

# Handbook of Research Design & Social Measurement

## EVALUATION RESEARCH AS A PROCESS

Contributors: Delbert C. MillerNeil J. Salkind

Print Pub. Date: 2002

Print ISBN: 9780761920458

Online ISBN: 9781412984386

DOI: 10.4135/9781412984386

Print pages: 79-100

This PDF has been generated from SAGE Research Methods Online. Please note that the pagination of the online version will vary from the pagination of the print book.

# EVALUATION RESEARCH AS A PROCESS

Every attempt to reduce or eliminate a social problem involves a theory, a program, and usually a large amount of money. The effectiveness of programs to reduce crime and delinquency, combat drug addiction, conquer health problems, and improve neighborhoods and communities and the quality of life generally—all pose problems of evaluation. Because these problems are so important to national and community life and are so costly, evaluation has been given a high priority and evaluation research is increasing.

Edward Suchman (1967) has written:

Evaluation always starts with some value, either explicit or implicit—for example, it is good to live a long time; then a goal is formulated derived from this value. The selection of goals is usually preceded by or concurrent with “*value formation*.” An example of “*goal-setting*” would be the statement that fewer people should develop coronary disease, or that not so many people should die from cancer. Goal-setting forces are always in competition with each other for money, resources, and effort.

There next has to be some way of “*measuring goal attainment*.” If we set as our goal that fewer people should die from cancer, then we need some means of discovering how many are presently dying from cancer (for example, vital statistics). The nature of the evaluation will depend largely on the type of measure we have available to determine the attainment of our objective. The next step in the process is the identification of some kind of “*goal-attaining activity*.” In the case of cancer, for example, a program of cancer-detection activities aimed at early detection and treatment might be considered. Then the goal-attaining activity is put into operation. Diagnostic centers are set up and people urged to come in for check-ups.

Then, at some point, we have the *assessment* of this goal-directed operation. This stage includes the evaluation of the degree to which the operating program has achieved the predetermined objectives. As stated previously, this assessment may be scientifically done or it may not.

Finally, on the basis of the assessment, a *judgment* is made as to whether the goal-directed activity was worthwhile. This brings us back to value formation. Someone now may say that it is “good” to have cancer diagnostic centers. At the end of the evaluation process, we may get a new value, or we may reaffirm, reassess, or redefine an old value. For example, if the old value was “it is good to live a long time,” the new value might be, “it is good to live until 100 if you remain healthy; but if you can't remain healthy it's better not to live past eighty.”

### 3.7.1 Design and Implementation of Evaluation Models

One way in which many social and behavioral science disciplines has changed is an increased call for accountability. One of the ways this has been expressed is through the design and implementation of evaluation models and strategies. In the following excerpt, Rossi and Wright discuss the role of evaluation in the assessment of programmatic outcomes and how evaluation has developed as a scientific approach to the answer of whether a program is effective.

## Evaluation Research: An Assessment

Evaluation research came into prominence as an applied social scientific activity during the Great Society programs of the mid-1960s. The distinctive feature of the past 25 years is the explicit recognition among policymakers and public administrators that evaluations could be conducted systematically using social scientific research methods and could produce results that had more use and validity than the judgmental

approaches used previously. During the Great Society era, Congress authorized many new programs and systematic evaluations were mandated in several of the more important pieces of legislation.<sup>1</sup>

The new administrative agencies set up to implement many of these programs were partially staffed by social scientists who had strong interests in applied work. The entire gamut of the social scientific disciplines was involved. Economists had a strong foothold in the Office of Economic Opportunity; sociologists, psychologists, and educators were ensconced in the Office of Education (later the Department of Education); the Department of Health, Education and Welfare (now Health and Human Services) was big enough to accommodate members of all of the social scientific disciplines in critical positions; and the Department of Labor's Manpower Research Division was also generous, providing opportunities for all.

The interdisciplinary character of this new social scientific activity was especially noteworthy. Economists, sociologists, psychologists, and educational researchers often found themselves bidding on the same contracts in competition with each other, a process that facilitated the transfer of knowledge, craft lore, and mutual respect across disciplinary boundaries. Research firms and institutes previously dominated by one discipline broadened their outlooks by hiring professionals from other social sciences, mainly in order to increase their competitive edge. Interdisciplinary professional societies were also founded, e.g., the Evaluation Research Society and the Evaluation Network.<sup>2</sup>

University-based social scientific researchers were slow to take advantage of the new opportunities for research funding, even though the topics involved were often of central interest, a reflection of the indifference (even hostility) to applied work that has characterized the academic social science departments until very recently (Raizen & Rossi 1981, Rossi & Wright 1983, Rossi et al. 1978). Private entrepreneurs, however, were quicker to notice and exploit the new emphasis on evaluation. Some existing firms that had not been particularly interested in the social sciences opened subsidiaries that could compete for social research contracts (e.g., Westing-house). Others greatly expanded their social science research sections (e.g., the Rand Corporation). In

addition, literally hundreds of new firms appeared on the scene, a handful of which became spectacular successes during the “golden years” (e.g., Abt Associates).<sup>3</sup>

By the middle of the 1970s, some 500-600 private firms existed primarily to bid on contracts for applied social research. As in other areas of corporate activity, a few firms garnered the majority of the available funds. For example, in the period 1975-1980, 6 large research firms received over 60% of the evaluation funds expended by the Department of Education (Raizen & Rossi 1981).

