ED 302 972 EC 211 859

AUTHOR Johnson, Lawrence J.
TITLE Program Evaluation: The Key to Quality Programming.
INSTITUTION Council for Exceptional Children, Reston, Va.; ERIC Clearinghouse on Handicapped and Gifted Children, Reston, Va.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE 88
NOTE 31p.; In: Jordan, June, Ed. And Others; Early Childhood Special Education: Birth to Three; see EC 211 851.
PUB TYPE Information Analyses - ERIC Information Analysis Products (071) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Data Collection; *Delivery Systems; *Disabilities; Evaluation Criteria; Evaluation Methods; Models; Preschool Education; *Preschool Evaluation; Program Effectiveness; *Program Evaluation; Research Design; Standards
IDENTIFIERS *Early Intervention

ABSTRACT

Part of a volume which explores current issues in service delivery to infants and toddlers (ages birth to 3) with handicapping conditions, this chapter presents program evaluation as a comprehensive process comprising three phases: input, process, and output. Three program evaluation models are summarized: Tyler's objective model, Scriven's goal-free model, and Stufflebeam's decision-making model. The latter is seen as the basis of the triphase evaluation process. Steps in the input evaluation phase are described in detail: (1) determine key elements; (2) identify sources of information; (3) develop a management plan; (4) collect data; (5) analyze and interpret data; and (6) develop an intervention program. The purpose of the second phase, process evaluation, is to monitor progress toward goals and objectives and to modify plans as needed. Research design considerations are explored in a discussion of the outcome phase, emphasizing the importance of a well-conceived, systematically implemented evaluation in order to determine the impact of the program on children, their families, and the community. Finally, standards for a high quality evaluation plan developed by the Joint Committee on Standards for Educational Evaluation are described, focusing on the four elements of utility, feasibility, propriety, and accuracy. References are appended. (JW)

Need for evaluation has intensified

Legislatures must be provided with reliable and valid data.

□ With the passage of P.L. 99-457, services for handicapped infants and toddlers, ages birth to 3 have reached a critical crossroad. Within the next 5 years we are likely to see a dramatic increase in services to these children. However, much still needs to be done before mandated services become a reality. Although states can receive financial support for providing services for handicapped infants and toddlers under the age of 3, they will not be mandated to do so. As a result, the need for systematic evaluation of programs serving these children has intensified. It is likely that policy makers will raise many questions about programming for these children. They will ask what programming options are available and what are the merits and drawbacks of each. They will wonder what impact these programs have on children, their families, and the community. Undoubtedly, they will eventually ask if the cost of establishing and operating such programs is justified. It is up to us to make use of comprehensive evaluation plans that can provide the answers to these and other questions that are sure to be raised. Legislatures must be provided with reliable and valid data when they consider alternatives for providing services to children from birth to age 3.

Data can be beneficial.

Although providing valid information to policy makers is an important function of evaluation, it is not the only function. Data collected from good evaluation plans can be beneficial to early childhood special education programs at many different levels. From an interviewer's perspective, it can provide information by which to make instructional decisions, monitor child and family progress, and document accountability. From a parent's perspective, it can be used to examine child and family programs and as an indication of program effectiveness. Finally, policy makers can use evaluation data to make informed decisions about program management, using information about the costs, benefits, and drawbacks of various program alternatives.

Interviewers are sometimes resistant to evaluation efforts.

Unfortunately, the development and implementation of good evaluation plans is one aspect of early childhood special education that has not always been adequate (see Dunst & Rheingrover, 1981; Odom & Fewell, 1983; Simeonsson, Cooper, & Schiener, 1982; White & Casto, 1984; White, Mastropieri, & Casto, 1984; Wolery, 1987; Wolery & Bailey, 1984). Several factors contribute to this situation. Administrators often lack the knowledge or resources to carry out a comprehensive evaluation and may also fear what such an examination might reveal. Interveners are sometimes resistant to participating in evaluation efforts, and they see program evaluation as an extra burden. They may believe that evaluation efforts interfere with what they are doing, but have no particular benefits for the program or them. At the same time, however, interveners have always evaluated what they were doing. They identify child needs, make plans to meet those needs, and monitor child progress, although the rigor with which this is done varies.

There should be a strong link between programming and evaluation.

One problem lies in the mistaken belief that evaluation is separate from intervention and essentially involves the collection of a series of pre/post measures. In actuality, current thinking on evaluation suggests that there should be a strong link between programming and evaluation. This notion was eloquently presented by Bricker and Littman (1982) in their article, "Intervention and Evaluation: The Inseparable Mix." They argued that evaluation data should provide the basis for intervention and help determine the value of the intervention for groups of children. The viewpoint presented in this chapter is congruent with Bricker and Littman and others who have stressed the link between evaluation and intervention (Goodwin & Driscoll, 1980; Isaac & Michael, 1981; Wolery, 1987; Wolery & Bailey, 1984). The evaluation process presented here has three phases—input, process, and outcome—and is based on the evaluation models of Tyler, Scriven, and Stufflebeam. The phases are interwoven into a single process that begins with program planning, continues through implementation, and then turns its attention to program impact. For clarity and efficiency, this evaluation process will be referred to as *triphase evaluation*; however, this author does not claim that this process represents a new model. Rather, it is a common-sense approach to conducting a comprehensive program evaluation.

Evaluation has three phases.

Evaluation models that form the basis of the triphase evaluation process are presented here; the triphase evaluation process is described in detail and examples are provided; and finally, critical components of a high-quality evaluation plan are discussed.

EVALUATION MODELS

□ In this section, three evaluation models are summarized. They are but a small sample of the many models that have been proposed for program evaluation (see Morris & Fitz-Gibbon, 1978 for a more complete description of evaluation models), but they have made significant contributions to thinking about program evaluation, and they form the basis of the triphase evaluation process. Strengths and weaknesses of the models are highlighted to give the reader a sense of their contributions to the triphase evaluation process.

TYLER'S OBJECTIVE MODEL

□ The Tylerian model focuses on the delineation of objectives and measurement of progress on these objectives (Tyler, 1942, 1958, 1971, 1974). Simply stated, a set of objectives is identified, procedures to assess their attainment are established, data are collected, and judgments are made as to the success of the program based on child and/or family performance on the identified objectives.

There are several advantages to this model. Its simplicity makes it easy to understand and interpret. Its focus on measurable objectives encourages accountability and provides teachers with a means to demonstrate progress to parents and administrators. Finally, it includes the intervener as an integral member of the evaluation process and employs more than just pre/post measures.

Simplicity makes it easy to understand and interpret.

Ironically, the simplicity of the model and reliance on behavioral objectives are also cited as weaknesses. Linking evaluation so closely to objectives prevents actions not easily measured by objectives from being included in the evaluation process. Many of the most important educational outcomes are not amenable to behavioral statements. The simplification of such outcomes into objectives can trivialize them, or worse, prevent them from being included in the program. Finally, outcomes not tied to objectives are not examined. This is a serious flaw, because a program can have a dramatic positive or negative impact that is not directly related to a specific objective.

Outcomes not tied to objectives are not examined.

SCRIVEN'S GOAL-FREE MODEL

Interviewers are not directly involved.

The evaluator is placed in a discovery role.

Lack of structure can be a liaoility.

❑ Concerned with the potential biasing and limiting impact of linking the evaluation process so closely to objectives, Scriven (1967, 1973, 1974) proposed goal-free evaluation. Unlike the objective-based model, interveners are not directly involved in the evaluation process; instead, an outside evaluator with little knowledge of the program is employed. This evaluator does not need to know what the objectives are, but is concerned with identifying the actual impact of the program, intended or unintended. Scriven (1974) believes that knowing the goals of the program encourages the evaluator to look for alleged effects instead of actual effects. The evaluator's role is to discover the actual effects of the program, which may differ markedly from the program's stated goals.

A goal-free approach to evaluation has several advantages. First, the evaluator is placed in a discovery role and is not limited to determining whether or not goals were obtained. Second, the search for unintended effects is positive and prevents tunnel vision. Someone with a new perspective can notice things about the program that those within the program or those focusing on the objectives of the program have missed. Finally, because the evaluator is independent from the program, he or she is in a better position to evaluate it critically.

Despite these advantages, the lack of structure can be a liability in this approach. Without clear objectives, the evaluation has no standard against which the effectiveness of the program can be consistently applied. This process does not include interveners in evaluation and is conducted after the fact, rather than being integral to the program from the beginning.

STUFFLEBEAM'S DECISION-MAKING MODEL

There are four kinds of evaluation.

Comprehensiveness is one of its greatest strengths.

❑ In this model, evaluation is defined as a decision-making process involving three steps: (a) delineating the information to be collected, (b) obtaining the information, and (c) providing the information to decision makers (Stufflebeam, 1971, 1974). Information collected through this process can then be used by decision makers to judge the merit of options presented to them.

Stufflebeam has stated that there are four kinds of evaluation: context, input, process, and product. Within each of these kinds of evaluation are four types of decisions that can be made in an educational setting. In context evaluation, the decisions to be made relate to planning. The primary purpose is to identify needs of individuals to be served by the program and identify objectives to meet those needs. The decisions of concern in input evaluation relate to the structuring of programs to meet the needs of the individuals to be served. Primary areas for examination are issues related to such areas as program management, staffing, and budgeting. In process evaluation, decisions relate to implementation of the program. Data are collected to determine any flaws in the program as it is being implemented. In product evaluation, decisions relate to what Stufflebeam has termed *recycling*, which refers to decisions being made to continue, terminate, modify, or refocus the program.

