population is, strictly speaking, not possible." A compelling argument for random selection is offered by Fisher:

> . . . The uncontrolled courses which may influence the research (of an experiment) are always strictly innumerable. When any such course is named, it is usually perceived that, by increased labour and expense it could be largely eliminated. Too frequently it is assumed that such refinements constitute improvements to the experiments . . . this equalization must always be to a greater or less extent incomplete, and in many important practical cases will certainly be grossly defective . . . the simple precaution of randomization will suffice to guarantee the validity of the test of significance, by which the result of the experiment is to be judged. (1966, p. 18).

This substantiates the notion that each sample element should have a known probability of selection. It is not sufficient, although often necessary in educational research, to randomly select classes or other existing groups. Groups, particularly classroom groups, possibly differ from one another in their initial assignment, and random selection of these groups is not always adequate, particularly when only a few groups are used in a research project. If the results of the research are to be applied to these groups only, this kind of selection may be proper. Generalizability of results is given up in the interest of specific application.

If the researcher has used one of the probability sampling techniques, he should be credited. Where **simple random sampling** is not possible or generalizability of results is not wanted, other procedures can be employed. Each of these has its own limitations.

Another form of **probability sampling** is **stratified random sampling**. Factors known to influence the behavior under observation identify the strata. The proportion of subjects randomly selected from each segment may be determined by its occurrence in the population of interest. Stratified random sampling usually is an improvement on simple random sampling only if the researcher is intelligent about determing relevant strata. Too often sex and socio-economic level are selected with little thought given as to why, and even less consideration to the relevance of other variables. **Cluster sampling** is yet another form of random sampling. Two or more sampling stages are involved, with a first cut which isolates specific clusters, and then a second stage in which elements within each cluster are randomly selected.

Nonprobability sampling is used for convenience, economy, or when probability sampling is inappropiate. As a class, non-probability sampling is less adequate than the methods already mentioned. **Accidental sampling** conscripts subjects because they happen to be available. The method suffers from the inability to describe the population from which the sample is drawn, and generalization of results is impossible. **Quota sampling**, like random stratified sampling, guarantees the inclusion of diverse population elements. Enough cases should be included from each stratum to represent the population, but it is not necessary that the sample strata be proportional to the population strata. A particular stratum is sought until some predetermined number of cases is found. This kind of sampling is popular in opinion polls. Finally, in **purposive**

sampling, cases are handpicked on the basis of some attribute. For instance, it would not be proper to randomly select from the population of all juveniles for a study of delinquency recidivism, for such sampling might not include a large enough number of delinquents for the study. Purposive sampling may be the only alternative. For a more complete discussion of this highly complex problem of sampling see Chein (1966, p. 509ff).

We must combine considerations of appropriateness and economy in assessing the sample used in any research project. Was the sampling method logical in terms of the research question, and if a better procedure exists was it economically feasible? Although the experimenter may have indicated how he selected his entire sample, he may then fail to discuss assignment of members of the sample to subgroups. No adequate interpretation of measurement results can be made without comparison groups, and in studies where we do not know how subjects were assigned to them, the results are suspect.

The merit of assigning subjects randomly to all groups—as well as selecting the total sample at random—should be obvious, but where this is not possible other sampling methods can be called upon. Self-selection is the circumstance of assigning oneself to participate in an experiment. It is not acceptable, for it creates conditions under which it is too easy for unwanted variables to effect experimental results. Matching subjects across groups without first selecting the groups randomly is similarly disadvantageous.

Randomization might be supplemented by matching individuals across groups case by case **(precision control)** or by matching experimental and control groups in terms of overall distribution of factors **(frequency distribution control)**. Matching is, at best, a questionable operation. As stated by Fisher (1966, p.18), one is never certain all relevant variables are considered and the process is tedious and often expensive.

Regression effect. Measuring a group of subjects on some dimension, selecting the extreme groups on this dimension for experimentation, and remeasuring on the same dimension is common practice. Taking children from a particular grade level, and selecting those who score at the extremes on an intelligence test is an example of this. A little thought indicates why the procedure is troublesome. Even without the intervention of an experimental treatment, how would the entire group redistribute itself on a second testing? Recognizing that no test score is a perfect measure, there will be some change in individual scores due to measurement error from the first to the second testing. Students scoring at the extremes in test one have no place to go in test two but back toward the middle (or to the average of the entire group). For example, this month a student takes a vocabulary test with a maximum score of 118. He scores 118. Next month the student repeats the test. He cannot score over 118 and is likely to score under the maximum simply because any test is somewhat unreliable. This is known as regression to the mean because the scores for individuals in any group who fall at the extremes on an initial testing of a dimension are apt to regress to the mean of the entire group on a second testing of that dimension. A researcher may conclude that the treatment given two groups selected this way had no effect, when in fact it did. The use of extreme groups as subgroups in "test-retest" studies is poor practice.

Sample size. For any research we use as large a sample as we can obtain, or as seems reasonable. The problem of ascertaining appropriate sample size has to do with the relationship of size to measurement error. The smaller the sample, the larger the probability of error in results. This is particularly important when the population from which the sample is drawn is small. For example, suppose we are interested in the political attitudes of living ex-presidents and vice-presidents of the United States (population). We go to one of the men (sample) and ask him if he has become more conservative since leaving office. He answers that he has, and we assume that all ex-holders of executive office in the United States become more conservative after leaving office. What we do not know is that this one person has been exposed to special circumstances contributing to conservative attitudes, and that none of the rest of the population has changed since leaving office.

When populations are large the researcher need not have selected commensurately large samples. The size of the smallest subgroup, the necessity for precision in estimating population characteristics, and available budget determines the size of the sample selected (see Hays, 1963, p. 204). Some writers arbitrarily select a sample size, like 10, which the smallest subgroup should equal or exceed. There is no known reason for this particular size except that most statistical analyses are inappropriate for small cell entries. Generally, the larger the sample the safer the assumption that all groups within it are equivalent on all possible factors which will influence results. This is one control on plausible rival hypotheses other than the one(s) tested. Samples should be large enough to allow the principle of randomization to work.