An additional large number of firms sprang up to bid on contracts for evaluation and other applied social research activities at the state and local levels. These research opportunities were neither as well funded as those on the federal level nor were the tasks as intellectually or technically challenging. There was (and continues to be) enough evaluation “business” on the state and local levels, however, to provide the essential “bread and butter” for a very large number of small-scale job shops.

Some of the existing university-based research institutes with histories of large-scale social research also prospered during this period. The National Opinion Research Center at the University of Chicago and the Survey Research Center at the University of Michigan both grew enormously in size. Their staffs eventually came to dwarf most academic departments in the relevant fields. New academic research organizations also were started to take advantage of the funding opportunities offered through the grant and contract mechanism.

A corresponding growth took place on the conceptual side of evaluation research. The publication in 1966 of Donald T. Campbell and Julian Stanley's seminal work on research designs useful in the evaluation of educational programs created an entirely new vocabulary for the taxonomy of research designs and for the discussion of validity issues. It also made the randomized, controlled experimental paradigm the method of choice for causal analyses. Both of these emphases came to dominate large portions of the evaluation field for the next decade. Evaluation research was initially seen as, quintessentially, the assessment of programs' net effects. Correspondingly, the main problem in designing evaluation research was to specify appropriate *ceteris paribus* conditions that would permit valid estimates of these net effects. Within this framework, the randomized, controlled experiment became the ruling paradigm for evaluation

research. The conceptual foundations had been developed many decades earlier, and this approach had been the ruling research paradigm in both psychology and biology for many years. The special contribution made during the period under current review was that the paradigm was taken out of the laboratory and into the field, and it was combined with the sample survey in studies designed to test the effects of the proposed programs. To many social scientists of a technocratic bent, the randomized field experiment promised to replace our bumbling trial-and-error approaches to forging social policy with a more self-consciously rational “experimenting society” (Campbell 1969).

By the early 1970s, an impressive number of large-scale experiments had been funded and started. These experiments covered a wide variety of topics: income maintenance plans intended to replace the existing welfare benefits system; housing allowances that might stimulate the market to produce better housing for the poor; health insurance plans that would not create perverse medical-care price effects; and so on through a veritable laundry list of field experiments. Ironically, most of them were designed and run by economists, members of a field not noted for its tradition of experimental work.

The realization quickly emerged, however, that randomized, controlled experiments could only be done correctly under very limited circumstances and that the demand for evaluation covered many programs that simply could not be assessed in this way. Not only were there frequent ethical and legal limitations to randomization, but many existing programs that had full (or almost full) coverage of their intended beneficiary populations could not be assessed using controlled experiments because there was no way to create appropriate control groups. It also turned out that field experiments took a longtime—3 to 5 years or more—from design to final report, a delay that was simply intolerable given the much shorter time horizons of most policymakers and public administrators.

Campbell and Stanley (1966) had provided one possible solution to this dilemma by coining the term *quasi-experiments* and using it to cover evaluation research designs that do not rely on randomization to form controls. Although they explicitly recognized the inferior validity of data generated in this way, they also discussed the conditions under which valid causal inferences could be drawn from evaluation studies using such designs. Their treatment of quasi-experimental research designs certainly

stimulated the use of such designs in evaluation studies, sometimes under conditions that Campbell and Stanley explicitly stated were potentially fatal. Indeed, the vast majority of the evaluations that have been carried out have been quasi-experiments, rather than randomized, “true” experiments, mainly because the latter have proven difficult, if not impossible, to implement in real world settings.

But even quasi-experimental designs have their limitations. For one thing, while not as expensive or time-consuming as “true” experiments, a well-conducted quasi-experiment may demand more funds, time and talent than are available. Another problem is that many of the more sophisticated quasi-experimental designs (in particular, interrupted time series designs) require long time series of data—ideally, series that contain a long run of observations prior to the introduction of a policy intervention and that continue for several years after that. Concerning the first, the necessary data often do not exist; and, concerning the second, the old problem of timeliness reappears. A final problem, of course—one Campbell and Stanley discussed in detail—is that there are potential threats to the validity of any quasi-experimental design. In using such designs, one always runs some risk of mistaking various artifacts for true program effects. Hence, quasi-experiments are almost always vulnerable to critical attack; witness the rancorous controversies surrounding some of the major educational evaluations (e.g., McLaughlin 1975, Mosteller & Moynihan 1972, Rossi & Wright 1982).