The comprehensiveness of this model is one of its greatest strengths. The interrelationship between the four types of evaluation encourages a focus beyond just pre/post measures. This model presents evaluation as a continuous cycle that builds on information collected in the other types of evaluation. Finally, it provides a vehicle to establish accountability in implementing the program as well as judging the impact of the program. However, the comprehensiveness of the model makes it complex, difficult to coordinate, and expensive.

TRIPHASE EVALUATION

❑ The basis of the Triphase evaluation process is Stufflebeam's decision-making model. As with Stufflebeam's model, the Triphase process is comprehensive and concerns itself with all aspects of the program. However, interrelationship between the phases is stressed more than in Stufflebeam's model. In Stufflebeam's model, evaluation is presented as the coordination of types of evaluation context—input, process, and product—that are used depending on the decision to be made. Evaluation from the Triphase perspective is seen as one process made up of three interwoven phases: input, process, and outcome. During each of these phases the evaluation plan focuses on a different aspect of the program. In the input phase, attention is directed at determining child, family, and community needs and developing a program to meet them. In the process phase, attention is directed at monitoring progress toward objectives and determining whether or not there are any discrepancies between what was proposed and what is being implemented. These phases build on each other, with the input and process phases being the most critical to the implementation of a good program. The influence of Tyler can be seen in the emphasis on behavioral objectives. The development of objectives and the monitoring of progress toward objectives is the backbone of the model. However, recognizing the concerns of Scriven's goal-free evaluation, efforts are not limited to performance on objectives.

Evaluation is made up of three interwoven phases.

The input and process phases are considered part of formative evaluation, which is the collection of evaluation data to aid in program planning and implementation. The outcome phase is part of summative evaluation in that the purpose of data collection is to provide information on the impact of the program. Unfortunately, people often think of evaluation as being equivalent to summative evaluation and do not consider the importance of formative evaluation. During formative phases, when problems are detected, changes can be made to the original plan to avoid potential disaster. However, in the summative phases, by the time problems are detected it is too late, and we must wait until next time to correct mistakes or change project orientation. On the other hand, it is not enough to document the proper implementation of a project; we must also determine whether or not it has a meaningful impact on the children, their families, and the community. Clearly all three phases are critical to the evaluation plan and the program. In the following sections, each of these phases is discussed in greater detail.

Outcome is part of summative evaluation.

All three phases are critical.

INPUT EVALUATION

❑ The focus of the input phase is on assessing the needs of children and their families and developing a plan to meet those needs. An important

Recommendations can be made for revisions in the plan.

Duplication of services is common.

step in this phase is to examine services that currently exist and compare them to what is being proposed to meet identified needs. In other words, after needs are identified we must determine whether or not there are any discrepancies between what is, what ought to be, and what is being proposed. Based on information obtained in this step of the evaluation plan, recommendations can be made for revisions in the proposed plan to address any discrepancies that are uncovered. This phase of the evaluation plan is vital to the development of a high-quality program. If the needs of children and their families are not adequately identified, everything we do in an attempt to meet their needs will be flawed. Beyond this problem, it is equally important to ensure that the program has the resources to carry out the proposed plan and that the plan is not a duplication of already existing services. Duplication of services is particularly common with programs serving exceptional children ages birth to 3. Many different agencies serve these children and their families, and unfortunately the linkages between these programs are not always strong. As a result, valuable resources are wasted, possibly preventing needed services from being instituted.

The intervener contributes the link between assessment and programming.

From an intervener's viewpoint, input evaluation is a concern every time a new child and family enter the program. The intervener must assess child and family needs and then develop a plan to meet those needs. This information can be used at a program level to keep in touch with the needs of the broader community. Essentially, the intervener contributes to the evaluation plan by forging a strong link between assessment and programming.

Data are collected from several sources.

From a program perspective, input evaluation is particularly important when a new program or a new component of an existing program is being developed. One of the first steps in program development is to conduct a needs assessment. Borg and Gall (1983) defined a need as being a discrepancy between an existing set of conditions and a desired set of conditions. Using this definition, conducting a needs assessment becomes more than providing parents or teachers a brief questionnaire to gather their perceptions of what is needed. Rather, it is a comprehensive plan by which data are collected from several sources. The steps outlined below can help ensure the systematic collection of needs assessment data. They are equally useful in collecting outcome evaluation data. They will be discussed in detail here and will be referred to in the section on outcome evaluation.

Determine Key Elements

Develop a set of goals.

□ The first step in the process is to determine the purpose of the needs assessment and the clients and audiences for the needs assessment. A helpful technique is to develop a set of goals or questions to be addressed and then prioritize the goals to ensure that the critical data are collected. In this way some of the less important goals can be sacrificed if the process becomes unduly complex or resources dwindle.

As an example, let us suppose that a small school decides to expand its preschool program to meet the needs of handicapped children from birth to age 3. Recognizing the importance of a good input evaluation, the administrators would probably decide to conduct a needs assessment. They might identify the *clients* as the handicapped children and their families within the community and the *consumers* of the needs assessment as parents of handicapped children, program administrators, and interveners. The following prioritized questions are examples of a set that the small school might use to guide their needs and assessment:

1. How many handicapped children ages birth to 3 need services?
2. What are the characteristics of these children and their families?
3. Who is providing services to these children and their families?
4. What alternatives within this community could meet the needs of these children and their families?

Identify Sources of Information

□ The next step in conducting the needs assessment is to determine the sources of information from which to answer identified questions. In addition, a data collection method must be developed that will obtain the needed information efficiently and accurately. Usually we must collect information from a variety of sources and therefore need a variety of methods for collecting data. For example, in the sample questions presented in the previous section, no one data sources would be able to provide information to answer adequately all the questions generated. Therefore, we must use multiple sources of data to be sure that we collect all the information to determine child, family, and community needs. Typical data collection methods include unobtrusive measures, observation, interviews, questionnaires, and tests. These methods are equally useful in the outcome phase of the evaluation plan.

We need a variety of methods for collecting data.

Unobtrusive Measures. These sources are classified as nonreactive because children and their families are not required to change their daily routine and are, for the most part, not aware of the data gathering. As Casto (in press) pointed out, unobtrusive measures have been used infrequently as an evaluation tool by programs serving handicapped infants and toddlers but could provide valuable, inexpensive information. For example, if we were interested in determining parent concerns we might examine the books checked out of a parent-resource library or the toys checked out of a toy-lending library.

Unobtrusive measures could provide valuable information.

Another important source of information can be the records and documents of agencies that might come into contact with children from birth to age 3 and their families. For example, as Casto (in press) noted, many of the children who eventually receive services in programs for handicapped toddlers and infants are graduates of neonatal intensive care units (NICUs). Fortunately, most of these units have computerized data bases and routinely collect extensive information on NICU patients. This information can be useful in locating children, determining numbers of potential clients, and providing critical family information. Much of this information, however, is confidential, so releases need to be obtained. When such releases are not feasible, a protocol can be developed with someone in the agency who can summarize the information of interest across clients, without violating individuals' rights of privacy.

Observations. Observations to collect needs assessment information are generally made by interveners at a programming level to determine performance of children and families in relation to specific objectives. The essence of behavioral observations is the systematic recording of operationally defined behaviors. When operational definitions are properly done, ambiguity is reduced to a minimum. Definitions should be based

Definitions should be based on observable characteristics of behavior.

on observable characteristics of the behavior, clearly stated, with variations of the behavior defined so that rules can be established for their scoring. Alberto and Troutman (198 delineated several dimensions of behavior that can be recorded, which depend upon the type of behavior targeted and the circumstances of the evaluation. For example:

1. *Frequency*: A count of how often the behavior occurs.
   Example: Susan had nine tantrums this week.
2. *Rate*: Frequency data expressed in a ratio with time.
   Example: On the average, Susan has six tantrums per week.
3. *Duration*: A measure of how long the behavior lasts.
   Example: Susan's last tantrum lasted 40 minutes.
4. *Latency*: A measure of how long it takes before a new behavior is started.
   Example: It took 20 minutes for Susan to stop her tantrum when she was removed from the other children.
5. *Topography*: A description of what the behavior looks like.
   Example: Susan shrieks, kicks her heels, and throws herself on the floor when she has a tantrum.
6. *Force*: A description of the intensity of the behavior.
   Example: Susan cries so hard during a tantrum that her veins stick out of her neck and her face turns bright red.
7. *Locus*: A description of where the behavior occurs.
   Example: Susan seems to have her tantrums in the bathroom or the hall.

The dimension of behavior recorded depends on the focus of the evaluation. The first four dimensions of behavior are useful when we want to quantify behavior, while the last three dimensions are of interest when we are interested in the quality of the behavior. The reader is referred to Alberto and Troutman (1982) for an excellent description of the issues and concerns of collecting observational data.

Conducting interviews is a powerful tool.

Interviews. Conducting interviews is an extremely powerful tool for the collection of needs assessment data. At its simplest level interviewing is simply asking questions and recording the responses. There are three basic interview structures: unstructured, semistructured, and structured (Patton, 1980).

