DOCUMENT RESUME

ED 335 487 CE 058 624

AUTHOR Dusewicz, Russell A.; And Others
TITLE How To Evaluate Adult Education.
INSTITUTION Research for Better Schools, Inc., Philadelphia, Pa.
SPONS AGENCY New Jersey State Dept. of Education, Trenton. Div. of Adult Education.
PUB DATE 87
NOTE 71p.
PUB TYPE Guides - Non-Classroom Use (055)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Adult Education; *Data Collection; *Evaluation Methods; Formative Evaluation; Measurement Techniques; Needs Assessment; Program Development; Program Effectiveness; *Program Evaluation; *Statistical Analysis; Summative Evaluation

ABSTRACT

This guide is designed to help educators plan and carry out evaluations of adult education programs. Following an introduction, Section 2 discusses how the evaluation process fits into the concept and normal cycle of program development. Section 3 describes the stages of the evaluation process, including needs assessment, program planning, process evaluation, and outcome evaluation. Section 4 provides examples of different designs that can be used in adult basic education--case designs, time series designs, and comparison group designs. Section 5 discusses these methods of data collection: questionnaires; interviews; observations; and tests. Section 6 describes statistical procedures for analyzing evaluation data, including descriptive statistics, tests of difference, and tests of relationships. Section 7 illustrates the preparation of the program evaluation outline, a device to help program staff sketch out the essential elements (objectives, measurement techniques, time schedule, evidence of accomplishment) of the evaluation plan. Section 8 outlines the different factors that objectively determine the effectiveness of a program. Section 9 describes various reporting and dissemination techniques for evaluation results. A list of 26 references is appended. (YLB)

HOW TO EVALUATE ADULT EDUCATION

With competition increasing for adult education resources, and the demand for services also increasing, adult educators must be assured that evaluation will produce benefits. Evaluation can fulfill funding mandates, determine benefits in relation to costs, and establish a data base for better management.

by Russell A. Dusewicz, Thomas W. Biester, and Jane L. Kenney

RBS

Research for Better Schools
444 North Third Street
Philadelphia, PA 19123

HOW TO EVALUATE ADULT EDUCATION

by Russell A. Dusewicz,
Thomas W. Biester, and
Jane L. Kenney

RBS
Research for Better Schools
444 North Third Street
Philadelphia, PA 19123

1987

The work upon which this handbook is based was funded in part by the Division of Adult Education, New Jersey Department of Education. The opinions expressed do not necessarily reflect the position or policy of the Department, and no official endorsement should be inferred.

Graphic Art by Peter Robinson
Word Processing by Carol Crociante

This is a product of the RBS Research and Development Project, Keith M. Kershner, Director and the School Improvement Services Project, Russell A. Dusewicz, Director.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION AND OVERVIEW

In adult education, as in other areas of education, the question is often asked: "Why evaluate?" What purpose does evaluation of a particular program or agency have? What benefits does evaluation provide to those associated with it? In view of the declining resources available for adult education and the increasing demand for services, adult educators must be assured that scarce resources devoted to evaluation activities will provide an equitable return in terms of benefits to administration, policy-making, and planning of adult education programs. Evaluation activities, if properly designed and executed, can provide such benefits and more.

As available resources for adult education and other areas of education dwindle, both agencies and programs will come under increasingly closer scrutiny with respect to their effectiveness. Any program or agency which may be called upon at some time or other to justify its existence must rely on more than just impressions and attitudes. It must have hard evidence to demonstrate the value and worth of its operations. Its services must be demonstrated to have a beneficial effect upon its intended clientele. In gathering and reporting such evidence, the program or agency must use some type of evaluation process. Thus, evaluation, in one form or another, can be said to be basic to the existence of any program or agency.

There are other reasons why an organization might wish to evaluate the programs it operates.

- To fulfill local, state, or federal mandates resulting from regulation or legislation.
- To determine the extent to which an existing program is accomplishing its objectives.

- To identify the strengths and weaknesses of a program.
- To determine the cost/benefit of an existing program.
- To establish a data base which management can use to make decisions on the productivity and efficiency of a program.

This guide is designed to help educators plan and carry out evaluations of adult education programs. It describes some of the more common approaches to program evaluation which can easily be adapted to and implemented by adult education programs.

The remainder of the guide is organized in the following way.