Due to the many evident problems of both experimental and quasi-experimental approaches to evaluation research, the need for methods of evaluation that were timely, relatively inexpensive, and responsive to many program administrators’ and officials’ fears that evaluations would somehow “do them in” quickly became apparent. This statement applies especially to evaluations that were mandated by Congress and that the program agencies themselves were supposed to conduct. Indeed, Congress—coupling its newfound enthusiasm for evaluations with a seriously flawed understanding of the time, talent and funding needed to carry out evaluations of even minimum quality—often imposed evaluation tasks on program agencies that far exceeded the agencies’ research capacities and then provided funds that were grossly inadequate to accomplish them.

The need for evaluations that could be carried out by technically unsophisticated persons and that would be timely and useful to program administrators fueled a strong

interest in qualitative approaches to evaluation research (Patton 1980, Scriven 1977, Guba & Lincoln 1981, House 1980). Qualitative research methods have always had some following in all of the social sciences, especially in sociology. Their special attraction in sociology is their presumed ability to stay close to reality and to promote an understanding of social processes through intimate familiarity with field conditions. In addition, for evaluation purposes, qualitative methods seemed to have the attractive triple advantages of being inexpensive, timely, and responsive to administrators' needs. These approaches were especially attractive to program sponsors and operators because they appeared to be flexible enough to cope with social programs that, once implemented, tend to vary sharply from one locale to another not only in their goals but also in the benefits and services that are actually delivered. The goals for some broad-spectrum programs (e.g., Model Cities) were not clearly defined by Congress or the administering agencies. Each operating agency thus defined its own goals and often changed them frequently (Kaplan 1973, Williams 1980). The appeal, at least initially, of qualitative approaches to evaluation is that they apparently had the potential to be sensitive to the nuances of ill-defined and constantly evolving program goals.

The great boom in evaluation ended in 1981 when the Reagan administration began to dismantle the social programs that had been developed over the previous 20 years. The extensive manpower research program of the Department of Labor was reduced to almost nothing and there were similar (although less drastic) cuts in the Departments of Health and Human Services, Education, and Agriculture, among others. The immediate consequence was a drastic reduction in the amount of federal money available for applied social research.

Ironically, the Reagan cutbacks occurred just as more and more academic departments began to discover that there was a nonacademic market for newly minted PhDs. Openings for evaluation researchers were a large component of this market. The American Sociological Association held an extremely well-attended conference in Washington, D.C., in 1981 (Freeman et al. 1983) on the appropriate training for careers in applied sociology. Many graduate departments throughout the country began programs to train applied researchers of all kinds, and there was an evident interest among at least some prominent sociologists. Indeed, both presidents of the American Sociological Association in 1980 and 1981 devoted their presidential addresses to applied work (Rossi 1981, Whyte 1982).



# The Intellectual Harvest of the Golden Years of Evaluation

The frenzied growth of evaluation research during the 1960s and 1970s produced a real increment in our knowledge about the relevant social problems and a decided increase in the technical sophistication of research in the social sciences. Both of these developments have already had some impact on the social sciences and will be increasingly valuable to our fields in the future.

## The Large-Scale Field Experiments of the “Golden Age”

Perhaps the most impressive substantive and technical achievements of the entire Golden Age were those of the large-scale field experiments. Most of these experiments were initially funded by the Office of Economic Opportunity and, upon the demise of that agency, by the Department of Health, Education and Welfare.

On the technical side, these experiments combined both sample survey techniques and classical experimental designs. Experimental and control groups were created by sampling open communities and then randomly allocating sampled households to experimental and control groups. Interviews with experimental and control households were then undertaken, using traditional sample survey techniques to measure responses to the experimental treatments. Looked upon as surveys, these experiments were long-term panels with repeated measurements of the major dependent (i.e., outcome) variables. Measures were taken as often as once a month in some of the experiments and extended over periods of up to five years. Viewed as experiments, the studies were factorial ones in which important parameters of the treatments were systematically varied.

Perhaps the best-known of the field experiments during the Golden Age were those designed to test various forms of the “negative income tax” (NIT) as a means of

maintaining a reasonable income floor for poor households. All told, there were five such experiments in the United States and one in Canada.

## Fixing Up Nonexperimental Designs

The discussion so far has been fairly narrowly focused on randomized experimental designs for impact assessment because (a) the technically most successful impact assessments were carried out using that design and therefore (b) the randomized, controlled experimental paradigm has dominated the evaluation scene for the last two decades. As detailed above, however, there are good reasons at least to modify the experimental paradigm, chief among them being that for most social programs evaluation must perforce use nonexperimental methods.