In unstructured interviews, the interviewer may have a general objective but believes this objective is best met by allowing respondents to respond in their own words in their own time. This interview structure is very useful to help identify issues for further examination that were previously unknown or when information to be collected is potentially damaging. However, the lack of structure makes such interviews vulnerable to bias and can often produce uninterpretable information.

Lack of structure makes interviewer vulnerable to bias.

Semistructured interviews are built around a core of questions that all respondents are asked, but they allow the interviewer to branch off and explore responses in greater depth. This structure helps ensure that information of interest is collected from all respondents and allows the opportunity to uncover issues or relationships that were unanticipataed or too complex to be identified by simple questions. Again, however, the unstructured component increases the chances of subjective biases. Because the interviewer follows up on responses to specific questions, there is the potential for the interviewer to lead the interviewee to a desired response.

Structured interviews are very similar to objective questionnaires. The interviewer reads a specific set of questions and might even provide the respondent with a set of responses from which to choose. Clarification of responses is not allowed or is restricted to very narrow limits. This structure reduces the potential of leading interviewees but precludes the uncovering of unexpected issues or complex relationships that are not easily represented in responses to simple objective questions.

Structured interviews are similar to objective questionnaires.

In most cases, the semistructured method has the best chance to provide the most useful information. To maximize this method's potential, steps should be taken to minimize biases and prevent the interviewer from leading the respondent. One helpful technique is to develop a set of acceptable probes that can be used to encourage the respondent to elaborate on responses. Also, the interview should be piloted. In this way a decision can be made as to whether questions in the interview elicit useful information and the interviewing technique of the interviewer can be examined. Based on the pilot test, questions can be modified, probes can be refined, and interviewers can receive feedback on their interviewing technique. By listening through an interview, good probes can be reinforced and leading probes can be identified and alternatives suggested. The following guidelines may be helpful in the development of good interview questions (adapted from Udinsky, Osterlind, & Lynch, 1981):

Develop a set of acceptable probes.

1. Word questions clearly and encourage effective communication between the interviewer and the respondent.
2. Make respondents aware of the purpose of each question they are asked.
3. Be sure that the population from which the respondents have been selected actually has the information being sought and that the interview questions permit the reasonable recovery of this information.
4. Avoid leading questions; that is, questions that suggest a desirable or preferred answer.
5. Ensure that a clear frame of reference is provided for each question, so that all respondents hear questions in the same way.

Another issue to be decided is how information obtained from the interview is to be recorded. Tape recording and writing summaries of each answer after the interview are methods generally used. Writing during the interview is discouraged because it tends to inhibit the interviewee. Tape recording is superior because it allows the interviewer and others to review the interview and prevents the possibility of bias that arises from having the interviewer summarize responses.

Writing is discouraged.

Although we generally think of interviews as being a one-on-one endeavor, a group interview can be extremely useful. Using this technique, the interviewer holds a meeting with the group from which information is being sought, such as parents, interveners, and administrators. The interviewer then explains the purpose of the meeting and breaks the group into a set of several small working groups, each one addressing a specific issue or question. The small groups present their responses to the larger group. Responses are then discussed and refined until there is a group consensus.

Problems are length and complexity of questions.

Questionnaires. One of the most commonly used data collection techniques for needs assessments is the questionnaire. Two of the greatest problems with questionnaires are length and complexity of questions. There is a tendency to keep adding questions to a questionnaire because the response on the question might be "interesting." It is important to keep the purpose of the questionnaire in mind and include only data specific to that purpose. It is equally important to state questions in unambiguous language. In other words, say it as simply as you can.

The majority of the questions on the questionnaire should be objective, with a set of alternatives. However, the inclusion of open-ended questions allows respondents to elaborate on answers and present concerns that were not reflected in the objective questions. Open-ended questions also help clarify ratings of respondents by providing a different source of data that reinforce interpretation of ratings or identify areas where caution should be exercised because of contradictions.

Often respondents are asked to rate specific statements along some kind of scale. In this case, one must decide to use an even-numbered or odd-numbered rating scale. For example, consider the following scale: SA = strongly agree, A = agree, U = undecided, D = disagree, and SD = strongly disagree, applied to the following statement using an odd and even response set:

I need information on appropriate feeding techniques.

Example 1	SA	A	U	D	SD
Example 2	SA	A	D	SD	

In the first example undecided responses can be confusing. Is the person truly undecided or does he or she choose the middle ground to avoid making an affirmative or negative decision? On the other hand, the second example forces the individual to make an affirmative or negative decision about the statement and increases the scorer's ability to interpret ratings of statements. In situations where a clear decision is wanted, an even-numbered scale is superior. Udinsky, Osterlind, and Lynch (1981) have provided detailed guidelines for the construction of questionnaires.

Importance of selecting instruments cannot be stressed enough.

Tests. Tests are another technique frequently used for collecting information in early childhood special education research. These can be particularly useful to interveners when they are attempting to assess child and family needs (see Chapter 3 for a complete discussion of this issue). Tests are usually easy to administer and score, and they have an aura of objectivity and rigor (Casto, in press). However, as many have pointed out (e.g., Garwood, 1982; Ramey, Campbell, & Wasik, 1982; Zigler & Balla, 1982), assessment devices used with handicapped toddlers and infants are unreliable, and they are often invalid for the purposes for which they are being used. Instruments that have the greatest potential are those that are developmentally based and can be used as a tool to help identify needs and then monitor progress throughout the intervention. The importance of selecting appropriate instruments cannot be stressed enough. The following questions can be useful in selecting appropriate tests:

1. Is this instrument appropriate for the population that it is to be used with?

2. What is the purpose of the instrument, and more importantly, is the purpose compatible with data collection needs?
3. Will this instrument provide the best set of information or is there a more appropriate instrument or data collection procedure?

Develop a Management Plan

□ A critical step is the development of a plan for collecting data from the identified sources. A schedule that delineates data-gathering procedures, data synthesis and analysis, and reporting activities is the backbone of the plan. Without a plan data may be collected haphazardly and key data may be missed. Often a time line such as the one in Figure 1 is helpful in summarizing when activities will be initiated and completed. In addition, it is important to delineate individuals who will be responsible for collecting specific data. The staff loading chart contained in Table 1 is an example of a simple way to keep track of these individuals and the data for which they are responsible.

A schedule is the backbone

Collect Data

□ Data should be collected according to steps delineated in the management plan. The time line and staff-loading chart should be referred

Figure 1. An Example of a Time Line for the Collection and Analysis of Needs Assessment Data.

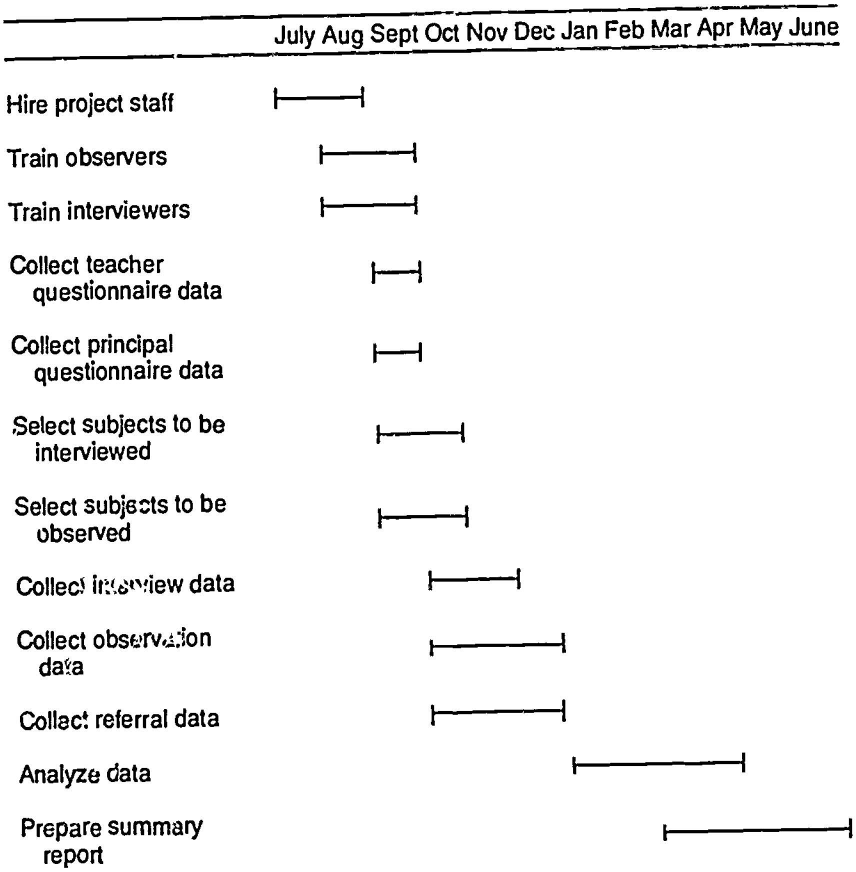

Table 1. Example of a Staff-Loading Chart for Collection of Evaluation Data.