Design validity. Campbell and Stanley (1963, p. 5) itemize twelve factors which jeopardize the validity of a research design. These factors pose threats to internal or external validity. **Internal validity** is concerned with whether or not the experimental manipulations work. It is a necessity; without internal validity experimental results are uninterpretable. Eight classes of extraneous variables are involved here, and if not controlled in the research design, they may produce consequences which confound the experimental results. These threats are often difficult for us to comprehend, but they should be considered in evaluating a project. Other events occurring (**history**) and changes in the subjects (**maturation**) between a first and a second measurement are two extraneous variables which can influence results. **Testing** the same subject at two different times with the same test, changes in calibration of the measurement instrument between two testings of the same individual (**instrumentation**), **statistical regression** (previously discussed) occurring as a result of comparing extreme groups, and biases resulting from **differential selction** (discussed under sampling) of comparison groups are also hazards to objective empirical results. Differential loss of respondents over time (**experimental mortality**) and the reciprocal influence of **selection and maturation** upon one another are the final threats to internal validity. If you find these factors difficult to grasp, Campbell and Stanley (1963) offer a more complete discussion of them.

External validity is concerned with the generalizability of results beyond the sample measured. **Measurement reactivity** or the **reactive effects of experimental manipulation** are factors here: the respondent knows he is being tested and reacts accordingly (Campbell, 1957; Orne, 1962; Selltiz **et al.**, 1959, Webb **et al.**, 1966). **Selection biases** (previously discussed) and the **experimental variable** itself may interact to produce uninterpretable results. Finally, in a design using **multiple treatments** (on several manipulations) there may be **confounding** among them because the effects of a previous treatment are not erasable. This interaction of the treatments and their combined effect on results may render the results difficult to interpret.

The sources of invalidity are nothing more than rival hypotheses to the hypothesis of chance results or to the hypothesis that the experimental variable has had an effect. Researchers typically try to control sources of invalidity, and our evaluations of projects should ask how well they succeeded.

Good and bad designs. Research designs have two fundamental purposes: to provide answers to research questions, and to control unwanted variance. (Variance refers to the spread of scores among individuals in a group.) A research design is the skeleton which tells us how observations are to be made and analyzed, and a good research design offers an adequate test of the relationship among variables.

Campbell and Stanley (1963) and Kerlinger (1964) present complete reviews of the limitations of various designs. We abstract here only a few of those most frequently seen. Knowing their assets and limitations should help to assess research using them.

Three often used procedures fail to meet any of the requirements for an experimental design. They do not meet the criterion of randomization and they fail to control for internal and external validity factors. The designs are:

1. The one-shot case study.
2. Observation of the group, application of some treatment, and reobservation of the same group.
3. The comparison of a group which has experienced some event with one which has not.

In all three cases, there is no opportunity to exert experimental control. An example of the first instance (the case study method) is a case history of a mother's feelings of inadequacy about coping with her difficult child. Given a complete history one might conclude that her feelings of inadequacy are based on the mother's own relationship with her parents, when, in fact, they are the result of a multiplicity of causes. We could test the mother before and after her exposure to her perplexing child and note the differences in her behavior (the second example). Little help is afforded here, for her behavior may have changed with or without the introduction of the child. If we compare a mother of a more docile child to our now haggard mother (the third example), and find differences in their behavior, we still cannot be certain the two mothers were alike prior to the advent of their children.

These designs should be negatively evaluated if used in research which has as its stated purpose extending scientific knowledge. Any one of them might be useful, however, in suggesting hypotheses or in helping us decide on policy matters in subject areas where little is known. They might have offered the only approach possible under a given set of circumstances.

The classic experimental design in the behavioral sciences is illustrated in Figure 1, where R indicates random selection of subjects, O indicates observation, and X indicates experimental treatment. (This notation from Campbell and Stanley, 1963.) This design includes an experimental and control group. A control group is a group equivalent to the experimental group (because of random selection) and exposed to all the conditions of the investigation except the experimental variable. The design controls for internal validity factors, but does not fare well in controlling external validity factors.

$$\begin{array}{cccc} R & O_1 & X & O_2 \\ R & O_3 & & O_4 \end{array}$$

Figure 1

True experimental design

The experimental effect may be limited to groups warmed up by the pre-test. The effect of awareness of being tested is unknown, and so is the effect of the interaction of selection and treatment. The design is often impossible to use because some variables of interest simply cannot be studied under experimental conditions. In assessing a report we can ask if the researcher could have randomized his subject selection, applied a treatment (manipulated the independent variable) and made observations (measured the dependent variable) in the manner required here. If the answer is "yes," he probably should have used this design or one like it. The design should be given a high rating but only if it is used where it is relevant to the question at hand.

We can illustrate the use of the "Figure 1" design by imagining how we might compare two methods for teaching reading. Students are randomly assigned to experimental and control groups with reading ability measured in both groups by a standardized test. A new instructional technique is introduced to one group, and the

other group continues to learn to read as before. Reading ability is again measured in each group, using the test previously used.

A more expensive and less often used method is an extension of the aforementioned design called the Solomon four group design (see Fig. 2). The previous model is extended by the addition of another experimental and control group, neither of which receives the pretest.

R	O_1	X	O_2
R	O_3		O_4
R		X	O_5
R			O_6

Figure 2

Solomon 4-group design

Threats to internal validity are controlled, and the external validity threats of testing and of the confounding of testing and treatment are determinable. This design can be varied further by adding additional experimental and control groups.

Although these two designs offer the greatest degree of control over experimental situations, they are less often found in educational research than are quasi-experimental designs. Quasi-experimental designs offer the advantage of introducing into natural settings something like experimental design in data collection, even though full control over the experimental setting is lacking. For a complete review of the quasi-experimental designs see Campbell and Stanley (1963, pp 39 ff). Two of these designs are worthy of mention here:

The times series design is indicated in Figure 3. It is characterized by periodic measurement of some group or individual, and change in an independent variable at some point. This design is appropriate where random selection is not possible, or where there is no control over the introduction of the experimental variable.

