1st LL 1-29-79

8978

EXPOSURE DRAFT

ASSESSING SOCIAL PROGRAM IMPACT EVALUATIONS: A CHECKLIST APPROACH

UNITED STATES GENERAL ACCOUNTING OFFICE
PAD-79-2
October 1978

Rpt. ff

108461

003354 QC

COMPTROLLER GENERAL OF THE UNITED STATES
WASHINGTON, D.C. 20548

B-161740

In a democracy such as ours, assuring that governments and agencies entrusted with public resources and the authority for applying them are held accountable is a never ending responsibility. Evaluators have a special role in meeting this responsibility, but they too must be held accountable for the quality of work performed with resources entrusted to them. Thus, those who evaluate must also be evaluated.

The Congress, executive policymakers, and program administrators need some assurance that the evaluations they wish to use were properly planned and conducted and that results were reported clearly, completely, and fairly. This checklist, which we believe should be applied in assessing all social program impact evaluations, provides a systematic framework for organizing the evidence necessary to support such assessments.

Our overall purpose in publishing this checklist is to encourage the conduct of high quality evaluations and to promote more effective use of those evaluations in public decisionmaking. The checklist is intended to be of value not only to those who reanalyze the evaluations of others, but to all those engaged in the evaluation process, including evaluators, sponsors, and decisionmakers who wish to use evaluations.

This document is being issued as an exposure draft to allow consideration of any problems auditors or evaluators may have in using it before it is finalized. We would appreciate receiving your comments on how it can be improved. Please send these comments to Harry S. Havens, Director, Program Analysis Division.

Elmer B. Staats
Comptroller General
of the United States

ACKNOWLEDGMENT

The General Accounting Office has drawn from a number of sources in preparing this checklist. In particular, Drs. Robert Boruch, David Rindskopf, and Eva Rezmovic of the Department of Psychology at Northwestern University contributed an original draft which was used substantially in the development of this checklist. Dr. Paul Holland of Educational Testing Service prepared a paper on causes of bias in data analysis which we used extensively in the data analysis section. Also, Jeffrey Rosenberg, now employed by Peat, Marwick, Mitchell, and Co. Inc., contributed a very useful paper on sources of bias in social research, while a student at the University of Chicago and student-intern with the General Accounting Office.

Contents

CHAPTER 1

INTRODUCTION

As resources for solving the problems of our society become increasingly scarce, the need to apply them more effectively increases. Impact evaluations are a vital part of that effort. 1/ Policymakers and administrators need them and the public interest demands them.

A survey recently completed by the Office of Management and Budget (OMB) showed that in fiscal year 1977 over $243 million was obligated by the executive branch for program evaluation. In addition, a recent study by the National Research Council showed that in fiscal year 1976 over $1.8 billion was spent on social knowledge production and application activities, including a substantial amount for evaluations of ongoing social programs. Public pressures to reduce the growth of Government programs and improve their effectiveness point to increased demands for evaluation in the future. Recent or proposed reforms in the congressional budget and oversight processes have also given stimulus to this trend.

The Congressional Budget Act of 1974 requires the Congress to consider particular spending demands in the light of overall national priorities and other claims on the budget. The Sunset Act of 1978 was passed by the Senate on October 11, 1978. This and other proposals introduced recently into the Congress would require scheduled evaluation and review of Federal programs. None of these reforms can be effectively carried out without evaluating the impact of current policies, programs, and activities.

Sophisticated techniques are constantly being designed to improve evaluation methodology and the accuracy of study findings. However, there is still a wide gap between our technical ability to evaluate programs and our ability to manage these evaluations so that their results directly aid decisionmakers and properly inform the public.

1/ Impact evaluation is the process of appraising the extent to which programs are (1) achieving their stated objectives, (2) meeting the performance perceptions and expectations of decisionmakers, individuals, and groups affected by the program, and others with a legitimate interest in it, and (3) producing other significant effects of either a desirable or undesirable character.

About 70 percent of the $243 million cited in the OMB study was for evaluations done under contracts or grants. This considerable Federal expenditure has made program evaluation a multimillion dollar industry. A large number of organizations, including universities, "think tank" institutions, private firms, and individual consultants have entered the evaluation arena.

The growth in this industry has been rapid and largely uncontrolled and has given rise to a number of persistent issues. These issues include the inadequacy of methods for assuring that evaluators are held accountable for their activities, and concerns over the lack of criteria to insure the quality of the evaluations. Partly as a result of these concerns, many evaluations are still done but lie dormant in remote files or on some decisionmaker's bookshelf.

While we recognize that there are other legitimate explanations for the decisionmakers' reluctance to use evaluations, we believe that to achieve the high quality necessary to make them useful, evaluations must themselves meet the following minimum criteria:

--Relevance--the evaluations must provide the information needed by a variety of audiences, especially decisionmakers, and must answer the right questions at the right time.

--Significance--the information must tell users something new and important; it must go beyond what is already apparent to them.

--Validity--the evaluation must provide a reasonably balanced picture of the real effects of the program or activity in question.

--Reliability--the evaluation must contain evidence that the conclusions are not based on variations in the data which are due to chance or inconsistent measurement.

--Objectivity--results must be conveyed in a complete and unbiased manner.

--Timeliness--the information must be available in usable form when decisions have to be made.

This checklist is a practical framework for determining whether these criteria are met. In our judgment, all of these items contribute to high quality impact evaluations, and are both legitimate and feasible in a variety of settings. In particular settings, some items may not be relevant, but their absence should be fully explained.

Chapter 2

CHECKLIST FOR ASSESSING SOCIAL PROGRAM IMPACT EVALUATIONS

The following checklist will enable the completeness of an evaluation effort to be determined quickly. The following notations can also be used for a preliminary quality assessment.

S--Task was performed in a satisfactory manner.

U--Task was performed but result appears unsatisfactory.

NP--Task considered necessary but was not performed.

NA--Task considered not applicable for the particular evaluation.

The preliminary quality assessment is a basis for planning a more indepth review of specific areas of the evaluation, and the checklist can be used again at the end of the review to summarize the assessment.

A. PLANNING

1. Have evaluation goals been defined and described?

   --Have decisionmakers been identified? _____

   --Have decisionmakers' needs been determined? _____

   --Have the perceptions and expectations of those affected by the program or others with a legitimate interest in it been considered? _____

   --Has the plan clearly stated what the evaluation seeks to accomplish? _____

2. Has the evaluability of the program been determined?

   --Has the problem the program is supposed to resolve been described? _____

--Have the reasons for believing the program will resolve the problem been stated? _____

--Have the program's objectives been made explicit? _____

--Has evidence been obtained to show that the program was implemented and specific activities are being carried out? _____

3. Has a clear evaluation approach been developed and justified, and potential threats to the validity of conclusions and inferences anticipated and accommodated?

--Has a literature search and synthesis of previous evaluations been performed? _____

--Have reasons for selecting a particular comparison method been stated? _____

--Have the limitations of the method been recognized for the given hypotheses and questions? _____

--Has the plan provided for identifying and weighing alternative causes of the measured impacts? _____

4. Has the method for sample selection been explained and justified?

--Has the program population been clearly described and provision made to accommodate possible bias caused by the program selection procedures? _____

--Have the sampling units been justified on the basis of the comparisons to be made? _____

--Have the sample selection procedures been justified? _____

--Have the required sample sizes been determined? _____

--Have incentives for participants to cooperate and mechanisms for monitoring and analyzing attrition been established? _____

5. Have measurement methods been identified and their validity and reliability assessed? _____

6. Have the frequency and timing of measurements been specified and explained? _____

7. Has the feasibility of performing the evaluation been examined?

   --Has the time required to obtain the needed information been estimated and compared to the required timing of decisions? _____

   --Have needed resources been identified and compared with those available? _____

   --Have legal constraints been identified and plans made to accommodate them? _____

   --Have political influences been identified and plans made to deal with them? _____

8. Has the necessary cooperation been obtained?

   --Has the cooperation of those administering the program been obtained? _____

   --Has community cooperation been obtained or constraints identified and plans made to accommodate them? _____

B. DATA COLLECTION

1. Have procedures for quality control of data been identified and implemented?

   --Have data collection instruments been pilot tested and standardized as appropriate? _____

   --Have qualifications and standards for employment been established and used in recruiting? _____

   --Have employees been provided with the necessary training? _____

--Have assignment procedures been established and adhered to? _____

--Have supervisory controls been implemented to identify those employees who need retraining or replacement? _____

--Have editing procedures been established and implemented to insure that inadvertent errors do not become a part of the data base? _____

--Have procedures been implemented for monitoring data collected in the program processes? _____

2. Have preliminary analyses been performed to detect missing or inconsistent information and correct deficiencies in the study plan?