Source	What	When	Who Responsible
1. Individual teachers	1.1 Attitude survey	1.1 As recruited	1.1 Child care liaison
	1.2 Individual plan (teacher objectives)	1.2 Postconsultation (when individual plan completed)	1.2 Child care liaison
		1.3 As recruited	
2. Groups of teachers in centers	2.1 Checklist of workshops	2.1 Returned by 9/30/86	2.1 Child care liaison
3. Directors (in directors' group)	3.1 Checklist of workshops	3.1 Returned by 7/30/86	3.1 Child care liaison
4. Workshop attendees	4.1 Postworkshop evaluations	4.1 Are attending workshops	4.1 Child care liaison, presenters
5. Individual teachers	5.1 Critical incidence questionnaires or multiple choice questions (satisfaction questions on posttest)	5.1 Recruited	5.1 Child care liaison
		5.2 Postconsultation	5.2 Child care liaison
6. Parents	6.1 Interviews	6.1 Midway through and after the intervention	6.2 Data collectors

to often to ensure that data are collected as planned. Changes in original plans should be thought out carefully. Once data collection is under way, there is a tendency to lose track of the original plan or to change the plan because of various data collection pressures. When this occurs, the quality of data invariably suffers, making interpretation impossible. Nothing is more frustrating than spending staff time and program resources collecting data and finding out after all the data have been collected that time and money have been wasted because key data are missing, or data collected are flawed, making interpretation impossible.

Analyze and Interpret Data

Analysis is bringing order to data.

□ The purpose of this step is twofold: to analyze data and to interpret the analysis. Analysis is the process of bringing order to the data by grouping them into meaningful descriptive units, examining data trends within units, and making comparisons between units. Interpretation involves attaching meaning to data trends within and between descriptive units. Techniques or tools are available to aid in the analysis of data, whereas interpretation relies on the evaluator's ability to see and explain meaningful trends and relationships.

Two distinct types of data have been discussed thus far in this chapter. One type lends itself to being quantified and includes such things as frequency counts of behavior, ratings on a scale, or scores on a test. This type of data is categorized as *quantitative* data. The second type is not as easily quantified and includes such things as responses to open-ended questions, descriptions of behavior, and written records.

This type of data is called *qualitative* because these sources provide an indication of the quality of the behavior under study. Both types of data are useful in determining needs and program impact. The techniques used to analyze and interpret these data sources, however, are different. A complete discussion of analysis techniques available for qualitative and quantitative data is beyond the scope of this chapter, and the reader is referred to Borg and Gall (1983) for a more complete discussion of data analysis techniques. In the following section, brief descriptions of analysis techniques for qualitative and quantitative sources are presented and the rationale for their use is examined.

Quantitative Analysis. The analysis of quantitative data is generally performed through some type of statistical procedure. Statistics can be a useful tool for summarizing large data sets, comparing groups, establishing causal influences, and predicting future performance. Statistical procedures can be broken down into three basic types: descriptive, inferential, and nonparametric. Nonparametric procedures are less commonly used and will not be discussed here; the interested reader is referred to Siegel (1956).

Statistics can be useful for summarizing large data sets.

The purpose of *descriptive* procedures is to summarize data systematically to make them more manageable and understandable (Kirk, 1978). As the name suggests, descriptive statistics are used to describe the data that have been collected. They are used to describe average scores (mean, median, or mode), the degree that scores differ from one another (standard deviation), and the degree of association between two groups of subjects (correlations). The advantage of descriptive procedures is that they enable us to summarize large amounts of data in a few descriptive statistics, which greatly aids our ability to interpret findings (Borg & Gall, 1983). A caveat should be noted, however; descriptive statistics often oversimplify data. Rarely are the mean, standard deviation, or other descriptive statistics representative of any one subject from which the data were collected.

Descriptive statistics are used.

Descriptive statistics often oversimplify data.

Common to inferential procedures are such statistical tests as *t*-tests and *F*-tests. The purpose of inferential procedures is to draw conclusions about the whole population from a sample or samples of subjects drawn from the population. These statistical tests are used to determine the likelihood of an observed occurrence happening by mere chance. If the chances of the occurrence are slim, then it is concluded that something systematic has happened, that is, that there is something different about the group that received intervention compared to the group that did not receive intervention. If the research design is sound, a convincing case can be made that it is the intervention that accounts for this difference.

Before we move on to qualitative procedures, it is important to spend some time discussing the limitations of statistics. An understanding of these limitations is crucial if one is to make intelligent decisions regarding their use. First, statistics will not compensate for an evaluation that is poorly designed and conducted. If the data are ambiguous, the statistical analyses will provide answers of equal uncertainty. Second, the absence of statistical confirmation does not "prove" that there was no relationship or impact. It is only through repeated analysis that confirmation of the lack of existence of a relationship is obtained. Finally, statistical significance tells very little about practical significance. With a large enough sample, small differences between groups can be statistically different but be of little practical significance.

Statistical significance tells little about practical significance.

This is not to say that statistical analyses should be discouraged in evaluations. Statistical procedures can have extraordinary power when properly applied. However, it is important to realize there are limitations to their use. For the field-based practitioner in the small class setting, practical significance should be the major consideration in evaluation of impact.

Qualitative Procedures. Qualitative procedures are used to analyze data collected from open-ended questions, interviews, observations, and other data collection methods that provide "softer" data. The procedures can provide a richness of information that is often difficult to achieve with quantitative methods. This richness, however, extracts a price. A qualitative data base typically consists of vast amounts of information from a variety of sources such as written notes on observations, interviews, written impressions, transcripts of electronic recordings, and anecdotal reports. Their management, reduction, and analysis represents a major challenge for the evaluator. The reader is referred to Miles and Huberman (1984) or Patton (1980) for specific guidelines for analyzing qualitative data.

Qualitative data consists of vast amounts of information.

Miles and Huberman (1984) described three components or activities for analysis of qualitative data: data reduction, data displays, and conclusion-drawing/verification. *Data reduction* refers to transforming the large body of written and verbal data collected during observations into clusters, themes, and summaries for the purpose of drawing conclusions. A common technique for the reduction or analysis of qualitative data is through a content analysis. Berelson (1952) described content analysis as a method by which the manifest content of communication can be described objectively and systematically. Typically, the manifest content of communication is clarified by a series of systematic procedures in which (a) the communication is divided into separate units or blocks for analysis; (b) coding categories are developed, defined, and refined; and (c) units of analysis are scored according to the previously developed categories.

Reduction leads to *data displays*, using matrices to organize the categories that most accurately characterize the data as a whole. Miles and Huberman (1984) have suggested that graphic, matrix, or charted displays result in greater accessibility of data than do narrative explanations alone. *Conclusion drawing* follows the data display component and is based on the evaluators' interpretation of data trends.

Conclusion drawing is based on interpretation of data trends.

Although the three data analysis stages occur one after the other, each phase impacts the other phases in a cyclical pattern. Thus, the ultimate interpretation of the data is achieved only after a number of cycles of interaction of data reduction/analysis, data display, and conclusions. The ongoing nature of a qualitative analysis is a critical feature of this approach. Interpretation is not a separate phase; rather, the evaluator attaches meanings and patterns to the data as they are being collected. (See Miles, 1979 for a more detailed discussion of problems associated with qualitative analysis.) Conclusions may be drawn, but they are subject to verification as observations proceed. Human beings are notoriously poor processors of information; judgment is readily flawed, and steps should be taken to prevent misinterpretations. Miles and Huberman (1984) have suggested some strategies, summarized here, that can be used to avoid such misinterpretations:

1. The evaluator should check for data representativeness, that is, assume the data base was derived from a nonrepresentative sample. For example, a check could involve the study of additional cases or the examination of contradictory cases. Similarly, the evaluator should check for reactivity effects of data collectors. In other words, are data representative of what actually occurs in the natural setting?
2. The evaluator should use multiple measurement techniques, referred to as *triangulation*. Since each form of data has its own special weakness, validity can be assessed by the convergence of different data types on the same observation. For example, the determination that an intervener is skilled would carry great weight if it were based on the evaluator's observations, comments from the intervener's peers, child progress, administrator reactions, and any number of other sources. Findings that cannot be substantiated by multiple sources might warrant further examination or be treated with caution.
3. The evaluator should weight items in the data base in terms of their "trustworthiness." A healthy attitude is to assume that data are questionable unless substantial evidence is provided to suggest otherwise.
4. Finally, there are a number of checks the evaluator can employ that are analogous to the considerations of an empirical study: (a) replicating a conclusion in other parts of the data, (b) checking out the plausibility of alternative explanations, (c) looking for negative evidence, and (d) ruling out spurious relationships.

The "fidelity" of qualitative data to reality will always be an issue. In the absence of a body of structured techniques and external checks, the method can very easily degenerate into meaningless, idiosyncratic observations. Qualitative evaluation techniques can be valid and systematic and can provide a rich source of information that is unlikely to be obtained from other sources. Moreover, they can enhance the meaning of quantitative findings and provide greater insight to statistically significant or nonsignificant findings (see Fujiura & Johnson, 1986, for a more complete discussion of this issue).

Qualitative evaluation techniques can be valid.

Develop the Program

□ The final step in the input phase is to develop an intervention program that will meet the needs of the community. This is an ongoing process, in that plans are being developed throughout the collection of information. As tentative plans are developed, they are revised as new data are obtained and summarized. Eventually tentative plans are refined into goals. Goals are then subdivided into more specific objectives. As an example, Figure 2 contains a program goal, related objectives, and activities that were developed in Project APPLE (Gingold & Karnes, 1986) to meet identified community needs. As can be seen in this example, these are management objectives that will be of primary concern during the process evaluation phase of the evaluation plan. In addition, a set of

Tentative plans are refined into goals.