O_1 O_2 O_3 X O_4 O_5

Figure 3

Time series design

The investigator makes a number of observations and infers that the experimental variable has an effect if immediately after its introduction the dependent variable changes. On the internal validity side, the design fails to control for the effects of history, and the instrumentation effects are unknown. Threats to external validity are unknown and the effects of the experimental variable may be limited to the sample at hand. The design is best in situations where the researcher is interested in the effects of a change which he expects at some future time—a change only applicable to a single group. An example of such a situation is the introduction of a new administrative system in a school district,

and its effects on teacher behavior.

A second quasi-experimental design is the nonequivalent control group design (Figure 4), common in educational research and often confused with a true experimental design (Figure 1). The difference between the two is that in the nonequivalent control group design groups are not randomly selected, and possibly not equivalent.

O_1 X O_2

O_3 O_4

Figure 4

Nonequivalent control
group design

Classrooms may function as the groups in this design, and the approach is problematic. First, selection and maturation can interact. An example of this occurs when one group has a faster maturation rate than the other. Second, if either group is selected on the basis of extreme scores on some variable, regression may be a problem. All forms of threats to external validity are present. While quasi-experimental designs are appropriate for studying problems where random selection is not possible, every attempt should be made by the researcher to use groups relevant to the purpose of the experiment. A research question using this design might be concerned with the effect of a new instructional method on the learning of arithmetic by ghetto children. Two neighboring ghetto classes are selected as the experimental and control groups. They are both tested on arithmetic ability. One class is subjected to the new instructional method and both are again measured on ability. The results of this kind of study are not generalizable to other groups unless students are randomly selected and then assigned to classes. The design then becomes a true experimental design.

The true experimental designs (Figures 1 and 2) are most often used in research which seeks to further our scientific understanding of a particular phenomenon, but they are so well controlled that they are often not as useful in policy determination as are less rigorous studies. We may actually need information about the most adequate method for teaching arithmetic to a particular group of ghetto children, and generalizing to other groups is unessential.

The virtue of a true experimental design is that independent variable variance is minimized. (The independent variable is the one manipulated.) Experiments using such a design are optimal for testing precise hypotheses and they are generally replicable. However, they are artificial, manipulation effects often are weak, and data analysis often demands the application of the most powerful statistical techniques. Their strength lies in overcoming threats to internal validity; their weakness lies in lack of control of threats to external validity. Field experiments are usually characterized by less control than are laboratory investigations. They present problems in the manipulation of independent variables, and there is the possibility of not being able to sample randomly. They are flexible and best suited to testing broad hypotheses. Sometimes they can follow true design principles, but more often they are quasi-experimental in nature or are *ex post facto* inquiries. In this latter case variables are not manipulated. Whether the investigator mentions it or not, his design considerations must go hand in hand with his thoughts about how to analyze his data. He should be evaluated on the appropriateness of his design to his problem, and the appropriateness of his data analysis to his design.

Independent and dependent variables. Variables are divided into at least two

classes, and researchers examine how changes in one class (independent variables) lead to changes in the other class (dependent variables.) In other words, independent variables are manipulated and their effects on dependent variables observed. A study's hypotheses teli us which variables are to be considered independent and which dependent. The goal of research is to show either association or a causative tie between independent and dependent variables. The appropriateness of the independent and dependent variable indicators in a particular study depends on the research problem and the hypotheses. We must consider these simultaneously, for research swings constantly back and forth between them.

> Despite the great importance of the procedure for moving from concepts to indicators and from indicators to concepts, precise rules for bridging this gap do not exist. There is no purely logical method of linking concepts and indicators; there is no logical way to determine whether an indicator is really measuring the theoretically defined concept. In fact, the nearest methodologists have come to codifying these procedures is to pose the questions: Does the interpretation make sense, and does it work? (Hirschi and Selvin, 1966, p. 193).

Studies using multipie indicators of all variables are usually more solid than those using only a single indicator of each variable. This means that we should look for studies which offer several operationalizations of each independent and dependent variable. Each measure will undoubtedly contain a great deal of error. Yet the error in each of several measures of any one variable should be relatively unique. The combined measures reflect, then, more of the true score than does any single measure. Extending this even further, indices of any one variable may be interchangeable, and we should not be hypercritical if the researcher used an indicator he cleverly conceived, rather than the definitive one we had in mind. We should assess both the efficiency and extensiveness of the measures used in any research problem. For instance, a study hypothesizing the relationship of problem solving ability and college academic performance should employ more than one measure of both variables, since no single measure of either is completely satisfactory. Research efforts in which indicators fully cover the domain of predictors and criteria deserve the highest rating.

Causality. The goal of research is to show either association or causation between independent and dependent variables, but demonstration of causality has to be regarded as more of an accomplishment. A demonstration of association may be important, though, when one wants to make a decision and needs to know about the relationships among variables. If the researcher's purpose is to show causality, he must establish that A and B are statistically associated, A is prior to B in time, and the association between A and B does not disappear when the effects of other variables prior to both A and B are removed. No one of these criteria is sufficient in itself to demonstrate causality. If the analyst finds an antecedent variable influencing the relationship between A and B, he should have declared the relationship questionable and perhaps moved on to consider a new independent variable in terms of the dependent variable. If, on the other hand, he has tested the effect of several variables and finds the original relationship still holds, he can only invoke a tentative conclusion that A caused B. (Another antecedent variable, of course, may turn up later and account for the relationship.)