There are many reasons why randomized experimental designs cannot be used in some evaluation studies. First, ongoing programs that cover most or all of their intended target populations simply do not admit of believable controls. For example, an estimated 5-10% of the persons eligible for Old Age and Survivors Insurance (Social Security) benefits have not applied for them. These nonapplicants cannot realistically serve as controls for estimating the effects of social security benefits, however, because the self-selection factors are undoubtedly strong. Comparing persons receiving social security benefits with those who are eligible but, for whatever reasons, have not applied for them violates the *ceteris paribus* condition.

Second, some programs, such as Head Start, fail to reach significantly large proportions of the eligible population—perhaps as much as 25% of the Head Start example. These children are not reached by the Head Start program because parents have not allowed their children to enroll or because the school systems involved have too few poor children to support Head Start projects. Clearly, strong self-selection factors are at work, and hence, contrasting Head Start participants with eligible nonparticipants would not hold constant important differences between the two groups.

Finally, it would be ethically unthinkable to use randomization in the evaluation of some programs. For example, a definitive way of estimating the relative effectiveness of private and public high schools would be to assign adolescents to one or the other

randomly and observe the outcome over an extended period of time. Obviously, there is no way that either policymakers or parents would allow such an evaluation to take place.

Thus, many of the evaluation studies of the past two decades have employed something other than classical randomized experimental designs. Unfortunately, these evaluations have not been technically successful on the whole. Each of the major nonexperimental evaluations has been shrouded in controversy—controversy that arises out of the political implications of the findings but that often centers on the technical inadequacies of the designs employed. Thus, Coleman's (1966) attempt to sort out schools' effects on achievement by analyzing a cross-sectional survey of thousands of students from hundreds of high schools was criticized mainly because of the statistical models he used (Mosteller & Moynihan 1972). Similarly, an evaluation (Westinghouse Learning Corporation 1969) of the long-lasting effects of participating in Head Start came under fire (Campbell & Erlebacher 1970) because the researchers compared youngsters who had attended Head Start preschools with "comparable" children who had not. According to the study's critics, confounding self-selection factors were undoubtedly at work that made the two groups incomparable in important respects.

The problem of administrative or self-selection of program participants and non-participants is at the heart of nonexperimental evaluation designs' vulnerability to criticism. To illustrate this point, we can consider Coleman and his associates' (1982) recent study of academic achievement in public and private (mostly Catholic) high schools. The critical comparisons in such a study are clearly plagued by self-selection factors: whether a child attends the Catholic parochial high schools or the public high schools cannot by any stretch of the imagination be considered a random choice. Parents often make the choice alone, although they sometimes consult the child; they make their education decisions on the basis of factors such as their anticipated income, their commitment to their religious group and its ideology, their assessments of their child's intellectual capabilities, the relative reputations of the local high schools, and so on. Nor are parents and child the only forces involved. Parochial high schools exercise judgment about whom they want to admit, selecting students on the basis of factors like their previous educational experience, the kind of curriculum the child or parents want, and the child's reputation as a behavioral problem. Some of these

factors are probably related to high school achievement; the extent to which these factors independently affect such achievement would confound any *simple* comparisons between the achievement scores of parochial and public high school students.

Obviously, one way out of the problem is to hold constant statistically those factors relating both to achievement and to school choice. The difficulties of doing so, however, are also obvious. First, it is necessary to specify the relevant factors correctly, a task that is usually difficult because of the absence of empirically grounded theory to aid in that specification. Secondly, if the element of choice is one of those factors (as in this example), it *cannot be* held constant since choice exists for one group but not for the other; in the present case, that is, non-Catholics would not have the option of sending their children to parochial schools. [See Rossi & Wright (1982) for a more detailed critique of Coleman along these lines.]

A potentially fruitful solution to this problem has recently been suggested by the econometricians (Goldberger 1980, Barnow et al. 1980, Berk & Ray 1982). They propose that researchers construct explicit models of the decision process and incorporate these models into structural equation systems as a means of holding constant the self-selection process. Although these proposals are somewhat more attractive than the usual approach of adding independent variables to a regression equation, they are still largely irrelevant because the appropriate decision models cannot be constructed except in special circumstances.

Another important development in the methodology used in nonexperimental evaluations has been the application of time series models to the assessment of the net effects of large-scale programs. [These models were originally developed in economic forecasting (Pindyck & Rubinfeld 1976) and subsequently applied specifically to evaluation problems (McCleary & Hay 1980, Cook & Campbell 1979).] First suggested by Campbell & Stanley (1966) as “interrupted time series” designs, the application of time series models has made it possible to assess the impact of new large-scale programs or the effects of modifying existing ones without recourse to classical randomized experiments. This approach is limited to programs that have long time series of data on their outcomes available and whose onset can be definitely located in time as, for example, with the enactment of new legislation.