Figure 2. Sample of Goals and Related Objectives, Activities, and Process Evaluation Activities Used in Project APPLE at the Developmental Services Center of Champaign, Illinois.

GOAL 4 To demonstrate comprehensive training and support services for parents of high-risk infants.

Objective 4.1 To develop and maintain a system of ongoing assessment of the education, training, and support needs of parents whose children are receiving early intervention services.

> *Activities 4.1* (1) The Needs Assessment Inventory is currently administered to families upon entering the program. (2) In addition, after 6 months in the program, a questionnaire will be administered which assesses parent satisfaction and addresses parent's interests in additional training and support.
>
> *Process Evaluation 4.1* (1) The Needs Assessment Inventory will be in each child's file within 2 weeks of the child's team assessment. (2) The Family Involvement Checklist will be in each child's file within 7 months of initial team assessment.

Objective 4.2 To maintain and enhance the range of parent activities which will satisfy the assessed needs for training, education, or support.

> *Activities 4.2* (1) In order to be able to meet expressed interests and needs of parents, the staff must be able to develop groups with variable schedules, addressing a variety of topics. Consequently, an annual schedule of parent activities cannot be arranged in advance, but staff can anticipate several short series of information-based meetings for parents. In addition, several support groups will be anticipated. These may be organized according to specific problems (e.g., acting out behavior) or parental characteristics (e.g., single parents). Informal parent-baby play groups may also be organized. (2) Parent-to-parent linkages will continue to be made at the request of parents and of other social service agencies. (3) Parent groups will be coordinated by a program development specialist.
>
> *Process Evaluation 4.2* (1) During the first 9 months of the project, at least four information-based meetings will be held at times convenient to parents, with child care provided, on topics of expressed interest to parents. (2) Documentation of all Parent-to-parent linkages will be kept on file.

Objective 4.3 To develop and maintain an individualized service program for each parent.

> *Activities 4.3* In order to plan each parent's involvement in the program, an individualized plan will be drawn up for the parent(s) of each child. This will be a simplified IEP which specifies activities which each parent will participate in. It will be mutually agreed upon by the parent and by the case manager. This plan will include possible participation in group activities, participation in the child's program, any particular training the parents want or need, and potential referral and linkage to other services.
>
> *Process Evaluation 4.3* (1) Within 2 weeks of each child's staffing, the parent's program plan will be in each child's chart. (2) Parent program plans will be monitored at 6-month intervals, as are the children's plans, for progress toward achieving the objectives set forth.

Objective 4.4 To develop and maintain special training and support services for parents who are identified as delayed, disabled, or potentially abusive/neglectful.

(Continued)

Figure 2. Sample of Goals and Related Objectives, Activities,and Process Evaluation Activities Used in Project APPLE at the Developmental Services Center of Champaign, Illinois. (Continued)

Activities 4.4 (1) The proposed project director and associates of Children's Services are currently developing materials for use with low-functioning or developmentally disabled parents. These materials are being developed with the assistance of a grant from the Governor's Planning Council. They are directed at helping adults understand normal child development and parenting issues and are to be used with small groups of parents.

Process Evaluation 4.4 Copies of session outlines including agendas, attendance records, and parent evaluations of sessions will be on file. Similar documentation will be on file as subsequent training sessions occur.

objectives related to child and family outcomes would be developed that would be of primary concern in the outcome phase of the evaluation plan.

In closing, the purpose of the input phase of the evaluation plan is to assess the needs of handicapped young children and their families. In a sense it is like developing a navigation plan for an ocean voyage. If the navigation plan of a voyage is flawed, the ship will never reach its destination, no matter how competent or diligent the crew. In the same way, if a program does not conduct an adequate input evaluation, the plans developed to reach its destination (to meet the needs of handicapped infants, toddlers and their families) will be flawed, preventing the program from ever reaching its goals.

PROCESS EVALUATION

□ In process evaluation the focus is on navigation toward the goals and objectives of the proposed plan. As information is obtained, adjustments can be made in the implementation process to keep the proposed plan on track. Furthermore, this process provides feedback to interveners on progress being made by specific children and their families as well as information on the overall progress of the program.

Adjustments can be made in the implementation process.

Program procedures and intervention methods or strategies that are employed to achieve program goals must be closely monitored. If the process is not monitored, the outcome evaluation of the program will be misleading. For example, suppose program objectives had not been met; we would probably conclude that the intervention used in the program was ineffective. The outcome, however, could also be attributed to inadequate implementation of procedures. For example, teachers might lack the time to complete interventions, materials might be insufficient, or a child's illness might preclude program completion. Negative findings may not be an indication of the program's "goodness"; rather, they may indicate the inadequacy of its implementation. One can see how evaluation of procedures supersedes the evaluation of objectives. If the procedures have not been monitored, then the evaluation of outcome is necessarily ambiguous.

Negative findings may indicate inadequacy of implementation.

Another concern in process evaluation is program management. Effective implementation of the program is intimately related to the adequate management of program resources. Again, the major concern is the identification of the relationship of management practices to program effectiveness. Management systems must efficiently allocate program resources such as personnel, equipment, and space. The basic evaluation question is, How do these and other resources constrain or enhance the implementation of the program?

How do resources constrain or enhance implementation?

Related to program management is the recent concern over program costs and costs in relation to program benefits. Cost effectiveness techniques have been developed to address this concern (see Levin, 1983, for a detailed discussion of cost effectiveness). These techniques fall somewhere between the process and output phases of evaluation. At one level, cost effectiveness techniques provide information that provides direction as to inefficient program components. However, we also obtain information concerning program effectiveness relative to costs. Although cost effectiveness evaluation is very popular, some have questioned its worth in early childhood special education programs (Strain, 1984).

Perhaps the most important aspect of the process evaluation phase is the monitoring of child or family progress toward objectives. This may be the first indication of faulty intervention plans that need modification. Furthermore, monitoring of progress creates a template that can be used to trace the effect of the program on children and families throughout the intervention.

Most important is monitoring child or family progress.

The purpose of this phase of the evaluation plan is to monitor progress toward goals and objectives and to modify the original plan when data indicate a need for a change. In the same way that a captain navigates a ship to its destination by taking frequent measurements and adjusting the ship's course as needed, the evaluator navigates the program to its destination by taking frequent measurements and adjusting the plan as needed. As an example, Figure 2 contains a sample set of goals, objectives, planned activities to meet goals and objectives, and possible process evaluation activities to be used to monitor progress toward goals and objectives.

OUTCOME EVALUATION

❑ The focus of this phase is to determine the impact of the program on children, their families, and the community. Such a view equates this phase of the evaluation with educational research; interpretation means determining the causal effect of the program on outcomes. That is, we attempt to determine the impact of the program on children and their families. Research methods are used to establish that the program is the most likely explanation for family or child outcomes. In other words, the purpose of outcome evaluation procedures is the elimination of as many rival explanations for child and family changes as possible. For example, with a "strong" research design, if we were to observe improvement in test scores after an educational intervention, we would infer that test score improvement was caused by the intervention. However, to the extent that other explanations can account for this improvement, we lack what is termed *internal validity*. The "stronger" a design is, the greater the internal validity or the more readily other explanations for the findings can be dismissed. In essence, we attempt to design our research so that all other explanations except the intervention are ruled out as causing observed changes. Described below are the eight threats to internal validity outlined

We would infer that test score improvement was caused by intervention.

by Campbell and Stanley (1963). Each threat can be logically controlled by the elements of evaluation design.

1. *Historical* threats are events unrelated to the program that affect outcomes. For example, the introduction of a child into a program may stimulate greater home involvement by the child's parents. Therefore, changes at postprogram assessments may be equally attributable to the program or the parents.
2. *Maturation* threats refer to various forms of growth by the child over the course of the program. If maturation is unrelated to the program, then the effect of the program is indeterminate. This is problematic in programs for children under age 3, for whom rapid change is expected over very short time periods.
3. *Testing* threats relate to the concern that the act of testing (or observing) may affect in some manner the postprogram assessment. This is most often seen in subjects becoming more "test-wise" after having been administered the preprogram assessment.
4. *Instrumentation* threats are changes occurring in the measurement. For example, if we have one observer rating a child's performance on a set of skills prior to intervention and a second observer rate the child's performance after the intervention, we may not be able to determine whether differences in pre/post ratings are attributable to differences in the interpretations of observers or differences in the child's behavior.
5. *Regression* is a statistical tendency for subjects with extreme scores at one time to score closer to "average" the second time. This has important implications for the evaluation of programs designed to intervene with children who perform differently than the "average" child (e.g., handicapped infants and toddlers). For example, subjects selected on the basis of low test scores in a screening may perform significantly better at postprogram assessment. The change may be due to regression and not the program.
6. *Selection* is a major threat in evaluation, particularly when we must use intact groups and cannot randomly select who will receive intervention. Since program effects are frequently inferred when differences are observed between subjects in the program and a comparison group excluded from the program, we must take steps to ensure that preintervention differences do not explain the postintervention differences. In other words, the quality of mother-infant interactions in the intervention group may have been superior to the mother-infant interactions of the comparison group prior to intervention. As a result, it will be difficult to conclude that the program accounted for discrepancy in mother-infant interactions between the control and intervention groups.
7. *Mortality* represents the loss of subjects during the course of the program. The remaining subjects may bias the outcome since the pre- and postprogram comparisons are based on different sets of subjects. For example, uncooperative families may withdraw from a program because of differences with the program staff. As a result, only cooperative families remain in the program; their postprogram scores are then compared to the preprogram scores, or the scores of another group that contains scores from both cooperative and uncooperative families.