The problem of determining causality necessitates holding constant variables which might contaminate the relationship between A and B. As an example, an investigator may have concluded that in nursery school, a child's likableness (A) determines his degree of leadership (B) in a play group. By holding degree of aggressiveness and intelligence constant he assures himself that the relationship between A (likableness) and B (leadership) is not explained by these other variables. He has properly

held constant the variables he thinks may influence the connection between his independent and dependent variables. As reviewers we should ask if additional variables might have caused a spurious relationship. For instance, likableness may, in part, be a matter of conformity to group norms. A child is seen as likable if he conforms to the group's goals, and the initial conformity later allows him to assume a leadership role. We should suggest that conformity to group norms be examined as a possible influence on the relationship between A and B. We rate most highly that research which takes into account the possible influences of a large number of variables on the independent-dependent variable relationship.

Specification. Some variables specify the conditions under which the association between A and B holds. Research may show that good grooming and good behavior are related in children, but only under conditions of certain parental values. When the research design considers such possibilities by measuring parental values as well as grooming and behavior in children, analytic techniques can be applied to assess conditions under which the relations holds between the two major variables.

The interview or questionnaire. A number of very good discussions of procedures for and value of obtaining data through interviews and questionnaires are available (Kahn and Cannel, 1957; Kerlinger, 1964; Kornhauser and Sheatsley, 1959). The investigator using either of these techniques should indicate the method he used in developing his form and he should include a copy of it in the appendix of his report. We must consider the following criteria in assessing the instrument.

1. Are questions related to the research problems and objectives?
2. Is each question necessary?
3. Are questions clear and unambiguous?
4. Are questions leading; that is, are they posed so that one answer seems most acceptable?
5. Are all possible response alternatives available to the respondent?
6. Are some of the possible answers more socially desirable than others?
7. Do the questions demand knowledge the respondent does not have?
8. Is personal and delicate material requested which the respondent might not want to risk giving?

Finally, we should ask if the return rate on the questionnaire was so low as to suggest a bias in the sample.

Tests and scales. Research may often include the use of published tests. The author must present reliability and validity data for the instruments he employs, so we can assess their usefullness. Buro's **Sixth Mental Measurement Yearbook** (1967) may be referred to for reviews of most published tests. As a minimum, we should be sure test reliability is acceptable, and that validity data are applicable to the sample at hand (Cronbach, 1960; Ghiselli, 1964; Guion, 1965). For a test to be considered acceptable its reliability should be .80 or above. (Reliability is the consistency or repeatability of a measure; test validity is the extent to which a test measures what we think it does.)

Often researchers develop their own tests or scales. Guion (1965, pp 187 ff) offers helpful remarks on the elements of test construction, as does Guilford (1954). Attitude scales are most frequently "home made" and they come in four major types. Summated rating scales or Likert scales (1932), the Semantic Differential (Osgood, Suci, and Tannenbaum, 1957), Thurstone scales, (1929), and Guttman scales (1944) are the most common forms of attitude measurement. The problems of scaling are thoroughly discussed by Torgerson (1958) and Guilford (1954, p 414 ff). We cannot intelligently review a piece of research without understanding the nature of the scale(s) used in it. A particular scale may not be appropriate to a specific study.

Ancillary data and norms. The extent to which generalizations can be drawn is

particularly important if the research purports to add to the scientific body of knowledge or is to be used in policy determinations for populations which are somewhat unlike the original sample. If the researcher reported suitable demographic data for his sample, we can estimate the kind of individuals to which the results of the research can be reasonably extended. A study might report that twelve year old children typically have vocabularies of 1200 words. Before denigrating the value of current educational practices it might help us to know that the sample in this research was drawn from backwoods Appalachian communities. Then we know that the results of the study should be generalized only to similar children.

Evaluating the findings of any research report without normative data is capricious. The results of standardized mathematics tests given at some grade level in a particular school are meaningless unless we know the national, state, and local averages (and dispersion around the average) for the same grade level. Where norms or values representing the usual performance of a group do not exist it might be possible for the researcher to collect his own, but a reviewer cannot generally do this to make a study more relevant to his own needs. It is not itself meaningful to know that 82% of a sample attends church on Sunday. The author could have enhanced the meaning of his results by also collecting data on parallel phenomena for the same sample and then running multiple comparisons. Attendance at movies, as well as church, offers a broader understanding of the behavior (church attendance) of interest. Collection of unnecessary data, however, is expensive and should be discouraged. Shotgun approaches are no substitute for rigorous thinking prior to data collection.

Pretests. A pretest is a test to determine some kind of performance or outcome which is given before the introduction of bona fide experimental conditions. Investigators should pretest both their instruments and designs, and the pretest results should be mentioned. Designs and tests can be modified on the basis of pretest results, and then the data for the experiment can be collected and analyzed.

We have now extended the initial link in the research chain (overall problem definition) to specific conceptualizations, hypotheses, specification of variables, indicators and design. Observations are made and translated into measures. Analyses are applied to reduce batches of data to manageable proportions, to summarize them, and to make inferences about populations from which they were drawn. It is these analyses which should next concern us.

RESULTS

A good research design will appear aesthetically balanced to the experienced reader. The design served its purpose if it controlled for variations caused by error. Various designs offer alternate routes to the same destination—reliable and valid statements about relationships among variables. Once the researcher has indicated the design of his study, he is then responsible for stating his method of turning observations into quantifiable measures, his means of categorizing these measures, and his approach to data summarization and analyses. Finally, he indicates the results of his study through supplementing text with tables, diagrams, charts, or some combination of these. This information comprises the "results" section of the research report, which often is the section most difficult to understand.

Another series of connected questions may be helpful. The relevant questions here are tied in many ways to what went before, and they assume a minimal knowledge of statistics. We recommend that readers lacking such information read carefully the extended section *and* a good statistics textbook. Several excellent texts are:

Walker and Lev: *Statistical Inference*
Wallis and Roberts: *Statistics: a New Approach*
Siegel: *Nonparametric Statistics*
Hays: *Statistics for Psychologists* (for the advanced reader)

First of all, we can make coarse guesses as to whether the collected data form nominal, ordinal, interval, or ratio scales. The characteristics of each of these types of measurement are discussed in the extended section on results. In the behavioral sciences we usually assume interval scale data and apply statistics appropriate to them. However, when data are clearly nominal or ordinal, such application is unacceptable.