--Have methods for determining the characteristics of nonrespondents been implemented? _____

--Has consistency of data collected from different sources been tested? _____

C. DATA ANALYSIS

1. Have the statistical methods and model for use in the analysis and the rationale for their selection been specified?

--Has the relationship of the statistical methods to the hypothesis or questions been specified? _____

--Has the relationship of the statistical model to the evaluation design been specified? _____

--Has the relationship of the statistical methods to the data been specified? _____

2. Has the unit of analysis been justified? _____

3. Have the assumptions essential to statistical methods and model been specified and have their conditions been met? _____

D. REPORTING RESULTS

1. Have the findings been presented clearly, completely, and fairly?

--Have all significant results of the analysis been discussed in terms of the decision-makers' needs? _____

--Have points of equal importance been given equal emphasis? _____

--Have all assumptions been made explicit? _____

--Has the evaluation approach been described in sufficient detail for users to understand what was done and why? _____

--Have issues and questions which need further study and consideration been identified and explained? _____

2. Have specific procedures been used to assure the report's quality? _____

3. Have followup provisions been made to assist decisionmakers in using the report? _____

E. DATA DISCLOSURE

1. Has adequate documentation been maintained?

--Have the purpose and sources of data and the methods and circumstances of their collection been documented? _____

--Have the content and organization of the documentation been clearly described? _____

2. Has a procedure been established for release of data for audit, reanalysis, other evaluations or research?

   --Have the location of data and officials authorized to release it been identified? _____

   --Have conditions for the release of data been specified? _____

CHAPTER 3

GUIDELINES FOR APPLYING THE CHECKLIST

This checklist is the result of a long evolution in evaluation methodology. To the extent that there are further advances in the state of the art of evaluation approaches, methods, and management, evaluation problems and their solutions will also change. Such an evolution will not invalidate this checklist but may justify its expansion and improve the extent to which particular items can be accommodated.

The task of establishing the degree to which each item in the checklist is accommodated is not a simple one. For example, many evaluations emphasize process and workload measures more than impact measures, thus, this checklist may be less applicable. The following guidelines are offered to explain the items and provide the basic criteria needed for a preliminary assessment. An indepth review of the items, however, will require more technical criteria and in some cases the assistance of experts.

A. EVALUATION PLANNING

Planning is the cornerstone of good evaluation management. Effective planning is an iterative process of continued interaction among evaluators, sponsors, and the appropriate decisionmakers.

Written documentation of the evaluation plan is fundamental. A substantial effort should be devoted to drawing up a comprehensive and thorough study plan because it will serve as a guide for subsequent work. The following are essential elements of the plan:

--A clear statement of the problem to be studied, questions to be answered, and decisions to be affected.

--A careful listing of constraints and assumptions.

--A statement and explanation of the evaluation approach and methods to be used.

--A specification of the resources to be committed, including identification of the key staff members.

--The frequency, format, and recipients of reports.

--Procedures for amending the study plan.

--The timeframe for the major components of the study and the final deadline.

1. Defining evaluation goals

Evaluation goals must be carefully defined because evaluations which address the wrong questions or do not provide adequate information for decisionmaking are not successful. The primary purpose of impact evaluations is to provide information for making a myriad of management and policy decisions

--program managers want to know what the programs entrusted to them are accomplishing and whether results might be improved,

--those contemplating similar programs want to know what techniques and methods work and why, and

--central agencies and legislative bodies need to know whether programs are implemented as intended, and which programs are accomplishing the most good.

The potential users of evaluation results consist of two major types of decisionmakers, program administrators who use results within the ongoing program to make decisions about modifying or restructuring it, and various interested groups outside the program setting. These groups include policy and budget officials and legislative bodies who may also consider possible expansion, reduction or termination of the program; agencies operating or contemplating similar programs; standard-setting or granting bodies; many other policymaking units at the Federal, State, or local level; and interest groups in the private sector.

It is obvious that these different users may have different needs for evaluations; sometimes their information requirements conflict because they do not all have the same interests. Agency evaluation criteria and standards may appear inconsistent with legislative intent or actual program activities and there may even be disagreement among agency officials, the Congress, and State/local officials about what the program is intended to accomplish, and what criteria should be used to define success.

It is the responsibility of sponsors and evaluators to recognize differences in opinion about what the evaluation should or can accomplish, to reconcile these differences where possible, and to provide a rationale for the evaluation. The key to accomplishing this would seem to be a process which produces discussion and agreement among sponsors, decisionmakers, and evaluators. The first step in this process is to identify the appropriate decisionmakers and determine their information needs. While the evaluator and sponsor may eventually have to make choices or rank the evaluation goals, it is important that the users' needs be identified and integrated as effectively as possible during the planning process.

Consideration should be given to the expectations and perceptions of individuals or groups affected by the program, and others with a legitimate interest. Although these people are not authorized to make decisions about terminating, modifying, or expanding the program, they can influence the decisions. This consideration may also enhance the evaluation value by contributing to a greater understanding of the social problem which the program is designed to improve.

The discussions with decisionmakers and other interested parties should enable the evaluator to determine their informational needs, but before the final study goals can be established, the evaluator must also determine if the program can be evaluated and if it is feasible to obtain the needed information. Matching informational needs with potentially available information is an iterative process which can require considerable discussions with decisionmakers, other interested parties, and the people who carry out the program on a day-to-day basis. It should enable the researcher to finalize the evaluation goals. These goals should be clearly stated in the study plan because the receptiveness of the final evaluation report will depend in part on the people's expectations.

2. Determining evaluability

The program's evaluability should be determined early in the planning process. This means determining whether the program has been implemented in such a way that its impact can be evaluated. Such a process requires an understanding of both the theoretical basis for the program and how program activities are actually accomplished.

To understand the theoretical basis for a program, the evaluator must first learn what is known about the problem the program is expected to resolve and the reasons why the program will resolve the problem. This information may be contained in a statement of the program objectives, but such objectives are quite often hazy, ambiguous, and hard to pin down. To clarify them, discussions with policymakers and program administrators may be necessary.

Before attempting to study the program's impact, the evaluator must also determine that the program has actually been implemented and the intended operations are being accomplished. Most programs require time to stabilize and go through a development phase where many changes in the process occur. During this phase, any impact evaluation would have to be based on trends and estimates drawn from meager data, thus, it may be impossible to provide the type of information needed for decisionmaking outside the program setting. On the other hand, formative evaluation which assists program administrators to modify, refine, and improve program processes should be very useful at this stage.

3. Developing an evaluation approach

The evaluators must develop an approach which will enable them to provide the needed information. This approach is usually concerned with the certainty and confidence with which the measured effects can be attributed to the program and the results can be generalized to other settings, respectively

No particular approach is inherently the appropriate one. The basic issue is one of fitting the research design to the goals of the evaluation and the availability of data. The evaluator should select the approach which yields the greatest certainty consistent with the evaluation goals, available time and funding, and the realities of the program. Although ideal models exist, most good approaches represent a compromise dictated by all the practical considerations that go into an evaluation.

There are several plausible explanations for any social phenomenon, but the evaluator must be able to identify the program's contribution to any effect measured. The certainty with which this cause and effect relationship can be established is referred to as the study's "internal validity." Without some degree of internal validity, the evaluation is useless; when the effects of the program remain confused with

those of other programs or of the environment, there is no adequate basis for making decisions about the relative value of the program or of its elements.