- Section 2 on evaluation and program development discusses how the evaluation process fits into the concept and normal cycle of program development.
- Section 3 describes the stages of the evaluation process including needs assessment, program planning, process evaluation, and outcome evalution.
- Section 4 on evaluation design provides examples of different evaluation designs which can be used in adult basic education. These include case designs, time series designs, and comparison group designs.
- Section 5 discusses various methods of data collection which can be used for evaluating adult education programs. The methods described are questionnaires, interviews, observations, and tests.
- Section 6 describes statistical procedures for analyzing evaluation data including descriptive statistics, tests of difference, and tests of relationships.
- Section 7 illustrates the preparation of the program evaluation outline which is a device to help program staff sketch out the essential elements (objectives, measurement techniques, time schedule, evidence of accomplishement) of the evaluation plan.
- Section 8 on determining program effectiveness outlines the different factors which objectively determine the effectiveness of a program.
- Section 9 on reporting evaluation results describes various reporting and dissemination techniques applicable to adult education programs.

- A list of references is included so that more detailed information on the different aspects of program evaluation can be acquired.

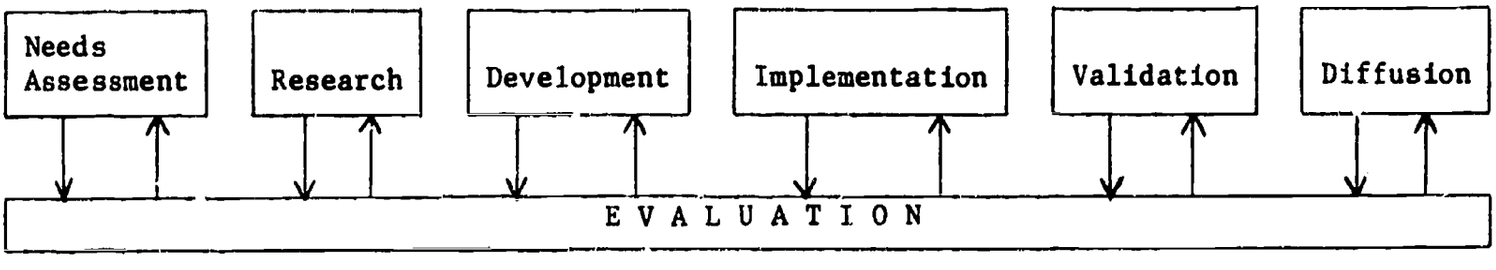
2. EVALUATION AND PROGRAM DEVELOPMENT

The term "evaluation" refers to the process of selecting, collecting, and interpreting information needed for decision making. It is essentially research applied to decision making. As such, evaluation goes hand-in-hand with program development. "Program development" is defined as a systematic process for creating or modifying methods, procedures, and/or materials to be applied toward the achievement of certain specified objectives.

Most successful programs follow the same cycle over the course of their development. Underlying the entire program development cycle is the evaluation process. Evaluation serves to provide objective information and feedback to each step within the program development cycle. This feedback enables program developers and administrators to make intelligent decisions about how program development should proceed. For example, it answers the questions of whether greater resources should be devoted within the same step of the program development cycle and when progress toward the next or succeeding steps should be made. The steps in the program development cycle, and examples of evaluation support relevant to each step, are described below.

Figure 1

Program Development Cycle



- Needs assessment -- Program development is usually initiated by some form of needs assessment, either formal or informal, by the local agency. Evaluation support at this stage involves the gathering and analysis of information related to program needs.

- Research -- Once the need for the development or modification of a program is established, some initial research is undertaken prior to any intensive development effort. This research may take the form of a review of the literature available in the area of interest, a mini-study conducted to test the feasibility of development of the intended program, or simply an informal survey of potential users of the program as to the practicality of undertaking the development effort. Evaluation support at this stage might be little more than providing a critical framework for analyzing previous research.

- Development -- Initial research is generally followed by an intensive development period in which the methods, procedures, and materials which are to constitute the principal elements of the program are created or modified based on an existing program. Evaluation support at the development stage may consist of little more than monitoring the timeline for program development and providing expert and consumer reviews of processes and procedures resulting from the development.

- Implementation -- Once development has been completed, the program is then implemented on a trial or pilot basis and closely scrutinized with respect to its operational effectiveness. At this stage, evaluation activities consist of implementation and progress evaluations of how the program is being implemented and the quality of implementation.

- Validation -- Results of the program are next validated by some systematic process in which the effects or results of the program are related to the original objectives upon which its development was based. Some objective method is used to determine that the program as developed and implemented meets the initial objectives for its development and hence satisfies the original identified needs of the organization. At this stage, evaluation activities would involve setting up an outcome evaluation design to assess the degree of effectiveness of the program in achieving its objectives.