8. *Selection interactions* are the interactions of other threats with selection. Some threats may be manifested with certain types of children. For example, a program may be composed of children equally deficient in some skill area. Half the group is chosen to receive a remedial intervention and the other half comprises a control group for purposes of comparison. If the intervener were to select children on the basis of their "promise," then a threat of selection-maturation exists.

Design Considerations

Research designs systematically examine effectiveness of programs.

□ Research designs are the structures by which the search for answers to questions about interventions are organized (Udinsky, Osterlind, & Lynch, 1981). In other words, they enable us to systematically examine the effectiveness of our programs and collect insights about how the programs might operate in other situations. The strength of a given design is determined by the design's potential to control for the threats to intemal validity. A "strong design" is one that allows us to conclude that changes in children and their families are most likely a result of the intervention rather than some unrelated factor.

Three dimensions differentiate most evaluation designs: (a) presence or absence of a preprogram measure on the outcome measure, (b) presence or absence of a nontreatment comparison group, and (c) whether groups are intact or randomly composed. A complete discussion of experimental design is beyond the scope of this chapter. The reader is referred to Campbell and Stanley (1963) for more information on issues related to research design. The designs included in this section are limited to those with the greatest potential for controlling the threats to internal validity.

Qualitative ideal is an extensive description of events in the natural setting.

Absence of Preprogram Measures and a Nontreatment Group. Under these conditions quantitative procedures are useless. The only potential for useful information is the use of qualitative methods. These methods have stimulated recent interest in the educational evaluation literature. What had been heresy years ago has achieved respectability. The qualitative ideal is represented by an extensive description of events in the natural setting.

The observer is in stark contrast to experimental tradition.

The field work of anthropologists perhaps best exemplifies the qualitative methodology. Of primary importance to this method is the attempt to faithfully and continuously record all events. This requires detailed descriptions of the setting and of the involved individuals and their interactions; usually generous quantities of quotations are used. Values of the observer must be "suspended" so that interpretations of events are not distorted by observer values. The observer who considers the context of the events being recorded is in stark contrast to the experimental tradition, where control of variables is paramount to the research effort.

The strength of qualitative approaches is the degree of detail that can be brought to bear upon the evaluation question. Rich portrayals of the subject matter and its associated context can be a source of valuable insights into process and possible causal relations. Furthermore, the researcher is less susceptible to being blinded by structured methods and is therefore more likely to be sensitive to unanticipated findings.

A serious weakness in the qualitative strategy is the difficulty of establishing the validity of the data. It is impossible for observers to be passive recorders of events; rather, they are filters through which considerable amounts of information do not pass. There are several explanations for this situation. First, it is not possible to accurately record every event in a given situation. Every situation presents far too many pieces of information; this is compounded by the exploratory nature of most qualitative studies, where there is uncertainty about which events are relevant and which irrelevant. Second, there are no guarantees that attitudes and biases do not distort the observer's perception of events. If information is selectively attended to, then it will very likely be the information most congruent with the observer's frame of reference. Third, the intimate involvement typically required of the observer can invite reactivity effects. In addition, the involvement can be emotional, which necessarily reduces the observer's objectivity. Fourth, many data bases must be constructed from memory, which compounds the problems of attitudes and biases.

A weakness is establishing validity of the data.

Although these problems may be more pronounced in the qualitative method, they are not unique to the approach. An evaluator conducting a traditional empirical study is just as susceptible to biases in the determination of what variables to manipulate and outcomes to assess. Regardless of the method employed, reality must be reconstructed, and biases and values of the individual doing the reconstruction will impact the effort.

Absence of a Nontreatment Comparison Group. As with the previous design dimension, under these conditions traditional quantitative procedures are of little value. Single subject designs, however, can control the threats to internal validity and be extremely useful in attempts to determine program impact.

Single subject methodology was developed to create conditions closely approximating those in control group designs when control groups are not available. Basically, children receiving the intervention are assessed repeatedly throughout the treatment period. Essentially, they serve as their own controls. This is a powerful design, whose logical strength rivals that of the true experimental design. Kazdin (1982) outlined three characteristics of the single subject design: continuous assessment, baseline assessment, and analysis of trend.

Children are assessed repeatedly.

The single subject design has many variations, and a systematic review of them would require many more pages than are available here. This variability reflects the adaptability of the repeated measures design to many different contexts and needs. It is an extremely flexible design.

1. Continuous assessment is the fundamental characteristic of the repeated measures design. Since no control group is employed, the evaluation of effect is based on performance changes that coincide with the onset of the intervention. There is strong basis for inferring effect when a series of assessments begins to yield different results after implementation of an intervention. Use of continuous assessment provides a control for maturational threats since program effects can be seen against the backdrop of growth prior to the intervention.
2. Baselines provide (a) an estimate of existing levels of performance, and (b) a "prediction" of what the future performance should be if the

intervention has no effect. Prediction is central to this design, since inferring effect requires changes in predicted performance at the point of intervention. Baselines provide a control for selection threats since treatment and nontreatment comparisons are within the same subject. In addition, regression effects are improbable explanations when stable baselines are achieved.

3. The notion of trend is related to predicted performance. If program effectiveness is inferred from departures from baseline performance, then performance trends over the repeated assessments have important analytic value. *Trend* refers to stable increases or decreases in performance. In the ideal evaluation example, baseline performance is stable (no change in the preintervention period), and with the onset of intervention, performance shows a marked trend.

A number of design options can help the evaluator better assess the impact of an intervention when a comparison group cannot be constructed. Some of the more commonly employed single subject designs are (a) reversal designs, (b) multiple baseline designs, (c) changing criteria designs, and (d) multielement designs. The reader is referred to Kazdin (1982) or Kratochwill (1978) for detailed reviews of single subject designs.

One group is given the intervention, one is not.

Intact Groups Pretest and Posttest. This situation allows us to use traditional quantitative procedures to establish that the program had a significant impact. At the simplest level, one group is given the intervention and one group is not; both groups are tested on a pre/post basis. This is a reasonably strong design that depends on how plausible the selection bias is an alternative explanation for findings. By analyzing pretest, however, the evaluator can determine whether or not groups were equivalent prior to the intervention. If they are equivalent prior to intervention, selection bias is much less plausible. History, maturation, testing, and instrumentation are controlled by the presence of a comparison group since each of these effects should operate equally on both groups. The pretest accounts for selection and mortality effects. However, regression is a threat, as it is in all intact group designs.

Regression is a threat.

Many educational researchers and evaluators have resorted to *matching* as an additional methodological control when intact groups exist. In matching, the evaluator selects children for the nonintervention group on the basis of their similarity to the intervention group members. The matching process is systematic in that behavior scales, test scores, or other quantifiable measures (rather than subjective judgments) are used to determine similarity. Having matched the children, the implicit assumption is made that the two groups are equivalent. Any changes observed at the posttest are presumed to be due to the intervention However, there may be an array of other relevant variables not considered, such as motivation and parental support, that may be equally as important as or more important than the variables used for matching. If we can be reasonably confident that no other variable is important to determining posttest skill, then the matching process adds to our confidence. It strengthens the inference only to the extent that the matching variable(s) represents the array of factors important to the outcome. Otherwise, selection-maturation interactions continue to be threats to this design.

Changes at posttest are presumed due to intervention.

Randomly Created Groups. Evaluations comparing groups are most conclusive when random assignment of subjects to groups is employed. Rather than employing an intact group for the intervention group, the evaluator would assign children to the treatment program and control group in some random fashion to control for a systematic bias. This is the most elegant and powerful design available. With the exception of mortality, the design effectively controls all threats to internal validity.

The design controls threats to internal validity.

However, random assignment guarantees only probabilistic equivalence, a notion many nonresearchers find less than compelling. Sampling variability can lead to initial group nonequivalence on critical variables (e.g., IQ, motivation, or any other key variable). In order to avoid this problem, many evaluators first match subjects and then randomly assign each member of a matched pair to either the intervention or nonintervention group. Again, what we have done by random assignment is eliminate any systematic bias in group membership.

A major impediment to the use of this design is the lack of control an evaluator typically has in the applied setting. This can be an ethical issue. Service delivery is dictated by chance rather than need. For this reason, we find relatively few true experiments in field situations. A situation that may allow us to use this design is when we have limited resources and are not able to serve all children or families who may need services. Random assignment of these individuals to a control or intervention group may be the most equitable distribution of limited resources.

A major impediment is lack of control in the applied setting.

Implementing the Outcome Evaluation Phase

□ As previously discussed in the input evaluation section of this chapter, it is critical that this phase of the evaluation plan be carried out in a systematic and careful manner. A poorly conceived or implemented outcome evaluation will obscure interpretation of program impact, with a disastrous effect on the program. With some slight changes, the steps outlined in the input evaluation section for determining needs are equally useful for the implementation of a good outcome evaluation. To review, the steps are (a) determine key elements, (b) identify information sources, (c) develop a management plan, (d) collect data, (e) analyze and interpret data, and (f) develop the program (this step is not included in the output evaluation phase). The slight changes in these steps in the outcome evaluation phase are described below.