Establishing internal validity requires comparison. The most rigorous form of comparison is the experimental method which attempts to measure the results of the program as though everything else is held constant. This is accomplished by measuring relevant group characteristics or behavior which has been exposed to the program and of a control group which has not. The classical experimental method requires that, except for exposure to the program, these two groups possess nearly identical characteristics that can only be achieved by random assignment to the two groups. More sensitive results can be obtained with variations of random assignment procedures. For example, eligible individuals may be matched on certain characteristics first, and then each pair randomly assigned to either the program or control group. The analytical strength of the experimental method makes it a very useful tool, but its value must be balanced against other considerations such as cost and ethical, institutional, or legal constraints.

When less certainty is acceptable, other methods are available. These include (1) comparing the program participants with a group of nonparticipants not randomly assigned, but matched as closely as possible in certain aspects such as age, sex, race, and socioeconomic characteristics, (2) comparing periodic measurements within the program group to detect changes over time, or (3) comparing measurements from one program with those from another. Many such methods are available and are most useful when the evaluation goals do not require avoiding every possible source of confusion in measuring the program's impact. The aim, however, still is to identify and eliminate those sources of confusion most likely to be significant in a given evaluation. The evaluator must always be alert to factors outside the program which could explain the measured impact and search for other analyses or extensions of the data which could help to rule out these factors.

The second issue which must be addressed in the evaluation approach is "external validity," or the extent to which results can be generalized to those settings not actually included in the evaluation. The success of any program or project depends, in part, on the setting in which it takes place and the program's impact; therefore, it may differ from one environment to another. The question of external

validity is never completely answerable; evidence about it often is obtained only gradually. To be most useful, the evaluator must provide convincing evidence that the results can be applied in those settings of interest to the decision-makers.

The primary tool for establishing external validity is replication of the evaluation in diverse settings, especially those settings which are thought to differ on crucial variables. Because both time and resources available for evaluation will limit replication, the interaction between the researcher and the user is crucial. The evaluator should determine which settings are of interest to the decisionmaker and select sites and populations which are as representative as possible of the environments in which the study findings are to be applied.

4. Method for sample selection

Sample selection is a critical element of most social program evaluations. Sampling is nothing more than a practical technique which makes measurement of social phenomena feasible. If measurements could be obtained from all those people who participate in a program and all those who did not, sampling would be unnecessary. Time and resource constraints almost always preclude such large measurement programs however, and the problem becomes one of measuring representative groups or samples.

Three questions must be addressed in arriving at a sampling method (1) what procedure can be used to assure that the samples fairly represent the populations to which observations will be attributed? (2) what can be done to assure that inevitable changes in the samples do not seriously affect their representativeness, and therefore, the quality of estimates of program effect? and (3) how large must the samples be to detect program impact with an adequate degree of certainty?

Assuring that samples are representative of the populations to which the observations will be attributed requires (1) an understanding of the criteria used to select the program participants and (2) appropriate techniques for selecting samples from those participants and other groups. The evaluator must attempt to identify ways in which these selection processes can affect the evaluation results.

If evaluators could decide which people or groups would participate in a program and which would not, assuring representativeness would be straightforward. Such is not the case, however. Frequently, for example, the criteria for program participation is established well before the evaluation begins and quite often participants have already been selected into the program. Even when evaluation needs are considered in planning and implementing programs, legal, ethical, and other constraints must be satisfied in selecting program participants.

When the evaluation design requires comparing program participants with nonparticipants, or when results must be generalized to a population other than the current participants, the evaluator must fully understand the selection process. This requires not only a description of the selection criteria, but a knowledge of how that criteria was actually applied since established criteria is often modified by program staff. In some cases, the selection procedures may be ambiguous and target population ill defined. Considerable effort may be needed to clarify these issues.

Techniques used to select samples from the program participants and other groups can significantly affect the evaluation results. Generally speaking, some form of random selection procedures are preferable because they are straightforward and permit greater confidence in generalizations. The guiding principle is that the subjects of the study should be typical of those people to whom the study findings will be attributed. The evaluation plan should clearly describe the sample selection procedures and the reasons for their use.

Some loss of representativeness in implementing sample designs is inevitable. In most instances, individual units selected to comprise the sample will have the option of cooperating or refusing to cooperate. Having initially agreed to cooperate, some will decide to drop out; others will remain, but for a variety of reasons will provide incomplete data.

Just as initial selection procedures can affect the evaluation results, subsequent changes in the composition of the samples can create differences which seriously bias the study. Mechanisms are generally needed to minimize attrition and determine whether nonrespondents differ from respondents in ways which will bias the estimate of program effect. These mechanisms should be identified and justified in the evaluation plan. Procedures for obtaining needed replacement units should also be specified.

The ability of the evaluation to detect subtle program effects depends in part on the sample size used. Some social programs are not expected to produce massive changes, but their impact though small may be very important. In these cases, rather precise measurement is needed to detect the effects and if samples are not sufficiently large, these programs may be improperly judged.

Also, these program effects measured in a sample are only estimates of the program's effects in a total population. This is true because in addition to statistical variation in the impacts themselves, no group is totally homogeneous and, therefore, the sample can never be a perfect representation of the population from which it came. The confidence with which the estimates of program effect can be attributed to the complete program population depends primarily upon the size of the sample.

The evaluation plan should specify the sample size necessary to achieve the required level of confidence. It should also demonstrate that the sample size is sufficient to detect subtle program effects of a magnitude which can reasonably be expected.

5. Choosing measurement methods

Evaluations are partially determined by the measurement methods and instruments used; if those methods are not both valid and reliable, measurements will be distorted and bad decisions may result. To be given credibility, the evaluation plan must provide evidence of the validity and reliability of the measurement methods.

The evaluation plan should specify reasons for selection of a particular measurement technique and evidence that the technique is adequate to achieve the goals of the evaluation. The concepts of validity and reliability are related to the measurement purposes and the circumstances in which they were made, therefore, many forms of evidence may be appropriate depending on the situation.

Validity

Validity refers to the extent to which a measurement represents what it is supposed to represent. Program success criteria are rarely measured directly because they are usually only concepts. For example, a preschool training program may seek to increase the "self-esteem" of participating students, but "self-esteem" is not subject

to direct measurement. It must be measured by proxies such as attitude tests or with multiple tests. The question becomes how well the tests represent the concept.

It is the evaluator's responsibility to provide evidence to decisionmakers that a high degree of correlation exists between the chosen measurement and the concept. In reviewing this evidence, two important factors are (1) evidence of validity is rarely definitive and (2) a method or instrument is valid only in relation to its purpose.

A measurement method can rarely be absolutely validated simply because the concept which it attempts to measure cannot be clearly defined. Validation brings together qualitative and quantitative evidence that supports the interpretations and uses to be made of the measure. Although not definitive, the evidence must be sufficient to persuade the well-informed reviewer.

No measurement method or instrument is valid for all purposes, in all situations, or for all groups of people. It is, therefore, not enough to say that a measurement is valid because its developer conducted a test of validity with favorable results. It must be ultimately shown that the measurement is valid for the particular situation and purpose for which it is used.

Reliability

Reliability is defined here as the degree to which a measuring instrument yields the same result in repeated applications to the same phenomena; it refers to the stability or consistency of the instrument. Reliability addresses the random or chance variation in a measurement and the evidence is usually quantifiable.

The evaluator who can find highly reliable, off-the-shelf measures appropriate to the circumstances may be spared some work. The samples of people and circumstances of program administration against which the instrument was standardized, however, must be comparable to those of the planned evaluation or the evaluator must reassess reliability.

6. Timing of measurements

Many social changes do not manifest themselves quickly; other changes are more immediate but deteriorate over time. In such cases, the timing of measurements will affect the

evaluation results. The evaluation plan should include the justification for the frequency and timing of measurements.