Among the best-known interrupted time series evaluations are the various assessments of the Massachusetts Bartley-Fox gun law (G. L. Pierce & Bowers 1979, Deutsch & Alt 1977, Hay & McCleary 1979). This law imposed a mandatory penalty for carrying guns without a license, with the objective of reducing the use of guns in crimes. Using time series models, the researchers modeled the trends in gun-related crimes before the Bartley-Fox law went into effect and compared the resulting projections with the trends observed after the law was enacted. The findings suggest that the law led to only a slight reduction in the use of guns in crimes. The times series models used (Box-Jenkins models) are composed of a family of frameworks, each differing from the others in its assumptions about the kinds of time-dependent processes at work. To some degree, the choice among models is a judgment call, a condition that has led to polemical exchanges among independent researchers about the law's true effects (e.g., Hay & McCleary 1979, Deutsch 1979).

The two developments just discussed have implications for sociology that go considerably beyond evaluation research per se. The conceptualization of the self-selection problem in evaluation research has direct applications to most sociological research that relies on cross-sectional studies. The data analysis problems encountered are identical, so solutions developed in the evaluation field have immediate applications in the many sociological studies in which self-selection issues complicate the interpretation of findings.

Time series of critical data are available on many of the substantive areas of interest to sociologists. Aggregate data on crime rates go back almost 50 years; unemployment rates have been available on a monthly basis for almost 40 years; and so on.

Research on program implementation is primarily research in public administration. Although good examples are rare, in principle it is no more difficult to test several alternative ways of delivering a program than to test several alternative programs; indeed, the two problems are formally identical. That implementation issues are often critical is widely recognized (Williams & Elmore 1976, Pressman & Wildavsky 1973, W. S. Pierce 1981), but the importance of research on the issues involved has not received the attention it deserves.

# The Future of Evaluation Research: An Addendum by Delbert C. Miller

Since the above assessment of evaluation research was written by Rossi and Wright in 1984, evaluation research has suffered a decline in fiscal support. The decline began to be evident as early as 1980, in the political climate imposed by the Reagan administration. Personnel of the U.S. General Accounting Office recently reported the following.

2. Between 1980 and 1984, the number of professional staff in all agency evaluation units decreased by 22%, from about 1,500 to about 1,200. In contrast, the total number of staff in these agencies decreased by only 6% during this period.
4. Between 1980 and 1984, funds for program evaluation were reduced by 37%, compared with a 4% increase for the agencies as a whole.
6. Information loss and distortion of findings were reported as the result of lack in assessment. These failures were shown to be most serious in the areas of defense, the environment, and labor and personnel (Chelimsky et al. 1989).

The future of evaluation research, in spite of recent declines, is promising.

Reducing the federal deficit and promoting public confidence in the federal government are two top concerns the incoming Congress and administration must face. Crucial to both is the availability of timely, technically sound information for legislative oversight, for program management, and for public awareness. Information for the first audience—Congress—answers questions about how money is being spent and managed, and what results have been achieved. Information for the second audience—program managers—answers questions about what needs to be done to comply with the law and to achieve greatest effectiveness and efficiency of operations. Information for the third audience—the public—answers questions about what it is getting for its money.

Program evaluation is an essential tool in providing information to all three audiences. (Chelimsky et al. 1989, p. 25)

These needs will not go away. If anything, as old social demands increase in severity and new social needs arise, and as budgets rise by the multibillions of dollars, evaluation research becomes ever more important to Congress, to program managers, and to the public. And the need for evaluation is not limited to the federal government; it is equally important for state and city governments.

Evaluation appears to have become part of the tools of government. Private research agencies will continue to receive important contracts for program assessment. Therefore, there will probably be a continuing need for personnel well trained in the social sciences to staff the research projects that will be undertaken, and sociologists may continue to find employment in evaluation research.

The accompanying list of the literature cited by Rossi and Wright is an outstanding compilation of both evaluation methodology and evaluation studies of social programs.

## Notes

1. Especially important were the evaluations mandated in the 1964 Elementary and Secondary School Education Act (McLaughlin 1975), in the Housing and Urban Development budget authorization of 1970 calling for the experimental evaluation of a proposed housing allowance program (Struyk & Bendick 1981), and in the enabling legislation for the Department of Labor's Comprehensive Employment Training Program (Rossi et al. 1980). Evaluation research is found today in all major fields of social intervention, including health, mental health, criminal justice, housing, and handicapped children and their families. The Department of Defense has used evaluation research increasingly.

2. A tabulation of the primary disciplines of the members of the Evaluation Research Society (Evaluation Research Society 1979) nicely illustrates the interdisciplinary character of the evaluation research field. Herewith, the breakdown of membership by

field: psychology 47%; sociology 10%; economics 4%; political science 6%; education 15%; and other 18%.