1. Are statistical analyses appropriate to the data?
   a. Are parametric statistics employed where data clearly do not meet the assumptions for their use?
   b. If means or other measures of central tendency were reported, are the appropriate indicators of variance also reported?
   c. If nonparametric statistics were used, does the researcher select the most powerful of these, given the constraints of his data? (He should have indicated what he thought his data constraints were).
   d. If the study included at least interval measurement, is a series of t-tests reported, or is the study designed so that an analysis of variance or multiple regression technique can be used?
   e. Are factor analyses used only under conditions where little is known about the domain of the independent variable? If factor analysis is employed, is the dependent variable measure thrown in? (This would make later testing for a relationship invalid.)
   f. If multiple indicators of the dependent variable are employed, are they weighted for combination in a single measure, is each separately analyzed against a single or a set of independent variables, or is something else done?

Different research approaches require different types of analyses, and the analytic demands imposed by various strategies are soon apparent to experienced readers. As stated previously, the objective of a laboratory experiment is to make the situation as sensitive as possible to small differences. The analytic technique must be powerful enough to indicate these small differences. Parametric statistics are the most powerful tools available, but they can only be applied to data meeting relatively stringent assumptions. A

field study, in which it is impossible to control all variables, where data fail to meet the assumptions of the more stringent analytic tools, and where large differences between groups exist, can often be subjected to less powerful analyses.

A number of other factors should not be overlooked:

2. Does the author analyze his data for obvious factors which contribute to spurious relationships?
3. Does the author specify the conditions under which his relationships hold?
4. If time was an important factor in the study, does the author do a time series data analysis?
5. If the study is one which lends itself to various *ex post facto* analyses, does the author do them?

Again, there are no rules for deciding what should be done. This is a logical problem and we can only assess on the basis of what we think should obviously have been looked at. Once satisfied with the categorization of measures and the analysis, we ask:

6. Do reported differences between or among groups reach statistical significance?

We must not ignore the presentation of results, for they contribute to the clarity of the report. As a minimum we ask:

7. Are tables and graphs presented in an understandable fashion?
   a. Is the meaning assigned horizontal and vertical axes in graphs understood? Generally, the independent variable is plotted on the horizontal axis, and the dependent variable on the vertical axis.
   b. Are all tables and graphs clearly titled and labelled?
   c. Where data are discrete rather than continuous, are bar graphs used instead of continuous graphs?

The results section should present the data analyses in a straightforward manner. To a considerable extent, the evaluation of a research project rests on the results section. Obviously, hypothesis testing which leads to significant differences between groups of subjects who experience different treatment conditions deserves a high rating. However, significant differences which are a result of improper data analysis are of no value whatsoever. On the other hand, proper analysis of data which finds no significant differences between groups may also be valuable, particularly in policy formation. For example, Chu and Schramm (1968) reviewed 421 studies comparing instructional television with conventional teaching. Twelve per cent of the studies favored conventional teaching, 15 per cent favored television, and 73 per cent showed no significant differences between the two modes of instruction. Extrapolating these data to hypothetical developing nations where education is needed and where the cost in time and money of training sufficient numbers of teachers is exorbitant, perhaps the wise policy decision would be to use instructional television. This would be a policy decision based on rigorous analyses that showed *no* significant differences.

We cannot assess the value of any analytic technique unless it is clearly described. Some readers may wish to examine the problems of data analysis more thoroughly before going on to consider the Discussion and Conclusions section of a report. The next section is to help them.

Results—From Categorization to Description and Inference

We are in a difficult position when we assess the results section of a research report. Data reduction and analyses demand technical expertise and the research consumer can either learn as much about data analysis as the researcher had to know to

carry out his study, or he can seek the aid of a consulting statistician. It is not the purpose of this paper to train reviewers in statistics. Only the very grossest hints can be offered here as aids. As stated before, a comprehensive understanding of statistics is essential in assessing research reports, and a number of excellent texts are available in the field.

Statistics serve us in at least two capacities. Descriptive statistics are for the purpose of organizing, summarizing and communicating data. A descriptive statistic is a single number which characterizes a larger series of numbers. Percentages, averages, and standard deviations are descriptive statistics. Inferential statistics are used to arrive at conclusions extending beyond the immediate data. The procedures of statistical inference introduce order into attempts to reach conclusions about populations, order based on evidence drawn from samples. Hypothesis testing involves statistical inference.

The known procedures for analyzing data dictate conditions for collecting evidence. In reverse, as stated previously, the conditions of data collection determine what statistical methods are applicable. Statistical tests tell us how large an observed difference must be before we have confidence that it represents a real difference in the groups from which the events were sampled.

Measurement scales. The appropriate use of a particular statistical test in a piece of research depends partly on the measurement scale adopted by the researcher. The weakest form of measurement **(nominal)** is the use of symbols to simply classify or identify objects. The naming of colors—red, blue, green, etc.—is an example of the nominal system.

Ordinal scales are measurement procedures for obtaining relations among a series of items. They indicate that one item is greater than, is preferred to, or is more difficult than another item in the same series. The system of school grade differentials is an example of ordinal scaling—the sixth grader is higher than the fifth, etc. It would be impossible, however, to say that the sixth grade is as different from the fifth as the fifth grade is from the fourth.

When a scale has all the properties of an ordinal scale and when the distrances between any two numbers are of known size and equal, an **interval scale** in achieved. The zero point and unit of measurement for interval scales are both arbitrary. Temperature is measured on an interval scale. Interval and ratio scales are truly quantitative. **Ratio scales** have all the characteristics of interval scales in addition to a true zero point. The number of oranges in a basket is measured by a ratio scale.