Methods of estimating program effects, such as the time series, generally require measurements before, during, and after participation. Repeated measurements may also be needed during each phase to provide more persuasive evidence or to identify abrupt changes and cycles in the program's effect. On the other hand, frequent measurement may fatigue respondents, provoking their discomfort or displeasure, and defeat the attempt to obtain new or useful information.

It is difficult to choose the ideal frequency of measurement because the rate of improvement or deterioration is not known beforehand. The evaluator may, however, make plausible guesses about the rates of change and should have some justification for the timing of measurements.

7. Feasibility of the evaluation

The feasibility assessment should be accomplished before substantial funds are committed to the impact evaluation. This assessment cannot be determined in an absolute sense, but is relative to the decisionmakers' informational needs and evaluation state of the art. Feasibility assessment is, therefore, the process of tentatively matching informational needs with available time, resources, and data. This assessment can be done in several ways, ranging from identification and assessment of similar studies on similar programs, to peer reviews and small pilot tests.

Timing is often a problem in performing evaluations. The ideal time required for meaningful evaluation may vary significantly from one program to another, but, political issues can quickly become sensitive and informational needs may be immediate. Under such pressures, decisionmakers must and will act--legislation to start programs will be passed and funds will be authorized, appropriated, and spent. After such decisions have been made, the opportunity to make effective use of the evaluation will be less. One purpose of the feasibility assessment is to match the evaluation time requirements with the time available for decisions and establish realistic expectations of what the evaluation can accomplish. In many cases, partial information available at the time the decision must be made is better than no information or late information. However, it is important that such decisions be documented so that any evaluation assessment will be based on realistic expectations.

Also, sufficient resources may not be available to complete the evaluation, particularly if its goals are complex. Contract bids may not be a valid indication of the required resources. Competition for contracts is often keen, and there are resulting pressures to bid low when costs are a factor in determining which firm receives the contract. However, if contractors are required to specify in their proposals how the evaluation will provide for items in this checklist, contractual commitments will be clearer and realistic cost estimates can be obtained. In any event, the resources needed to complete the evaluation should be carefully estimated and compared to available funding so that the evaluation scope can be appropriately tailored.

The character of legal restrictions and the extent to which they can be waived or accommodated depends heavily on the nature of the program and the information needed for evaluation.

The Privacy Act of 1974 (5 U.S.C. 552a), the Family Educational Rights and Privacy Act of 1974 (20 U.S.C. 1232g), or other legislation may restrict the evaluator's ability to employ the most effective methods and to capitalize on available data. The Privacy Act, for example, may prevent the use of individually identifiable data in Federal agency archives without the consent of the individuals involved. Institutional review boards required by the Department of Health, Education, and Welfare may prohibit the collection of individually identified information if they believe it is too great a risk to participants in experimental and research programs or violates ethical principles over which the institutional review boards have jurisdiction. Other legislation may also inhibit evaluation. For example, in testing the impact of reduced training times for new recruits, the Army was unable to randomly assign people to the new program because soldiers destined for overseas assignments were given a minimum training period prescribed by the existing laws.

Legal, procedural, and statistical methods exist to overcome some of these restrictions, but, their use may significantly increase the cost of the evaluation or reduce the degree of certainty. The feasibility assessment should provide for identifying such restrictions and methods for overcoming them before the evaluation is begun.

Because evaluations yield information about the worth of a program and have potential for affecting the allocation of resources, they have explicit political implications. While political attention to evaluation results is generally viewed as desirable, pressures detrimental to the evaluation may be brought to bear. These pressures may manifest themselves in a number of ways such as attempts to influence site selection. Recent Federal and State legislative initiatives to require evaluation as a provision of "sunset" laws may increase both the potential for use of evaluation results and pressures which could be detrimental. The evaluator must attempt to understand and deal with the political forces at work before beginning the study.

8. Obtaining cooperation

The successful performance of an evaluation depends on obtaining the cooperation of those who are likely to be affected by it. This includes program administrators and staff, community organizations, and other interest groups. Such groups often have the capacity to frustrate or enhance the evaluation effort and the evaluator must search for common ground with them.

Friction between program administrators and evaluators is common in impact evaluations. Administrators usually have a strong commitment to programs under their control and may view evaluation as threatening or at best obtrusive and of little value in their administration. When the administrators sponsor the evaluation, they often do not have the experience to appreciate what is required if the evaluation is to be carried out properly, and policymakers do not always have the knowledge or authority to impose conditions necessary for rigorous evaluation. In addition, administrators may view the demands of data collection as an unnecessary burden. Refusals to permit random assignments to program and control groups, attempts to conceal data believed to show the program in a bad light, or failure to pay sufficient attention to important recordkeeping functions are not uncommon when program practitioners are not convinced of the value of the evaluation or its methods.

These potential conflicts cannot always be eliminated but they can often be reduced or accommodated when evaluators and program personnel sit down together before the evaluation takes place to discuss the goals and to plan study procedures. Administrators may find through such discussion that the

benefit of understanding the nature, magnitude, and relative costs of program effects justifies their support of the evaluation.

Many evaluations also require strong support from the community in which the program operates. For the most part their interest will be in the delivery of services, and they are likely to resist efforts which they believe threaten those services. Plans should include efforts to obtain their cooperation so that the evaluation will be no more difficult than necessary. Some communities such as the Woodlawn section of Chicago and Roxbury, Massachusetts, have organized groups which can approve or disapprove research in the area of their purview. Unless such groups are satisfied, it may be difficult or impossible to elicit information directly from individuals in the community.

Even when such organized groups do not exist, community support is important. The knowledge of key people in the community can be useful in dealing with a wide range of operational difficulties, and can reinforce the legitimacy of the research and enhance its realism. Because community attitudes are difficult to anticipate, good evaluation planning generally involves the use of fieldwork to identify the relevant groups, their interests, and likely sources of resistance to the evaluation.

B. DATA COLLECTION

No amount of planning will be sufficient unless controls are established and maintained during the operational phase to insure that the evaluation plan is properly implemented and that data collected are accurate and complete. Careless evaluations waste time and money and result in doubtful or misleading information.

Inaccurate or incomplete data can result from a number of causes including the data collection instruments, the data collectors, the program participants (data sources), or those administering the program. Effective procedures for quality control of data must be implemented and maintained to identify and minimize potential sources of bias. In addition, preliminary analyses should be performed as soon as possible to identify missing or inconsistent information and to correct deficiencies in the evaluation plan.

1. Procedures for quality control of data

Quality controls are needed to insure the integrity of raw data because errors, omissions, or misrepresentations at that point will bias the statistical analyses. Good intentions are not enough; structural controls are required, and the effectiveness of these controls should be carefully documented because they will help provide the credibility necessary to use the evaluation results.

Any quality control procedure must be specifically tailored to the particular circumstances and setting of the evaluation. However, some form of the following are considered good practice for most impact evaluations.

Pilot tests

Pilot testing of data collection instruments on a sample drawn from the program to be evaluated is generally advisable. Pilot testing is especially important when the data collection instruments are new and untested. The purpose of the test is to determine if the data collectors and program participants, or other respondents, understand the instrument and have the knowledge or information required to use it. During the test, specific item deficiencies such as lack of precision or ambiguity should be identified and corrected so that the data eventually collected will be interpretable.

Recruitment and training

The quality of an evaluation depends on the effort put into performing it. A team studying any complex policy or program should be composed of experienced persons from various disciplines, with the stature required to obtain the needed information and assure the credibility of the study. In complex studies, large numbers of employees must be recruited to collect, record, edit, file, and analyze data. They should be selected with care. During the recruiting process, qualifications and standards should be established and followed.

Once recruited, some employees must also be trained. The amount of training required will vary with the initial ability of the employees and the complexity of tasks, but usually both general and specific training will be required. Employees should be oriented to the evaluation goals and the policies and procedures to be employed in the study. They

must also learn the specifics of their assignment--what to do and how to do it. Periodic reinforcement of the training may also be necessary because of staff turnover or the need to remedy deficiencies brought to light by quality controls. These procedures should be documented.