3. Some of the spectacular successes of those prosperous times, of course, have been greatly diminished by the reverses of today's harder times. At its height, Abt Associates employed more PhDs in the social sciences than any one of the Boston area universities and more than most combinations of universities. In the past few years, its PhD workforce has been reduced by almost 50%.

SOURCE: From Peter H. Rossi and James D. Wright (1984), "Evaluation Research: An Assessment," *Annual Review of Sociology*, 10, 332-352. Reprinted by permission of the Annual Review,

## Literature Cited<sup>#</sup>

*Human Action*

### 3.7.2 Defining Evaluation and a Comparison of Internal and External Evaluation

Evaluation, according to Sonnichsen (2000), is the process of collecting and analyzing evidence, then disseminating the findings to identified audiences so that policy and programmatic judgments and decisions can be made.

Traditionally, most evaluations have been external in nature, performed by an evaluator selected from outside the organization, but this has changed. One of the major advantages of internal evaluation, in today's environment of tighter budget constraints, is the lower costs associated usually with internal evaluations. Several other factors also come into play.



Sonnichsen (2000) presents the following comparison of internal and external evaluation, including the advantages and disadvantages.

*Internal Evaluations* *External Evaluations*

- Commitment to the organization
  - Knowledge of the organization's personnel and operations
  - Quick response to evaluation requests
  - Greater perceived credibility by organizational components
  - Function as an institutional memory
  - Frequent client contact
  - Support of the decision-making process
  - Access to data
  - Lower costs
  - Greater ability to observe the organization's operations
  - Flexibility to assume other than evaluation tasks on short notice
  - Greater ability to monitor recommendation implementation
  - Potential to educate organization regarding value of evaluation
  - Serve as change agents
  - Continuity of evaluation effort
- 
- May possess superior evaluation skills
  - Perceived as more independent
  - Bring fresh perspective to organizational issues
  - Greater objectivity
  - Less susceptible to co-optation
  - Can objectively assess organization-wide programs that may include the internal evaluators as participants
- 
- Possible lack of power in the organization
  - Possible lack of independence
  - Ethical dilemmas
  - Burden of additional tasks
  - Perceived organizational bias
  - May lack technical evaluation expertise
- 
- Lack knowledge of the organization

- Limited access to organizational data
- More expensive

### 3.7.3 The Sequence of Evaluation

Regardless of what it is that evaluators evaluate (such as programs, policies, organizations, products, or individuals), and be the evaluation internal or external, the sequence of steps tends to remain the same. Sonnichsen (2000) outlines a series of steps (see table on previous page) that should take place in any internal evaluation that, for the most part, can be applied to external evaluations as well.

This “methodical sequence of the significant components required for an internal evaluation” almost guarantees a thorough and complete set of outcomes. The following sequence, easily modified depending upon one’s particular problem at hand, is an excellent starting point.

## Significant Components of an Internal Evaluation

2. Study initiation
  - Opening memorandum. Contains sufficient information to convey, to interested stakeholders and the evaluation staff, the commencement of the evaluation
  - Authority
  - Purpose
  - Deadlines
  - Team members
  - Administrative requirements
  - Computerized control log entries
  - Meetings with affected executives and managers
  - Notification of department heads and appropriate stakeholders
  - Contacts with subject matter experts

4. Literature search
  - Review both organizational policies and paperwork and the substantive literature
6. Formulation phase
  - Develop issues and frame evaluation questions
8. Evaluation plan
  - Prepare workplan
  - Design evaluation
  - Prepare design matrix
10. Data collection and analysis
  - Determine quality and availability of data
  - Ascertain precision needed in evaluative data
  - Resolve quantitative and qualitative data collection strategies
  - Decide on random or purposeful sampling methodologies
  - Plan and match appropriate statistical and analytical tools to anticipated data
12. Communicating evaluation results
  - Determine if report will be written
  - Prepare appropriate briefings
  - Resolve dissemination procedures for evaluation findings
14. Write recommendations
  - Identify options if appropriate
16. Closing procedures
  - Workpaper preparation
  - Retention and disposition of workpapers
  - Report annotation
  - Classified material handling (if necessary)
  - Control log updating
18. Follow-up
  - After six months, determine the status of suggested changes, approved recommendations, and attempt to measure impact of the evaluation

SOURCE: From Richard Sonnichsen (2000), *High Impact Internal Evaluation: A Practitioner's Guide to Evaluating and Consulting Inside Organizations* (Thousand Oaks, CA: Sage). Reprinted with permission.

## 3.7.4 The Concept of Forms in the Evaluation Process

John Owen and Patricia Rogers (1999) have created what they refer to as a meta-model of evaluation that is based on five evaluation forms, each with a “defining orientation and a focus on a set of common issues, which provide guidance for the planning and conduct of investigations.”