Determine Key Elements. As in the input phase, we must determine the purpose of and audience for the outcome evaluation. We must also develop a set of questions that should be answered. For example,

1. What impact did the support groups have on families that participated in the program?
2. Do children make significant progress as measured by the Bailey Scales?
3. Are parents satisfied with the program?

Identify Information Sources. As in the input phase, we must determine the sources of information needed to answer our evaluation questions. It is important that our data collection efforts go beyond just collecting child change data. Programs for handicapped infants, toddlers, and their

Data collection efforts go beyond change data.

families have impact beyond those limited to children, and we must go beyond them as a data source so that we can assess these impacts.

An additional concern in this step is the selection of a research design. We must select the design that will give us the greatest control over the internal threats to validity and still be within the limitations of the situation (e.g., is there a comparison group or can we randomize?).

The methods available to collect data are essentially the same as those described in detail in the input evaluation section of this chapter.

Develop a Management Plan. The importance of a management plan as described in the input evaluation phase applies equally to outcome evaluation. Steps described to help manage data collection should also be employed.

Collect Data. Again, the issues and concerns discussed in regard to the input evaluation are equally applicable to the output evaluation.

Analyze and Interpret the Data. As with the previous phases, issues related to analysis and interpretation have been discussed in detail earlier in this chapter.

In closing, the purpose of this phase is to determine the impact of our program on children, their families, and the community. In the voyage analogy, the plan developed by the mayor was considered successful only if the health of the port improved. In the same vein, even the best-designed program, appropriately implemented, would be of little value if it didn't have the desired impact on children and their families.

A HIGH-QUALITY EVALUATION PLAN

□ How do we define a high-quality evaluation plan? The Joint Committee on Standards for Educational Evaluation (1981) was formed, under the direction of Daniel Stufflebeam, to develop a set of standards to which a good evalaution plan must conform. This group was made up of representatives from some of the most prominent educational organizations: National School Boards Association, National Educational Association, National Association of Elementary School Principals, Education Commission of the States, National Council on Measurement in Education, American Association of School Administrators, American Educational Research Association, American Federation of Teachers, American Personnel and Guidance Association, American Psychological Association, Association for Supervision and Curriculum Development, and Council for American Private Education.

RATIONALE FOR DEVELOPING STANDARDS

□ Standards were developed for two basic reasons. First, it was felt that the technical quality of many evaluation studies was insufficient to provide adequate data. As previously discussed, this concern has also been raised with regard to evaluation studies in early childhood special education (Dunst & Rheingrover, 1981; Odom & Fewell, 1983; Simeonsson, et al., 1982; White & Casto, 1984; White, et al., 1984; Wolery, in press; Wolery & Bailey, 1984). Second, it was realized that program evaluation could be corrupted to produce results that reflect the program's bias and serve the needs of the program. Suchman (1967) grouped such manipulations into four categories: eyewash, whitewash, posture, and postponement.

"Eyewash" is a technique by which an ineffective program is made to look better by selecting those aspects of the program that will make the program look good and ignoring those aspects that will not. A common technique is to collect many pre/post measures and then report only those measures on which significant growth was shown. If enough measures are used, significant changes can be found by mere chance.

Significant changes can be found by mere chance.

"Whitewash" takes the deception a step further than eyewash by presenting misleading or inaccurate data. An often-used method is to present glowing testimonials on the impact of a program without presenting data to support the claims. Anyone who has watched more than an hour of television should be familiar with this technique. "Posture" is used by a program to give the impression of a rigorous evaluation design and quality program. One method frequently used is to report complex data collection or analysis procedures that are difficult to understand. The complexity of the analysis sounds good and makes it appear that the program is being rigorously evaluated.

"Postponement" is used to avoid or delay some action that the program administration does not want undertaken. By suggesting that an evaluation study be conducted before a decision can be made, the administration can stall until the storm blows over and they are no longer receiving pressure to implement the action.

PURPOSE OF EVALUATION STANDARDS

□ The committee felt that a set of standards could help improve the professionalism of program evaluation by giving people benchmarks for developing and judging the quality evaluation plan. It was the hope of the committee that these standards would reduce the number of technically inadequate evaluation plans and help ferret out reports of evaluation plans that have been corrupted. The committee concluded that a high-quality evaluation plan has (a) utility, (b) feasibility, (c) propriety, and (d) accuracy. Each of these elements has a set of more specific features that the evaluation plan must have in order to be considered as meeting the requirements of that standard.

Standards would reduce technically inadequate evaluation plans.

Utility

□ For an evaluation plan to have utility, data collected from the evaluation plan must have potential usefulness to the program and/or consumers of the program. Several steps should be taken to ensure the utility of the evaluation plan. The audience for the evaluation must be identified and steps should be taken to ensure that the plan is appropriate to meet their needs. Furthermore, information must be of a broad enough scope to answer all the pertinent evaluation questions. When the results of the plan are written, information must be clear and easily understood. Otherwise, the report will sit on a shelf and be of little use. Finally, it is critical that results of the evaluation plan be disseminated promptly. Nothing detracts more from the impact of a good evaluation plan than the presentation of the findings after people are no longer concerned with the outcomes.

Feasibility

□ Feasibility refers to the plausibility of implementing the evaluation plan. A major concern is the practicality of components of the plan. For example, asking interveners to give a battery of tests in addition to their normal duties is probably not practical. They will feel pressured to collect the data and will probably hurry through their administration. As a result, the morale of the interveners will be hurt and the data collected will be of poor quality.

Not every program can afford a rigorous evaluation.

New innovations should be based on solid data.

In a similar vein, it is important that the cost of the evaluation plan can be in tune with the benefits to be obtained from the plan. Not every program can afford to develop a rigorous evaluation that employs a solid experimental design to establish program impact. In fact, if a program does not have the rescurces to conduct an adequately controlled study, it should not be undertaken. Results from technically unsound investigations are bound to be specious and add to the confusion regarding the efficacy of early intervention (see Wolery & Bailey, 1984 for a complete discussion of this issue). Furthermore, it is far more important that the program document a high-quality implementation of intervention, rather than the impact of the intervention. If a program can establish that the intervention being used represents "best practice" and that the intervention is implemented properly, results are bound to occur. If they do not, however, this is not an indictment of the program. By documenting that what the interveners are doing represents "best practice" and that the program was adequately implemented, the program planners have established accountability. That is not to say that a question could not or should not be raised as to why there was little positive impact. This question may then be examined through an evaluation plan that uses good experimental (Campbell & Stanley, 1963) or at least good quasi-experimental design (Cook & Campbell, 1979). Furthermore, it is also important that when new innovations are proposed, they be based on solid data so that we are not in the business of creating new educational myths. For the typical program, however, efforts should be directed at identifying client needs, developing a plan to meet those needs, documenting the plan that represents "best practice," and monitoring progress on the plan.

Propriety

It is a rare endeavor that has all positive outcomes.

□ This standard relates to how equitable and ethical the evaluation plan is. Evaluators, like everyone else, have a responsibility to respect the rights of individuals connected with the program, and the evaluation plan should reflect this responsibility. Readers of evaluation reports should beware of reports that have nothing but positive findings. It is a rare educational endeavor that has all positive outcomes. Readers should also be concerned when a report does not seem to have a breadth of measures included in the evaluation plan. Sometimes we must ask what is not included in the report. We can have greater confidence in a report that includes both positive and negative findings and will be less likely to believe that damaging results were withheld.

Accuracy

□ For an evaluation plan to have accuracy, steps must be taken to ensure that the data collected are correct and representative of the program.

Perhaps the most important consideration for an evaluation plan is the validity of information obtained. Validity can be thought of as the degree to which a test or procedure provides information relevant to the decision to be made. In other words, do the tests or procedures used in the identification plan measure what they purport to measure?

Validity...degree which a test provides relevant information.

Several steps should be taken to help ensure the development of a valid evaluation plan (see Goodwin & Driscoll, 1980, for a detailed discussion of validity).

1. It is imperative that multiple sources of information be used. Using multiple sources of information maximizes opportunities for children and their families to demonstrate their growth and thereby enhances the program's ability to monitor progress toward objectives and determine the impact of the program.
2. The selection of formal sources of data (e.g., standardized tests or published criterion-referenced tests) should be based on the degree of validity that has been established for these sources. Either they should be highly correlated with established tests that measure the same trait (concurrent validity), or they should be good predictors of the child's future behavior (predictive validity). Formal sources of data that only report face or content validity are suspect and should be avoided. Technical manuals of tests should include a discussion of the tests' validity.
3. Formal sources of data chosen should be used with the population for which they were intended as well as in the manner in which they were intended to be used.
4. Informal methods of data collection (e.g., intervener observations, intervener developed tests or checklists, interviews, etc.) should have good face validity. That is, the information obtained from the informal source should be relevant to the trait or traits being measured.

A second consideration, of equal importance, is reliability—the extent to which variations in data reflect actual variations in the phenomena under study rather than being a result of measurement error (see Goodwin & Driscoll, 1980, for detailer discussion of reliability). In other words, can we be assured that the test or procedure being used will consistently produce the same results given the same input? As with validity there are steps that can be taken to ensure the development of a reliable identification and plan. First, selection of formal sources of data should be based on the degree of reliability established for each source. Reliability coefficients should be found in the test's technical manual. Second, programs can take steps to examine the reliability of informal sources of data they are using. For example, both parents could be asked to fill out checklists, or it may be possible to have an intervener and an aide complete the intervener checklist independently. By examining the same informal source of data completed by two individuals regarding the same child, one can determine whether or not information obtained from this source is consistent across individuals.