Assignment procedures

A number of studies have shown that data collectors' expectations, style, and appearance can affect the data obtained from respondents and cause the collection of biased information. When the potential exists for bias emanating from these sources, procedures should be employed to eliminate or minimize it. Some techniques commonly used are blind assignment of observers or raters so that they do not know which people received the program and which did not, use of combined measurements obtained from more than one independent observer of the same event, and the random assignment or systematic rotation of data collectors. The state of the art for such techniques is advancing and should be exploited to reduce the risk that data will be contaminated.

Supervision

Adequate supervision is essential to identify those employees who are performing satisfactorily, those who need retraining, and those who must be replaced. One common procedure is for supervisory personnel to verify a portion or sample of the information. In the interview situation, for example, the supervisor may recontact a sample of the program participants to verify that (1) they were indeed interviewed, (2) certain questions were asked, and (3) they provided particular responses to key questions. Another form involves the supervisor occasionally accompanying data collectors to obtain more insight into how planned procedures are being carried out. The state of the art is advancing in this area also and should be exploited where possible.

Consistency checks or side studies are another important means of supervisory control. The employees' work may be screened for internal consistency or correspondence with available external data sources. These procedures usually require, however, that data collection instruments be carefully designed to include items with a high degree of correlation or consistency or items corresponding to information in external data sources.

Data editing

Data editing is essential to insure that inadvertent errors do not become a part of the data base. Formal, well documented checks are especially important when the data set is large and automatic data processing equipment is used in the study. Typical procedures include range checks to assure that all transcribed observations fall within predetermined plausible bounds, and internal consistency checks to assure that responses are consistent or reasonable as well as checks on keypunching and typing.

Monitoring program processes

The evaluator does not normally control the delivery of program services and in many cases does not control the collection of data to be used in assessing the program's impact. In such cases, the evaluator must take action to identify sources of systematic bias and attempt to have it eliminated. This usually requires the evaluator to monitor the program processes and data collection activities to verify that the program is being delivered as planned and that complete and accurate records are being maintained. When program recipients are to be compared with a control or comparison group, the evaluator must also verify that conditions necessary for valid comparison are maintained throughout the study.

2. Preliminary analyses

Preliminary analyses should be performed during the early stages of data collection, particularly if the evaluation effort is large and the issues are complex. Such early analyses can help to identify missing or incomplete information, inconsistencies, or other discrepancies in the data, and will help to guide changes required in the study methods.

Characteristics of nonrespondents

Nonresponse or attrition is a special problem in impact evaluations. Inevitably, some participants will leave the study while others will remain, but provide incomplete information. Nonresponse can often be minimized with properly supervised followup procedures, and the attrition can sometimes be accommodated by increasing the sample size. If, however, subjects who leave the study have some distinguishing characteristics and particularly if these differ between the program participants and the control or comparison group, the results of the evaluation may be seriously compromised.

Procedures must be established to carefully study the character of attrition and, if possible, to determine its cause and the likely effect on study results.

Testing consistency of data

Successful evaluation planning is rarely a one-shot affair, and no data collection process is perfect. Inevitably, deficiencies will occur in the implementation of the study, but will not surface until the evaluator begins to manipulate the data. In some instances, whole pieces of information may be missing and additional questions may have to be answered before reliable conclusions can be reached. The analyses may reveal that data gathered from different sources conflict and the conflicts will have to be resolved.

Preliminary analyses should be undertaken as soon as possible so that necessary action can be taken in time to obtain the data, resolve conflicts, or modify the study plan.

C. DATA ANALYSIS

Data analysis is the tabulation, organization, and summarization of the raw information collected during the evaluation. The analysis should relate data to the basic questions of the study, shape it into some digestable form, and determine through use of an appropriate statistical model if the indicated findings are significant or due to random variation.

1. Statistical methods and model

Based on the evaluation goals, plans should have been made to test specific hypotheses or answer specific questions, and to collect the data needed. Fitting the data to a statistical model directs the analysts' attention to certain aspects of the data and suggests inferences which may be drawn from it. This model should be made explicit, and its use should be justified by its relationship to the evaluation goals, design, and the nature of the data collected. The analysis may be seriously misleading if (1) these relationships are not recognized in the choice of statistical methods and model, (2) the level of data aggregation is not appropriate, or (3) the conditions necessary for employing the statistical model are not reasonably met.

Searching through data haphazardly is likely to be unproductive or misleading. This does not mean that the analyst cannot search through the data for evidence with which to test additional hypotheses, but, great care should be used in interpreting such evidence if the study is not designed to answer these questions. If the analyst investigates the data hard enough, refines variables, regroups units, and narrows the scope of the data included in the analysis, findings will probably emerge, but they may not all be genuine. The primary question and one which is not easily answered as the study nears completion is whether each finding is significant or the result of chance variation in the data.

The statistical model must be related to the evaluation design. For example, when people have been randomly assigned to program and control groups, simple comparisons of average values pertaining to these groups and the use of appropriate statistical tests may be sufficient to state with high confidence that the program did or did not have an effect. If random assignment is not employed, however, or for some reason not successful, more care in drawing conclusions would be required. This is because the evaluator cannot assume that people not randomly selected into program and control groups are on the average equivalent or comparable, in either their preprogram conditions or the rate at which they change in response to the program. If the program participants are less capable initially than the control group, the evaluation could be biased unfairly toward an unfavorable conclusion; if they are more capable initially, the evaluation could be biased in the other direction. When random assignment is not used, careful and sophisticated analysis and very careful judgments must be combined to arrive at a relatively unbiased conclusion. This is one of the most critical aspects of most data analysis because randomization is difficult to maintain even under the best of conditions.

Appropriate application of the statistical model must also be based on consideration of the data actually collected. For example, results of an evaluation may differ significantly as a function of which individuals or subgroups are included in the final analysis. Thus, an analysis which included only those who successfully completed the program may differ notably from one which also includes those who initially began the program but later withdrew for some reason.

In applying the model, the analyst must also consider the particular meaning assigned to variables and characteristics during the data collection. Depending on the questions or hypotheses being evaluated, the same variable could have a different meaning or status in different studies. Variables are classified as measures of program impact or one of several categories which help explain the cause of the impact. In social program evaluations, variables which help explain the impact are (1) population defining variables such as age or sex, (2) exposure variables which are the program inputs, or (3) covariates which are measures of variation before exposure to the program. Confusing the meaning of these variables in an analysis can cause very misleading results and conclusions. The use to which the variables are put is especially important in interpreting program effects when other than random selection procedures are used. For example, differences in test scores before exposure to a remedial education program (a covariate) could be used to properly interpret differences in later scores. If such a preprogram difference were ignored, the evaluation could be completely misleading.

2. Unit of analysis

Unit of analysis refers to the level of aggregation at which comparisons will be made, i.e., individuals or groups of individuals. There is no one unit of analysis proper for all evaluations. The level of aggregation during analysis should be based on the specific design and conditions of the evaluation. The essential criteria include how units were initially selected into the samples, whether groups differed initially in significant ways, and whether their experiences during the evaluation period differed significantly. If people were selected into the program or control groups on an individual basis then the individual is probably the proper unit of analysis. If, however, they were selected in groups, e.g., classroom by classroom, and the groups initially differed in significant ways or the group experiences differed significantly during the evaluation period, then the groups should probably be maintained as the unit of analysis. Such criteria is vitally important in the selection of appropriate statistical methods and models for the impact evaluation but the data may be regrouped and several units of analysis used for more speculative or exploratory auxiliary analyses.

3. Assumptions essential to statistical methods and models

All statistical models involve assumptions, e.g., that subjects were randomly assigned to treatment or control groups. Violation of some of these assumptions to a greater or lesser extent is to be expected. Some violations, however, may seriously bias the evaluation results. For this reason, the model assumptions should be made explicit and the evaluator should carefully document the extent to which the assumptions are satisfied. The analyst should also indicate to the extent possible, the probable effect of violations on the analyses.