These five categories or forms are

- Proactive
- Clarificative
- Interactive
- Monitoring
- Impact

As you will shortly see, they can all be compared across a variety of criteria including the purpose of the evaluation, the issues the different forms address, the approach taken during the evaluation, the major focus of the evaluation, and others. Table 3.1 organizes these five forms of evaluation as a function of these criteria.

## 3.7.5 Selected References on Evaluation Research

Sage Publications continues a very active publication effort in the field of evaluation, with coverage of theory, method, and utilization. The student or researcher interested in operational aspects of evaluation research should first examine volumes in the Program Evaluation Kit and then continue with the readings listed in this section for other selected examples of evaluation research focused on specific problems. The “general references” are directed to the student who seeks a fuller understanding of theory, method, and research advances.

Interest in evaluation research is exploding, both in scope and in publication. One good starting point is the ERIC Clearinghouse on Assessment and Evaluation, which can be found at [eric.ed.gov](#). Here, you can find entire books, journal articles, and other resources on assessment, evaluation, and associated research topics.

## 3.7.5.1 Program Evaluation Kit (Sage Publications)

This kit, first published in 1978 with Joan L. Herman as series editor, provides information on many different techniques for evaluating many different types of programs.

- Volume 1—*valuator's Handbook* (1988) by Joan L. Herman, Lynn Lyons Morris, and Carol Taylor Fitz-Gibbon. This first volume is the heart of the Program Evaluation Kit and provides a broad overview of evaluation planning and a practical guide to designing and managing programs.
- Volume 2—*How to Focus an Evaluation* (1988) by Brian Stecher and W Alan Davis. This volume provides a broad overview of evaluation planning and a practical guide to designing and managing programs.
- Volume 3—*How to Design a Program Evaluation* (1988) by Carol Taylor Fitz-Gibbon and Lynn Lyons Morris. This volume reflects the tremendous explosion of interest in this vital area of the evaluation process and recognizes that deciding what to evaluate is a complex negotiation process that involves many different factors.
- Volume 4—*How to Use Qualitative Methods in Evaluation* (1988) by Michael Quinn Patton. Introduces the reader to qualitative approaches.
- Volume 5—*How to Assess Program Implementation* (1988) by Jean A. King, Lynn Lyons Morris, and Carol Taylor Fitz-Gibbon. Extensively revised to reflect modern views of program implementation, this volume introduces the variety of functions served by implementation studies and the roles played by qualitative and quantitative data.
- Volume 6—*How to Measure Attitudes* (1988) by Marlene E. Henerson, Lynn Lyons Morris, and Carol Taylor Fitz-Gibbon. An important part of any evaluation process, this book focuses on the assessment of attitudes.

- Volume 7—*How to Measure Performance and Use Tests* (1988) by Lynn Lyons Morris, Carol Taylor Fitz-Gibbon, and Elaine Lindheim. The evaluator's role in performance measurement is a critical element, and this volume focuses on ways an evaluator can select, develop, and analyze tests.
- Volume 8—*How to Analyze Data* (1988) by Carol Taylor Fitz-Gibbon and Lynn Lyons Morris. This is a basic introduction to a variety of elementary statistical techniques, including those for summarizing data, for examining differences between groups, and for examining relationships between two measures.
- Volume 9—*How to Communicate Evaluation Findings* (1988) by Lynn Lyons Morris, Carol Taylor Fitz-Gibbon, and Marie E. Freeman. This volume includes examples from a wide range of disciplines and shows the reader how to communicate results to users and stakeholders throughout the evaluation process.

**Table 3.1 EVALUATION FORMS: ORIENTATION, TYPICAL ISSUES, AND**

	<i>Proactive Evaluation</i>	<i>Clarificative Evaluation</i>	<i>Interactive Evaluation</i>
Orientation	Synthesis	Clarification	Improvement
Typical issues	<ul style="list-style-type: none"> <li>• Is there a need for the program?</li> <li>• What do we know about this problem that the program will address?</li> <li>• What is recognized as best practice in this area?</li> <li>• Have there been other attempts to find solutions to this problem?</li> <li>• What does the relevant research or conventional wisdom tell us about this problem?</li> <li>• What do we know about the problem that the program will address?</li> <li>• What could we find out from external sources to rejuvenate an existing policy or program?</li> </ul>	<ul style="list-style-type: none"> <li>• What are the intended outcomes, and how is the program designed to achieve them?</li> <li>• What is the underlying rationale for this program?</li> <li>• What program elements need to be modified to maximize the intended outcomes?</li> <li>• Is the program plausible?</li> <li>• Which aspects of this program are amenable to a subsequent monitoring or impact assessment?</li> </ul>	<ul style="list-style-type: none"> <li>• What is this program trying to achieve?</li> <li>• How is this service delivered?</li> <li>• Is the delivery cost-effective?</li> <li>• Is delivery consistent with the program plan?</li> <li>• How could delivery be changed to make it more effective?</li> <li>• How could this organization be structured so as to make it more effective?</li> </ul>