Broad considerations for application of specific statistical tests. Several criteria beyond the nature of the measurement are important in selecting the statistical test appropriate to any set of data. The nature of the sampling and the population from which the sample came should both be considered. The power of a statistical test must be looked at in relation to the power of alternative tests. A statistical test is powerful if, for the particular data set, it has a small probability of rejecting the null hypothesis when it is true, and a large probability of rejecting it when it is false. We must know something of the logic involved in assessing the power of a given test for a given set of data if we are to adequately evaluate research reports. The most powerful available appropriate test should be used in any project.

For any study, the nature of the population, the manner in which the sampling was done, and the nature of the measurement determine the appropriate statistical analyses. Associated with every statistical test are some sampling and measurement requirements. A statistical test is valid for observations if they meet these requirements. Requirements include such assumptions as normality of distribution of the measured factor in the population from which data were drawn, equal measurement variances among subgroups in the sample, and specification as to whether the test is valid for nominal, ordinal, interval, or ratio data. A test is "robust" if some of these requirements can be ignored without seriously invalidating the statistical results.

Parametric versus non-parametric tests. Inferential statistics can be divided into two major types. Parametric statistics include product moment correlations, t-tests, analyses of variance, and other analysis techniques. They can only be used with data meeting fairly rigorous assumptions. Nonparametric statistics include chi squares, sign tests, ranking statistics, etc. and make weaker assumptions about data. They are also less powerful than the parametric tests.

Both types of analysis require that:

1. Observations be independent of one another.

Additionally, all applications of parametric tests technically require that.

2. Observations be drawn from normally distributed populations.
3. Populations from which the observations were drawn have equal variances.
4. Measurement of variables is on at least an interval scale.

The central tendency (a measure which typifies the distribution) for interval or ratio scale data is best described by the mean (if there are no extreme scores). The appropriate measure of dispersion about the mean is the variance. All the common parametric tests (product moment correlations, t-tests, F-tests) can be applied to **interval** or **ratio** scale data only. The median is the most appropriate description of the central tendency for ordinal data, and an appropriate measure of variance is the interquartile range. A large group of nonparametric tests called ranking statistics can be applied to **ordinal** data. Hypotheses can sometimes be tested for **nominal** data. Nonparametric statistics such as chi squares or tests based on the binomial expansion and a measurement of association called the contingency coefficient are appropriate to certain forms of categorization. We negatively evaluate analyses using statistical tests making different assumptions than the data fulfill.

There are several rules regarding appropriate distinctions for the use of parametric and nonparametric tests:

1. If sample sizes as small as six are used, parametric statistics are not applicable.
2. If the observations are drawn from different populations, nonparametric statistics sometimes are applicable, but parametric tests are almost never appropriate.
3. If data are ranked, nonparametric tests are usually suitable.
4. Nonparametric, but not parametric, tests are available for data which are simply classifiable.
5. If the assumptions of parametric tests are met, the application of nonparametric tests is wasteful unless an analysis which can quickly be done by hand is desired.

Analyses of variance and factor analyses are the most sophisticated of the parametric tests and are, strictly speaking, applicable only to data meeting the relatively rigorous assumptions of parametric testing. However, these tests are relatively robust and they can sometimes be used with data ignoring certain assumptions (Guilford, 1954, Hays, 1963).

Analyses of variance are appropriate where a group can be subdivided on one or more independent variables, and measured by at least an interval scale on the dependent variable. Data on the learning ability of three groups, each exposed to a different method of teaching history, might be analyzed through analysis of variance. Factor or cluster analyses are usually seen where we know little about the relationship among variables in a

domain. An example of this is the current interest in delinquency research of the effect of anomie (A) on delinquent behavior (B). What variables constitute anomie? A factor analysis of variables thought to contribute to anomie might increase understanding, but if the dependent variable in the A-B relationship is delinquent behavior, the factor analysis should not include a measure of delinquent behavior as a possible aspect of anomie. Factor analyses should never include the variable to be predicted by the resulting factors.

A discussion of multivariate analyses (analyses combining sets of variables to predict some phenomenon) is beyond the scope of this paper. Faced with a project requiring such analysis, you might well refer to an advanced text such as Cattell (1966).

Other analyses. The researcher wishing to assess the viability of a relationship between two variables should have held other variables constant, one at a time, and observed whether the relationship still exists or is spurious. This is generally accomplished by using partial correlation coefficients. Partialling identifies spurious relationships and conditions for the existence of relationships.

The most common form of time series analysis is the correlation of events occurring at one time with those taking place at some different time. Such analyses are usually done with time lag correlations, which correlate changes in a variable occurring at time one with changes in the same or some other variable occurring at a different time.

Significance levels. Since most tests of hypotheses report the "significance" of the difference between two groups, or differences among groups, it is essential that we know what this means. One kind of error is that of "seeing too much in the data." The other kind of error is "not seeing enough in the data." The significance level tells us the proportion of times the experimenter could be expected to err in the direction of seeing too much in the data. If he accepted a significance level of .05, he decided to regard as real those effects which could have been produced by chance five times in one hundred. It is traditional in the behavioral science to accept significance levels of .05 or .01. Under either of these decision rules the probability of accepting a difference which in fact does not exist is low. We might be willing to take somewhat greater chances, and significance levels should be looked at with an eye to the consequence of the decision to be made on the basis of the research result. If a school district is to adopt an innoculation program against polio, the person ultimately making the decision should be unwilling to adopt the program unless the vaccine research indicates with a high degree of certainty that the vaccine is not dangerous to life. On the other hand, evidence of the benefits of a proposed manual arts training program can be accepted with considerably less certainty if it is not too expensive and if it is known to do no harm.

Tables, graphs, and figures. Tables, graphs, and figures should supplement, not duplicate the text. Good tabular presentation displays quantitative data systematically, precisely, and economically. Designing a clear, uncluttered table is an art, and we should not be too hard on the researcher who is unable to do this. Tables should be self-explanatory and should note significant differences between groups. We should be able to compare data both within and among tables.