D. REPORTING RESULTS

For most people, statistical analyses will become useful only with an interpretation provided by the evaluator. This interpretation is the oral or written reports of the evaluation. Careful attention is required to assure that the reports, especially the final one, are not misleading. Many users will judge the entire study effort on the basis of the final report and if it is not clear and complete or does not fairly reflect all of the findings and limitations of the evaluation, the information may be ignored or misused.

1. Clear, complete, and fair reporting

There is no foolproof way of course to assure that evaluation results are not misused. However, full and open disclosure can help to assure that findings are not misunderstood. The report itself should be sufficiently clear to its audience in its description of purposes, procedures, and findings about what was done, why it was done, and what was learned.

Evaluators along with study sponsors often have considerable latitude in highlighting what they believe to be important in the analyses and ignoring other aspects. Subtle pressures may be applied to avoid or bury unpopular facts and conclusions. The report should discuss all significant findings of the study, even those which may be unpopular with special interests. Moreover, all points of equal significance should be given equal emphasis in the report.

Human judgments or assumptions are involved in the conduct of an evaluation. As a result, evaluation findings are never totally free of the leanings and biases of people sponsoring

or conducting the evaluation. Clearly stating the assumptions and reasons for making them will enable decisionmakers to better deal with their inherent subjectivity.

The report must be equally clear and forthright about the limitations of the evaluation. Even with good planning, things will go wrong during the course of the study. These things are often attributable to factors such as funding or field conditions outside the control of the evaluator. If these limitations have the potential for altering study results in any significant way, they should be identified in the report and, if possible, their potential impact described.

In addition, all evaluation findings have limits beyond which reasonable inference cannot be made. While the evaluator could not possibly list all conditions or situations to which study results are not applicable, the findings should be sufficiently qualified to help readers avoid drawing improper inferences. Evaluators should not expect users to accept on faith that the study results are valid and complete. The report should contain enough information about the evaluation's scope and procedures for the users to understand what was done and why.

During the process of analyzing and reporting results, evaluators often recognize factors which should have been examined but were not, thus indicating the need for further evaluation. The report should identify and explain those issues and questions which need further study and consideration.

2. Assuring report quality

Procedural controls are available to help ensure the quality of the final report. These controls include independent verification of the facts, findings, conclusions, and recommendations contained in the report; careful review by those responsible for the report; and advanced review by program administrators and other officials responsible for the program being evaluated.

The independent verification of the accuracy of all facts contained in the report is an important quality control procedure. If carried out by someone not directly associated with the study, the verification procedure also helps determine whether the raw data supports the findings, conclusions, or recommendations of the report and will enhance its credibility.

Also, reports should be carefully reviewed by top officials responsible for the evaluation. Particular attention should be given to:

--Reasonableness and appropriateness of the findings and recommendations.

--Clarity of presentation.

--Potential adverse reactions and possible ways to handle such reactions.

Another effective way to insure that reports are fair, complete, and objective is to obtain comments on a draft of the report from program administrators and other officials of the organization responsible for the program which was evaluated. It is useful to the recipients of the report to know what these people think about the evaluation results and what actions they may be planning to take based on the evaluation.

3. Followup

Usually, some decisionmakers will need assistance in (1) interpreting the report, (2) clarifying aspects of it, (3) getting answers to questions raised by it but not answered, and (4) developing a reasoned reaction to it. Briefings, and supplementary written materials, may be needed to help the decisionmakers in understanding and using the study.

E. DATA DISCLOSURE

In some cases, information provided in the evaluation report will be sufficient to assure the use of evaluation results. In others, particularly, when evaluation findings are controversial, additional assurances in the form of audit or reanalysis of the evaluations, will be necessary before decisionmakers are sufficiently confident of the results to use them. To facilitate these activities, the evaluator must be careful to document important information and data. Such data may also be useful in research work or other evaluations and it should be maintained in an easily accessible form.

1. Documentation

Deciding what information to document and preserve is not easy. In large and complex evaluations, paperwork tasks can be staggering and comprehensive documentation is not without cost.

Insufficient documentation, however, will greatly limit any subsequent use of the data such as for the audit or reanalysis of the evaluation. As a general rule, the purposes and sources of the data, methods and circumstances of collection, and any suspected corruption of the data items should be documented.

Making data available also requires that it be in an accessible and understandable form. No standardized format or content will be sufficient for all evaluations. The data should be logically organized and the physical characteristics of the data set described in sufficient detail to permit analysts who did not participate in the original evaluation to use it.

2. Procedures for data release

As a general rule, outside analysts should have access to any data which feeds into the public decisionmaking process. On occasion, however, evaluators must collect sensitive information or information which could be embarrassing or damaging if made public in an identifiable form. In order to obtain a high response rate and candid answers, the evaluator may have to offer confidentiality pledges.

These pledges must be honored; provisions must be made to protect the privacy of individuals who provide information for the benefit of society. Evaluations are too important to the common welfare to risk eroding public trust in the activities which produce them.

Quite often statistical records will be sufficient for use in audit, reanalysis, research, or other evaluations. The release of records which contain no individually identifiable information will not jeopardize individual rights or violate the confidentiality pledges and they should be made available as soon as possible after the evaluation is completed. Procedures for the release of the data, including its location and officials authorized to release it, should be contained in the evaluation report.

In some cases, however, it will be necessary for auditors or analysts to access individually identifiable data. In those cases, Privacy Act requirements must be considered and a proper balance found between the individual's right to privacy and the public's right to know.

The purpose of most impact evaluations is to influence public policy, therefore, the public has a right to know that the evaluations were conducted properly and that results are valid and reliable. Audit and reanalysis provide such an assurance because they serve as a check on methodological and statistical procedures and provide an independent assessment of the credibility of the evaluation findings. An independent review of evaluations may also be necessary to assure that requirements of particular policy-making bodies are met. For example, the Congress may not always be willing to rely on analyses provided by the Executive since those analyses might be perceived as biased, or at best, emphasizing only aspects which the Executive might wish to highlight.

Auditors and evaluators should agree on procedures for assuring that there is no intent to violate the confidentiality pledges. Even though the auditor or analyst will not use research information to make determinations about individuals, conflicts may still arise if the evaluator has made unqualified assurances of confidentiality. No person should be persuaded to provide information under assurances of confidentiality which cannot be maintained.

To minimize the problems, confidentiality pledges should be made only when they are necessary to obtain sufficient and reliable information. When it appears that a pledge of confidentiality might conflict with the statutory rights and responsibilities of GAO or others, evaluation sponsors should consult with those parties before the pledges are made.

BIBLIOGRAPHY

1. Adams, S. Evaluative research in corrections: A practical guide. U.S. Department of Justice, Law Enforcement Assistance Administration, Washington, D.C., 1975.

2. Babbie, E. R. Survey research methods. Wadsworth, Belmont, CA, 1973.

3. Bailar, B. A. & Lanphier, C. M. Development of survey methods to assess survey practices. American Statistical Association, Washington, D.C., 1978.

4. Bennett, C. A. & Lumsdaine, A. A., eds. Evaluation and experimentation. Academic Press, New York, 1975.

5. Bernstein, I. N. & Freeman, H. E. Academic and entrepreneurial research, Sage, New York, 1975. p. 187.

6. Boruch, R. F. On appropriateness and feasibility of randomized tests of social programs. In L. Sechrest, ed. Evaluating emergency medical services. Government Printing Office, Washington, D.C., 1977.

7. Boruch, R. F. & Gomez, H. Sensitivity, bias, and theory in impact evaluations. Professional Psychology, 1977. pp. 411-443.

8. Borus, M. E., ed. Evaluating the impact of manpower programs. D. C. Heath, Lexington, MA, 1972.

9. Breger, M. J. Legal aspects of social research. Paper presented at Symposium on Ethical Issues in Social Science Research, University of Minnesota, MN, Apr. 1976.