Key approaches	<ul style="list-style-type: none"> <li>• Needs assessment</li> <li>• Research review</li> <li>• Review of best practice (benchmarking)</li> </ul>	<ul style="list-style-type: none"> <li>• Evaluability assessment</li> <li>• Logic/theory development</li> <li>• Accreditation</li> </ul>	<ul style="list-style-type: none"> <li>• Responsive</li> <li>• Action research</li> <li>• Quality review</li> <li>• Developmental</li> <li>• Empowerment</li> </ul>
State of program	None	Development	Development
Major focus	Program context	All elements	Delivery
Timing [vis-à-vis program delivery]	Before	During	During
Assembly of evidence	Review of documents and databases, site visits, and other interactive methods. Focus groups, nominal groups, and Delphi technique useful for needs assessments.	Generally relies on combination of document analysis, interview, and observation. Findings include program plan and implications for organization. Can lead to improved morale.	Relies on intensive studies, including observation. Degree of data structure depends on approach. May involve providers and participants.



## 3.7.5.2 Other Sage Publications in Evaluation

The Evaluation Studies Review Annuals are published by Sage Publications. Information on them is available at [. The series includes the following volumes:](#)

Volume 1: edited by Gene V Glass (1976, 704 pages)

Volume 2: edited by Marcia Guttentag with Shalom Saar (1977, 736 pages)

Volume 3: edited by Thomas D. Cook and Associates (1978, 783 pages)

Volume 4: edited by Lee Sechrest, Stephen G. West, Melinda A. Phillips, Robin Redner, and William Yeaton (1979, 768 pages) Volume

5: edited by Ernst W Stromsdorfer and George Farkas (1980, 800 pages) Volume

6: edited by Howard E. Freeman and Marian A. Solomon (1981, 769 pages) Volume

7: edited by Ernest R. House and Associates (1982, 752 pages)

Volume 8: edited by Richard J. Light (1983, 672 pages)

Volume 9: edited by Ross F. Conner, David G. Altman, and Christine Jackson (1984, 752 pages)

Volume 10: edited by Linda H. Aiken and Barbara H. Kehrer (1985, 650 pages) Volume

11: edited by David S. Cordray and Mark W Lipsey (1986-1987, 757 pages) Volume

12: edited by William R. Shadish, Jr., and Charles S. Reichardt (1988, 704 pages)

### 3.7.5.3 Selected Examples of Evaluation Research

### 3.7.5.4 General References for Evaluation Studies

### 3.7.5.5 Specialized Evaluation Journals

The following journals are all published by Sage Publications. Information is available on them at .

*Evaluation: The International Journal of Theory, Research, and Practice*

*Evaluation and the Health Professions*

*Evaluation Review: A Journal of Applied Social Research*

The following evaluation journals are published as noted.

*Evaluation Practice* (JAI Press)

*New Directions for Evaluation* (Jossey-Bass)

*Evaluation and Program Planning*(Pergamon)

*Journal of Policy Analysis and Management* (John Wiley)

*Canadian Journal of Program Evaluation* (University of Calgary Press)

*Evaluation Journal of Australia* (Australian Evaluation Society)

*Educational Evaluation and Policy Analysis* (American Educational Research Association)

*Assessment and Evaluation in Higher Education* (Carfax Publishing Ltd.)

## 3.7.5.6 The Survey Kit

What follows are the references included in the Sage Survey Kit, a collection of books published by Sage Publications in 1995 that treats the essential topics related to using surveys to collect data.

Bourque, L., & Fielder, E. P. *How to conduct self-administered and mail surveys.*

Fink, A. *How to analyze survey data.*

Fink, A. *How to ask survey questions.*

Fink, A. *How to design surveys.*

Fink, A. *How to report on surveys.*

Fink, A. *How to sample in surveys.*

Fink, A. *The survey handbook.*

Fink, A. *The survey kit.*

Frey, J., & Oishi, S. M. *How to conduct interviews by telephone and in person.*

Litwin, M. *How to measure survey reliability and validity.*

### 3.7.5.7 Professional Organizations for Program Evaluators and Policy Analysts

American Evaluation Association at

Association for Public Policy Analysis and Management at

Canadian Evaluation Association at

Australian Evaluation Association at

European Evaluation Association at

United Kingdom Evaluation Association at

Italian Evaluation Association at

### 3.7.6 Selected Bibliography on Evaluation

*Evaluation and the health professions*  
*Evaluation review: A journal of applied social research*  
*Evaluation: The international journal of theory, research, and practice*