Graphs and other figures should be intelligent additions to the understanding of data. We ought to have no difficulty assessing whether a graph or other figure helps us to make sense of the text.

Once we have digested the results of a research report, we must consider its meaning. The final section of a report details the author's summary of his work and his thought about its implications.

DISCUSSION AND CONCLUSIONS

The results section of a report focuses on the cake, and the discussion section adds the icing. We can never accept an investigation as valuable simply because the author says it is. The discussion section helps us make a final decision about the worth of the report. We ask:

1. Does the author state inferences which could be drawn from his findings?
   a. Do these inferences seem logical in light of the limitations of the study? (These limitations depend on the way the problem was presented, the design, sample selection, the way data were collected, data analyses, etc.)
   b. Are the inferences of any practical or scientific merit? Was the study worth doing?
2. Does the author indicate the qualifications to his study which limited the inferences he was able to draw?
   a. Does he commit the common fault of making too much of his data?
   b. Or, does he seem too modest in his conclusions?
3. Does the author relate his findings to other work in the field?
4. Is a discussion of still unanswered or new questions included?
   a. Are new approaches to the problem considered?
   b. Are the questions brought out by the research important ones?

Finally, our job is not complete until we assess the tradeoff between the cost and value of any study. As we place the report back on the shelf we should ask ourselves:

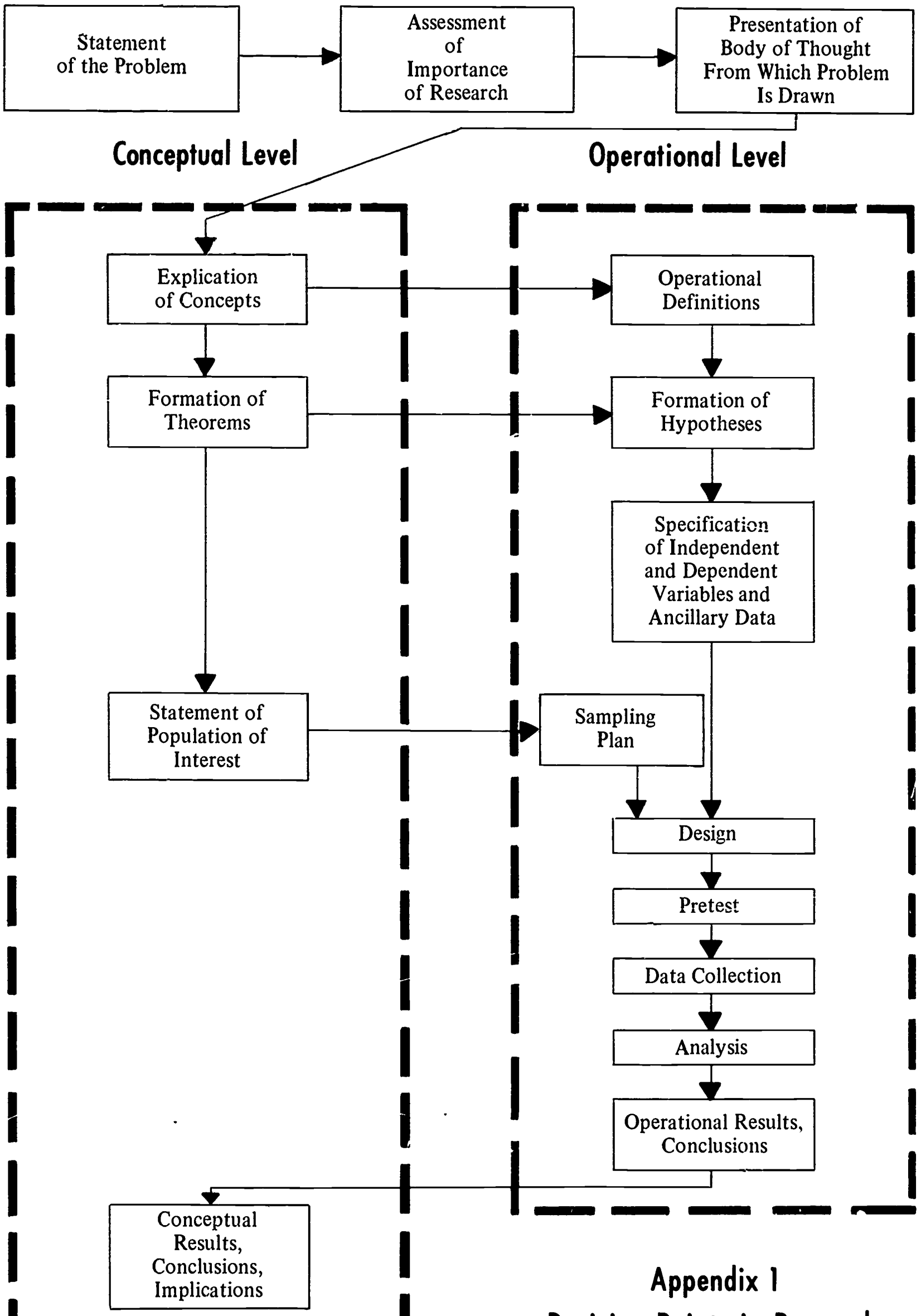
5. Do I now know anything about the subject matter I did not know before?
6. Can I trust that new knowledge and, if it is trustworthy,
7. Can someone make use of what was presented either in terms of advancing knowledge or deciding upon policy?

If an investigator has asked an important question, completed an adequate piece of research, provided appropriate inferences, and then suggested reasonable alternative or novel approaches to his problem, he has made an especially useful contribution.

EPILOGUE

The purpose of this paper has been to underscore the basic points in research evaluation, and to indicate the relationships among practical need, theory, hypotheses and data collection.

Problem solutions cannot always await the development of appropriate theory. So as implementers of behavioral science research we must not ignore available data, but we must use approaches and techniques having less than complete research support. Our societal problems are so acute that we simply must support and demand competent research related to them, and then jump in with action programs based on evidence gathered in less than a lifetime.