10. Bryant, F. B. & Wortman, P. M. Secondary analysis: The case for data archives. American Psychologist, 1978, in press.

11. Bryk, A. S. & Weisberg, H. I. The implications of non-random assignment in comparative studies involving growth systems. Proceedings of the American Statistical Association, Social Statistics Section. ASA Washington, D.C., 1975.

12. Bryk, A. S. & Weisberg, H. I. Use of nonequivalent control group design when subjects are growing. Psychological Bulletin. 1977, 84, pp. 950-962.

13. Bunker, J. P., Barnes, B. A., & Mosteller, F., eds. Costs, risks, and benefits of surgery. Oxford University Press, New York, 1977.

14. Campbell, D. T. & Boruch, R. F. Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine, Evaluation and experiment. Academic Press, New York, 1975.

15. Campbell, D. T. & Stanley, J. C. Experimental and quasi-experimental designs for research. Rand McNally, Chicago, IL, 1966.

16. Caplan, N. Social research and national policy: What gets used, by whom, for what purpose, and with what effects? In S. Nagel, ed. Policy Studies Review Annual, 1, Sage, Beverly Hills, CA, 1977.

17. Caro, F. G., ed. Readings in evaluation research. Sage, New York, 1977.

18. Carter, L. F. Federal clearance of educational evaluation instruments: Procedural problems and proposed remedies. Educational Researcher, 1977, 6(6), pp. 7-12.

19. Cleary, T. A., Linn, R. L., & Walster, G. W. The effect of reliability and validity on power of statistical tests. In E. F. Borgatta and G. W. Bohrnstedt, ed. Sociological Methodology. Jossey-Bass, Inc., San Francisco, CA, 1970.

20. Cline, M. G. The "what" without the "why", or evaluation without policy relevance. In C. C. Abt., ed. The evaluation of social programs. Sage, Beverly Hills, CA, 1976, pp. 367-374.

21. Cochran, W. G. Sampling techniques. Wiley, New York, 1963.

22. Cochran, W. G. Some effects of measurement error on multiple correlation. Journal of the American Statistical Association, 1970, 65, pp. 22-35.

23. Cohen, J. Statistical power analysis for the behavioral sciences. Academic Press, New York, 1969.

24. Cook, T. D., et al. Seasame Street revisited. Sage, New York, 1975.

25. Cook, T. D. & Gruder, C. L. Metaevaluation Research. Evaluation Quarterly. 1978, 2(1), pp. 5-51.

26. Coulson, J. E. National evaluation of the Emergency School Aid Act (ESAA): A review of methodological issues. Journal of Educational Statistics, 1978, 3(1), pp. 1-60.

27. Cronback, L. J., Rogosa, D. R., Floden, R. E., & Price, G. G. Analysis of covariance in nonrandomized experiments: Parameters affecting bias. Occasional Paper, Stanford Evaluation Consortium. Stanford University, Stanford, CA, Aug. 1977.

28. Datta, L. Does it work when it has been tried? And half full or half empty? Journal of Career Education, 1976, 2(3), pp. 38-55.

29. Datta, L. The impact of the Westinghouse/Ohio evaluation on the development of Project Head Start. In C. C. Abt., ed. The evaluation of social programs. Sage, Beverly Hills, CA, 1976, pp. 129-191.

30. Dillman, D. A. Mail and telephone surveys. Wiley, New York, 1978.

31. Dunn, E. S. Social information processing and statistical systems - Change and reform. Wiley, New York, 1974.

32. Erickson, E. P. Some lessons learned from conducting Federally sponsored surveys. Proceedings of the American Statistical Association: Social Statistics Section (Part 1), ASA, Washington, D.C., 1977, pp. 173-182.

33. Fairweather, G. W. & Tornatsky, L. G. Experimental methods for social policy research. Pergamon, New York, 1977.

34. Field, C. G. & Orr, L. L. Organizations for social experimentation. In R. F. Boruch and H. W. Riechen, eds. Experimental testing of public policy. Westview, Boulder, CO, 1975, pp. 68-99.

35. Gilbert, J. P., Light, R. J., & Mosteller, F. Assessing social innovations: An empirical base for policy. In C. A. Bennett & A. A. Lumsdaine, eds. Evaluation and experiment. Academic Press, New York, 1975.

36. Gilbert, J. P., Mosteller, F., & Tukey, J. W. Steady social progress requires quantitative evaluation to be searching. In C. C. Abt., ed. The evaluation of social programs. Sage, Beverly Hills, CA, 1976, pp. 295-312.

37. Gilbert, J. P., McPeek, B., & Mosteller, F. Progress in surgery and anesthesia: Benefits and risks of innovative therapy. In J. P. Bunker, B. A. Barnes, F. Mosteller, eds. Costs, risks, and benefits of surgery, Oxford University Press, New York, 1977, pp. 124-169.

38. Gilbert, J. P., McPeek, B., & Mosteller, F. Statistics and ethics in surgery and anesthesia. Science 1977, 198, pp. 684-698.

39. Glaser, D. Routinizing evaluation: Getting feedback on effectiveness of crime and delinquency programs. National Institute of Mental Health, Rockville, MD, 1973.

40. Glass, G. V. Primary, secondary, and meta-analysis of research. Educational Researcher, 1976, 5(10) pp. 3-8.

41. Hardin, E. On the choice of control groups. In M. E. Borus, ed. Evaluating the impact of manpower programs. D. C. Heath, Lexington, MA, 1972, pp. 41-58.

42. Hedrick, T. E., Boruch, R. F., & Ross, J. Policy and regulation for ensuring the availability of evaluative data for secondary analysis. Evanston, Illinois: Psychology Department, Northwestern University, Policy Sciences, 1978, 9, pp. 259-280.

43. Hyman, H. H. Secondary analysis of survey samples. Wiley, New York, 1972.

44. Jabine, T. B. & Pigman, N.M. Some lessons learned from SSA experience in contracting for surveys. Proceedings of the American Statistical Association: Social Statistics Section (Part 1). ASA, Washington, D.C., 1977, pp. 173-182.

45. Kenny, D. A. A quasi-experimental approach to testing treatment effects in the nonequivalent control group design. Psychological Bulletin, 1975, 82(3), pp. 345-362.

46. Kiesler, S. B. & Turner, S., eds. Fundamental research and the process of education: Final report of the Committee on Fundamental Research Relevant to Education. National Academy of Sciences, Washington, D.C., 1977.

47. Kirusek, T. J. & Lund, S. H. Process and outcome measurement using goal attainment scaling. In G. V. Glass, ed. Evaluation Studies Review Annual, Sage, Beverly Hills, CA, 1976, 1, pp. 383-398.

48. Kish, L. Survey sampling. Wiley, New York, 1965.

49. Mattick, H. W. & Caplan, N. S. The Chicago Youth Development project. Institute for Social Research, Ann Arbor, MI, 1964.

50. McKay, H., Sinisterra, L., McKay, A., Gomez, H., & Lloreda, P. Cognitive growth in Colombian malnourished preschoolers, Science. 1978, 200, pp. 270-278.

51. McLaughlin, M. W. Evaluation and reform: The elementary and secondary education act of 1965/Title 1. Ballinger, Cambridge, MA, 1975.

52. Moynihan, D. P. & Mosteller, F. On equality of educational opportunity. Vintage, New York, 1972.

53. Nay, J. N., Scanlon, J. W., Schmidt, R. E., & Wholey, J. If you don't care where you get to, then it doesn't matter which way you go. In C. C. Abt., ed. The evaluation of social programs. Sage, Beverly Hills, CA, 1976, pp. 91-96.

54. Nejelski, P. Social research in conflict with law and ethics. Ballinger, Cambridge, MA, 1976.

55. Porter, A. C. & Chibucos, T. R. Selecting analysis strategies. In G. D. Borich, ed. Evaluating education programs and products. Educational Technology Publications, Englewood Cliffs, NJ, 1974, pp. 415-464.

56. Privacy Protection Study Commission. Personal privacy in an information society. U.S. Government Printing Office, Washington, D.C., 1977.