Reliability coefficients should be found in the test's manual.

An often overlooked concern is *harmony*, or the degree to which the evaluation plan is associated with the goals of the program. Harmony depends on whether the tests and procedures chosen are appropriate for matching the child with the program. It is possible that specific tests or procedures within an evaluation plan are reliable and valid but are not compatible with the goals of the program. Although data collected through these procedures will provide what appear to be good data in the sense that they are derived from reliable and valid practices, the data are not useful in determining the impact of the program.

A concern related to harmony is the collection of defensible information sources. In other words, does the information source have the potential to provide good information about the activity being judged? For example, one program goal might be to improve parent-child interactions during play periods. The intervener who has worked on this goal all year may not be the best source of data to judge whether or not any growth has occurred. The intervener may be too invested to make an unbiased judgment. On the other hand, asking the program administrator who has limited contact with the parents would be even worse. This individual would not have adequate knowledge of parent-child interactions to make such a judgment.

When reading evaluation reports we should examine the methods section carefully and not just read the conclusions. We must look for a systematic data collection procedure that relates to the intentions of the program.

We must determine whether conclusions are justified.

It is critical that we determine whether or not the conclusions are justified based on the data collected. Without examining how data were collected and analyzed we have no basis from which to make such a judgment. As a general rule, it is best to be guarded when interpreting the results of any evaluation study that does not adequately describe how data were collected and analyzed.

CONCLUSION

□ In this chapter, program evaluation has been presented as a comprehensive interwoven process comprising three phases: input, process, and output. In the input phase, evaluation efforts are directed at the identification of needs and the matching of program capabilities to identified needs. In the process phase, the focus of evaluation efforts is on the monitoring of progress toward objectives and program implementation. In the outcome phase, program impact is determined. The first two phases are critical to the development of a high-quality program. The emphasis of most programs for handicapped infants, toddlers, and their families should be placed on the input and process phases of evaluation. Without these phases a program is sure to have problems. Moreover, a program should not attempt to undertake an outcome evaluation for which it does not have the resources or expertise. The literature is full of confusing findings with regard to the impact of early intervention. A poorly conceived outcome evaluation produces confusing findings. On the other hand, a good comprehensive evaluation plan can greatly enhance our ability to meet the needs of handicapped infants, toddlers, and their parents; help us establish accountability; and provide us with the ammunition to convince policy makers of the need for and benefits from early intervention.

REFERENCES

Alberto, P. A., & Troutman, A. C. (1982). *Applied behavior analysis for teachers: Influencing student performance.* Columbus, OH: Merrill.

Berelson, B. (1952). *Content analysis in communication research.* Glencoe, IL: Free Press.

Borg, W. R., & Gall, M. D. (1983). *Educational research.* New York: Longman.

Bricker, D., & Littman, D. (1982). Intervention and evaluation: The inseparable mix. *Topics in Early Childhood Special Education, 1*(4), 23-33.

Campbell, D. T., & Stanley, J. C. (1963)., *Experimental and quasi-experimental designs for research.* Boston: Houghton, Mifflin.

Casto, G. (in press). Research and program evaluation in early childhood special education. In S. L. Odom and M. B. Karnes (Eds.), *Research in early childhood special education.* Monterey, CA: Brooks Cole.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Dunst, C. J., & Rheingrover, R. M. (1981). An analysis of the efficacy of infant intervention programs with organically handicapped children. *Evaluation and Program Planning, 4,* 287-323.

Fujiura, G. T., & Johnson, L. J. (1986). Methods of microcomputer research in early childhood special education. *Journal for the Division for Early Childhood, 10*(3), 264-269.

Garwood, S. G. (1982). (Mis)use of developmental scales in program evaluation. *Topics in Early Childhood Special Education, 1*(4), 61-69.

Gingold, W., & Karnes, M. B. (1986). *Program progress report: Project APPLE.* Springfield, IL: Illinois Governor's Planning Council on Developmental Disabilities.

Goodwin, W. L., & Driscoll, L. A. (1980). *Handbook for measurement and evaluation in early childhood education.* San Francisco: Jossey-Bass.

Isaac, S., & Michael, W. B. (1981). *Handbook in research on evaluation: For education and the behavioral sciences* (2nd ed.). San Diego, CA: EdITS.

Joint Committee on Standards for Educational Evaluation. (1981). *Standards for evaluations of educational programs, projects, and materials.* New York: McGraw-Hill.

Kazdin, A. E. (1982). *Single-case research designs.* New York: Oxford Press.

Kirk, R. E. (1978). *Introductory statistics.* Monterey, CA: Brooks/Cole.

Kratochwill, T. R. (1978). *Single-subject research: Strategies for evaluating change.* New York: Academic Press.

Levin, H. M. (1983). *Cost-effectiveness: A prover.* Beverly Hills, CA: Sage.

Miles, M. B. (1979). Qualitative data as an attractive nuisance: The problem of analysis. *Administrative Science Quarterly, 24,* 590-601.

Miles, M. B., & Huberman, A. M. (1984). Drawing valid meaning from qualitative data: Toward shared craft. *Educational Researcher, 13,* 20-30.

Morris, L. L., & Fitz-Gibbon, C. T. (1978). Evaluator's handbook. In L. L. Morris (Ed.), *Program evaluation kit* (pp. 1-133). Beverly Hills, CA: Sage.

Odom, S. L., & Fewell, R. R. (1983). Program evaluation in early childhood special education: A meta-evaluation. *Educational Evaluation and Policy Analysis, 5,* 445-460.

Patton, M. Q. (1980). *Qualitative evaluation methods.* Beverly Hills, CA: Sage.

Ramey, C. T., Campbell, F. A., & Wasik, B. H. (1982). Use of standardized tests to evaluate early childhood special education programs. *Topics in Early Childhood Special Education, 1*(4), 51-60.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gragne, & M. Scriven (Eds.), *Perspectives on curriculum evaluation,* (pp. 39-83). AERA Monograph Series on Curriculum Evaluation No. 1. Skokie, IL: Rand McNally.

Scriven, M. (1973). Goal-free evaluation. In E. R. House (Ed.), *School evaluation: The politics and process.* Berkeley, CA: McCutchan.

Scriven, M. (1974). Evaluation perspectives and procedures. In W. J. Popham

(Ed.), *Evaluation in education: Current applications*, (pp. 1-93). Berkeley, CA: McCutchan.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill.

Simeonsson,, R. J., Cooper, D. H., & Schiener, A. P. (1982). A review and analysis of the effectiveness of early intervention programs. *Pediatrics, 69*(5), 635-641.

Strain, P. S. (1984). Efficacy research with young handicapped children: A critique of the status quo. *Journal of the Division for Early Childhood, 9,* 4-10.

Stufflebeam, D. L. (1971). The relevance of the CIPP evaluation model for educational accountability. *Journal of Research and Development in Education, 5*(1), 19-23.

Stufflebeam, D. L. (1974). Alternative approaches to educational evaluation: A self-study guide for educators. In W. J. Popham (Ed.), *Evaluation in education: Current applications,* (pp. 95-143). Berkeley, CA: McCutchan.

Suchman, E. (1967). *Evaluation research. New York: Sage.*

Tyler, R. W. (1942). General statement on evaluation. *Journal of Educational Research, 35*(7), 492-501.

Tyler, R. W. (1958). The evaluation of teaching. In R. M. Cooper (Ed.), *The two ends of the log: Learning and teaching in today's college,* (pp. 164-176). Minneapolis: University of Minnesota Press.

Tyler, R. W. (1971). Accountability in education: The shift in criteria. In L. M. Lesinger & R. W. Tyler (Eds.), *Accountability in education.* Worthington, OH: Charles A. Jones.

Tyler, R. W. (1974). Introduction: A perspective on the issues. In R. W. Tyler & R. M. Wolf (Eds.), *Crucial issues in testing,* (pp. 1-10). Berkeley, CA: McCutchan.

Udinsky, B. F. Osterlind, S. J., & Lynch, S. W. (1981). *Evaluation resource handbook: Gathering, analyzing, reporting data.* San Diego, CA: EdITS.

White, K. R., & Casto, G. (1984). An integrative review of early intervention efficacy studies with at-risk children: Implications for the handicapped. *Analysis and Intervention in Developmental Disabilities, 5,* 7-31.

White, K. R., Mastropieri, M., & Casto, G. (1984). An analysis of special education early childhood projects approved by the joint dissemination review panel. *Journal of the Division for Early Childhood, 9,* 11-26.

Wolery, M. (1987). Program evaluation at the local level: Recommendations for improving services. *Topics in Early Childhood Special Education, 7*(2), 111-123.

Wolery, M., & Bailey, D. B. (1984). Alternatives to impact evaluation: Suggestions for program evaluation in early intervention. *Journal of the Division for Early Childhood, 9,* 27-37.

Zigler, E., & Balla, D. (1982). Selecting outcome variables in evaluations of early childhood special education programs. *Topics in Early Childhood Special Education, 1*(4), 11-22.