Appendix 1

Decision Points in Research

APPENDIX 2–NOTES ON ASSESSING RESEARCH REPORTS

The following points should be considered in assessing any research project. The appropriate answer to each question is "yes," and the italicized questions should be given special emphasis. From these questions, the reviewer can develop his own checklist. The number of "yeses" required for any project to be considered adequate depends on the reviewer's standards for acceptable research, and on his purpose in reviewing the project.

Problem Presentation

Is the area of interest presented in an understandable way?
Is the question asked an important one?
Will the answer to the stated problem aid in furthering scientific understanding?
Will the answer to the problem provide useful guidelines to decision-making?
Is the problem well explicated, given the limitations of the research area?
Is sufficient information provided so that one can understand the concepts involved?
Are variables operationalized such that they maintain their relevance to the overall concepts and purposes of the research?
In light of the concepts and variables presented, are reasonable, testable propositions framed?

Method

Are the hypotheses stated in terms of expected differences?
Could someone reading this report replicate the research?
Is the population of interest defined?
Is sampling from it adequate?
Is the regression effect avoided?
Given the limitations imposed by the nature of the research problem, is the design adequate?
Are the variable measures appropriate?
Are appropriate ancillary data collected?
Is the study designed to provide evidence of causation or correlation?
Considering the design and sampling techniques, are the results generalizable to groups other than those studied?

Results

Are observational categories relevant?
Are subcategories for any one variable mutually exclusive?
Are they exhaustive?
Are the categories for different variables independent of one another?
Are the statistical analyses appropriate to the data?
Do reported differences reach statistical significance?
Are the results presented so they are understandable?

Discussion and Conclusion

Are logical inferences drawn from the findings?
If only correlational analyses were possible, did the author avoid inferences about causation?
Are the inferences of any practical or scientific value?
Are the generalizations appropriate?
Are the limitations of the research spelled out?
Are still unanswered questions considered?

BIBLIOGRAPHY

American Psychological Association Council of Editors. Publication manual of the American Psychological Association. *Psychological Bulletin,* 1967.

Anderson, B. F. *The Psychology Experiment.* Belmont, California: Brooks/Cole, 1968.

Blalock, H. M. *Causal Inferences from Nonexperimental Research.* Chapel Hill: University of North Carolina Press, 1964.

Campbell, D. T. Factors relevant to the validity of experiments in social settings. *Psychological Bulletin,* 1957, *54,* 297-312.

Campbell, D. T. and Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (ed.), *Handbook of Research on Teaching.* Chicago: Rand McNally, 1963, pp. 171-246.

Cattell, R. B. *Handbook of Multivariate Experimental Psychology.* Chicago: Rand-McNally, 1966.

Chein, I. An introduction to sampling. In Selltiz, Claire *et al., Research Methods in Social Relations.* New York: Holt, 1959.

Chu, G., and Schramm, W. Learning from television: what the research says. U.S.O.E. Contract OEC 4-7-0071123-4203. 1967.

Cronbach, L. J. *Essentials of Psychological Testing.* New York: Harper, 1960.

Fisher, R. A. *The Design of Experiments.* London: Oliver and Boyd, 1951.

Ghiselli, E. E. *Theory of Psychological Measurement.* New York: McGraw-Hill, 1964.

Guion, R. M. *Personnel Testing.* New York: McGraw-Hill, 1965.

Guilford, J. P. *Psychometric Methods.* New York: McGraw-Hill, 1954.

Guttman, L. A basis for scaling qualitative data. *American Sociological Review, 1944, 9,* 139-150.

Hays, W. L. *Statistics for Psychologists.* New York: Holt, 1963.

Hempel, C. G. Fundamentals of concept formation in empirical science. In *International Encyclopedia of Unified Science.* Volumes I and II. no. 7. Chicago: University of Chicago Press, 1952.

Hirschi, T., and Selvin, H. C. *Delinquency Research: An Appraisal of Analytical Methods.* New York: Free Press, 1967.

Hyman, H. *Survey Design and Analysis.* New York: Free Press, 1955.

Kahn, R. L., and Cannell, C. F. *The Dynamics of Interviewing.* New York: Wiley, 1957.

Kerlinger, F. N. *Foundations of Behavioral Research.* New York: Holt, 1964.

Kornhauser, A., and Sheatsley, P. B. Questionnaire construction and interview procedure. In Selltiz, Claire, *et al., Research Methods in Social Relations.* New York: Holt, 1959.

Likert, R. A technique for the measurement of attitudes. *Psychological Archives,* No. 140, 1-55.

McNemar, Q. *Psychological Statistics.* New York: Wiley, 1962.

Orne, M. T. On the social psychology of the psychological experiment: with particular reference to demand characteristics and their applications. *American Psychologist,* 1962, *17,* 776-783.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. *The Measurement of Meaning.* Urbana: University of Illinois Press, 1957.

Selltiz, Claire, *et al., Research Methods in Social Relations.* New York: Holt, 1959.

Siegel, S. *Nonparametric Statistics for the Behavioral Sciences.* New York: McGraw-Hill, 1956.

Thurstone, L. L. Theory of attitude measurements. *Psychological Review,* 1929, *36,* 222-241.

Torgerson, W. S. *Theory and Methods of Scaling.* New York: Wiley, 1958.

Underwood, B. *Psychological Research.* New York: Appleton, 1951.

Walker, Helen, and Lev, J. *Statistical Inference.* New York: Holt, 1953.

Wallis, W. A., and Roberts, H. V. *Statistics: a New Approach.* Glencoe, Illinois: Free Press, 1962.

Webb, E. J. *et al., Unobtrusive Measures: Nonreactive Research in the Social Sciences.* Chicago: Rand-McNally, 1966.

Whitehead, A. N. *An Introduction to Mathematics.* New York: Holt, 1911.

This paper is distributed pursuant to a contract with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Points of view or opinions do not, therefore, necessarily represent official Office of Education position or policy.