57. Riecken, H. W., Boruch, R. F., Campbell, D. T., Caplan, N., Glennan, T. K., Pratt, J. W., Rees, A., & Williams, W. Social experimentation: A method for planning and evaluating social intervention. Academic Press, New York, 1974.

58. Rivlin, A. Systematic thinking for social action. Brookings, Washington, D.C., 1971.

59. Rivlin, A. & Timpane, M. Ethical and legal issues in social experimentation. Brookings, Washington, D.C., 1975.

60. Rivlin, A. M., ed. Protecting individual privacy in evaluation research. National Academy of Science, Washington, D.C., 1975.

61. Rossi, P. H. & Lyall, K. C. Reforming public welfare. Sage, New York, 1976.

62. Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 1974, 66(5), pp. 688-701.

63. Rubin, D. B. Formalizing subjective notions about the effects on nonrespondents in sample surveys. Journal of American Statistical Association, 1977, 72, pp. 538-543.

64. Rubin, D. B. Assignment to treatment group on the basis of a covariate. Journal of Educational Statistics, 1977, 2(1), pp. 1-26.

65. Rubin, D. B. Bayesian inference for causal effects: The role of randomization. Annals of Statistics, 1978, 6(1), pp. 34-58.

66. Rutman, L., ed. Evaluation research methods. Sage, Beverly Hills, CA, 1977.

67. Schoenfeldt, L. F. Data archives as resources for instruction, research, and policy planning. American Psychologist, 1970, 25, pp. 609-616.

68. Scriven, M. Maximizing the power of causal investigations: The modus operandi method. In G. V. Glass, ed. Evaluation Studies Review Annual, Sage, Beverly Hills, CA, 1976, 1, pp. 75-99.

69. Sechrest, L., ed. Emergency medical services. U.S. Government Printing Office, Washington, D.C., 1978.

70. Sewell, W. H., Hauser, R. M., & Featherman, D. L., eds. Schooling and achievement in American society. Academic Press, New York, 1976.

71. Stanford Evaluation Consortium. Review essay: Evaluating the "Handbook of evaluation research." In G. V. Glass, ed. Evaluation Studies Review Annual. Sage, Beverly Hill, CA, 1976, 1, pp. 195-217.

72. Stanford Program on Teaching Effectiveness. A factorially designed experiment teacher structuring, soliciting, and reacting. Stanford Center for Research and Development in Teaching, Stanford, CA, 1976.

73. Struening, E. & Guttentag, M., Eds. Handbook of evaluation research. Sage, Beverly Hills, CA, 1975, 1, pp. 289-354.

74. Suchman, E. A. Evaluation research. Sage, New York, 1967.

75. Sudman, S. & Bradburn, N. M. Response effects in surveys: A review and synthesis. Aldine, Chicago, IL, 1974.

76. Thistlethwaite, D. L. & Campbell, D. T. Regression-discontinuity analysis: An alternative to the ex post facto experiment. Journal of Educational Psychology, 1960, 51(6), pp. 309-317.

77. Tiku, M. L. Tables of power of the F test. Journal of the American Statistical Association, 1967, pp. 525-539.

78. Tuchfarber, A. J. & Klecka, W. R. Random digit dialing. Police Foundation, Washington, D.C., 1976.

79. Timpane, M. Evaluating Title 1 again? In C. C. Abt., ed. The evaluation of social programs. Sage, Beverly Hills, CA, 1976, pp. 415-424.

80. U.S. Census Bureau. Indexes to survey methodology literature: Technical paper #34. Social and Economic Statistics Division, U.S. Department of Commerce, Washington, D.C., 1974.

81. U.S. General Accounting Office. Standards for audit of government organizations, programs, activities, and functions. U.S. GAO, Washington, D.C., 1972.

82. U.S. General Accounting Office. Evaluation and analysis to support decisionmaking. U.S. GAO, Washington, D.C., Sept. 1, 1976.

83. U.S. General Accounting Office. Federal program evaluations. 1977 Congressional Sourcebook Series, PAD-78-27, U.S. Government Printing Office, Washington, D.C., 1978.

84. U.S. General Accounting Office, Program Analysis Division. Background paper for use by the SSRC Committee on Audit and Research on the need for access by GAO auditors in the audit of social research and experiments. U.S. GAO, Washington, D.C., Apr. 8, 1977.

85. U.S. General Accounting Office. The Concord--Results of a supersonic aircraft's entry into the United States. CED-77-131, U.S. GAO, Washington, D.C., Sept. 15, 1977.

86. U.S. General Accounting Office. An assessment of the Department of Housing and Urban Development's Experimental Housing Allowance Program. CED-78-29, U.S. GAO, Washington, D.C., Mar. 8, 1978.

87. U.S. General Accounting Office. Finding out how programs are working: Suggestions for Congressional oversight. PAD-78-3, U.S. GAO, Washington, D.C., Nov. 22, 1977.

88. Van Horn, C. D. & Van Meter, D. S. The implementation of intergovernmental policy. In S. S. Nagel, ed. Policy Studies Review Annual, Sage, New York, 1977, 1, pp. 97-120.

89. Watts, H. & Rees, A., eds. The New Jersey income maintenance experiment. Academic Press, New York, 1977.

90. Weikart, D. P. & Banet, D. P. Planned variation from the perspective of a model sponsor. Policy Analysis, Brookings, Washington, D.C., 1975, 1(3).

91. Weiss, C. H. Evaluation research: Methods of assessing program effectiveness. Prentice-Hall, Englewood Cliffs, NJ, 1972.

92. Weiss, C. H. Evaluation research in a political context. In E. L. Struening and M. Guttentag, eds. Handbook of evaluation research. Sage, Beverly Hills, CA, 1976, 1.

93. Zeisel, H. Reducing the hazards of human experiments through modification in research design. Annals of the New York Academy of Sciences, 1970, 169, pp. 475-486.

CROSS REFERENCE: BIBLIOGRAPHY TO CHECKLIST 1/

REF. ITEMS	CHECKLIST ITEM																
	A-1	-2	-3	-4	-5	-6	-7	-8	B-1	-2	C-1	-2	-3	D-1	-2	E-1	-2
1			X														
2				X					X								
3				X					X	X							
4			X														
5																	
6							X						X				
7				X			X										
8			X						X								
9							X										
10																	X
11													X				
12													X				
13				X													
14			X								X	X	X				
15			X	X					X		X	X	X				
16																	
17			X	X	X		X						X				
18																	
19											X		X				
20																	
21				X									X				
22												X	X				
23				X													
24	X																X
25																X	X
26							X										
27				X							X	X	X				
28	X						X										
29	X						X										
30				X					X	X							
31																	X
32				X													
33			X						X		X	X	X				
34							X										
35			X								X	X	X				
36				X													
37							X										
38																	
39			X						X								
40																	X
41			X														
42																X	X
43																X	X
44			X														
45			X														
46																	
47	X						X										
48			X						X		X	X	X				
49																	
50	X																

REF. ITEMS	CHECKLIST ITEM																
	A-1	-2	-3	-4	-5	-6	-7	-8	B-1	-2	C-1	-2	-3	D-1	-2	E-1	-2
51																	
52																	X
53	X						X										
54							X										X
55			X								X		X				
56							X		X								X
57	X		X	X	X		X	X	X								X
58																	
59							X										
60								X									
61	X																X
62			X								X		X				
63											X		X				
64				X													
65																	
66	X	X					X										X
67																	
68																	
69			X														
70																	
71																	
72																	
73			X	X													
74	X	X		X			X										
75					X												
76				X							X		X				
77				X													
78				X													
79	X																
80					X												
81																X	
82	X	X			X			X	X	X	X		X	X		X	
83																	
84																	X
85																	
86																	X
87		X			X		X				X			X			
88																	
89	X		X														
90																	
91				X				X									
92																	
93							X										

1/ Indicates places where further discussion on an item can be found but other sources listed may also touch on the item.

(97451)