- Diffusion -- Once validation has been accomplished, the program then undergoes diffusion. Diffusion is simply the process of disseminating information about the program to other parts of the same agency, or other agencies

reflecting similar needs, and the provision of technical assistance necessary for replication of the program at other sites within the same agency or within other agencies. Evaluation support at this stage might consist of assessing the degree to which the dissemination of materials about the program, the training offered by program developers, the implementation of the program at other sites, and the outcomes achieved by the program at other sites were effective and, where relevant, were comparable to the original program results.

For more detailed information concerning the evaluation process and program development, consult: Alkin, 1969; Kent, 1973; Provus, 1971; Stufflebeam, Foley, Gephart, Guba, Hammond, Merriman & Provus, 1971; and Suchman, 1967.

3. STAGES OF EVALUATION PROCESS

The evaluation process can be categorized into four general areas or stages of activity: needs assessment, program planning, process evaluation, and outcome evaluation.

Needs Assessment

Needs assessment involves a determination of what goals and objectives a new program should try to achieve or how well an existing program is achieving its pre-specified goals and objectives. Information supplied through a needs assessment process can provide the basis for making decisions regarding the initiation of a new program or making changes to an existing program.

Needs assessment usually involves the following stages.

1. Determining needs.
2. Assessing the relative importance of those needs.
3. Evaluating the degree to which established programs can address those identified needs.
4. Deciding whether new programs should be initiated or established programs modified to address such identified needs.

Needs can be established through several sources including:

- o federal and state legislation, guidelines, and regulations
- o concerns at the local, state, or national level appearing in published editorials, articles, and legislation
- o parent and community concerns expressed at public meetings and conferences
- o statistics kept by local, state, and federal agencies
- o educational or job qualifications

- opinions of adult education staff and students.

The importance of these identified needs can be assessed through a survey aimed at a broad range of persons (e.g., adult education participants or the general public), or by a rating or ranking procedure using a small group of informed persons (e.g., staff of an adult education program). If a large survey is to be conducted, and there are very few content items, respondents may be called upon to rank order the problems or issues from most important to least important. If the number of content items is moderate to large, it is best to adopt a procedure which involves a five-point Likert scale with a range of response categories as follows: (1) very unimportant, (2) somewhat unimportant, (3) not sure, (4) somewhat important, (5) very important. A small group of people assembled for the purposes of rating potential problems and issues needing to be addressed may either rank order such needs or use a set of cards or envelopes to sort needs into categories. Suggested categories for this purpose are as follows: (1) unimportant/inappropriate/irrelevant, (2) below average importance, (3) average importance, (4) above average importance, (5) very important/critical/essential.

It may also be desirable to survey or sample several different groups of individuals. If more confidence is placed on the opinions of one group over another, differential weightings can be given to the results from each group. To accomplish this, the mean ratings of each group sampled are multiplied by the weighting value assigned to each group. The potential needs then receive a ranking based on the summed weighted means across all groups surveyed.

The important things to remember in assessing needs are: (1) the needs assessment effort should be appropriate to the size of the program intended, (2) the needs assessment should be broadly based, and (3) the needs assessment should involve a survey or sensing of needs among individuals, persons, or groups expected to benefit from the program. For more detailed information on needs assessment, consult: Scriven & Roth, 1978; Spitzer, 1979.

Program Planning

Once needs have been identified and prioritized, program planning can begin. An important first step in program planning is translating identified needs into a set of goals and objectives. Goals are rather broad and general statements of what a program is intended to achieve. Objectives, on the other hand, are more specific statements describing in detail how portions of goals will be achieved. Depending upon the identified needs and the priorities set among such needs, it may be advisable to outline the general overall goals of a program before specifying its objectives. This is especially true for more complex programs. For simpler programs, it may suffice to directly state the objectives. Ideally, the objectives should be stated in measurable form. Figure 2 illustrates a set of goals and measurable objectives specified for an adult education program funded under Section 310 of the Adult Education Act.

Figure 2

Program Goals and Objectives

Goals	Objectives
1. To develop a basic skills instructional model addressing occupationally-referenced skills necessary for success in entry level and semi-technical occupations.	a. Develop a list of occupations. b. Identify and sequence skills and specify instructional objectives for each skill. c. Select or develop instructional materials to accomplish objectives. d. Develop an assessment system for student placement within the skill sequence and for measuring mastery. e. Develop a record-keeping system (student prescription form) to document program progress.
2. To implement the model with 40 unemployed adults.	a. Forty adults will receive instruction referenced to their occupational goals and skill levels. b. Each trainee will achieve 85% mastery of a minimum of 80% of the assigned skill sequences.
3. To document the effectiveness of the model.	a. Fifty percent of the trainees will be placed on jobs consistent with their occupational goals. b. Eighty percent of the placed trainees will maintain their jobs for at least three months.
4. To document and disseminate the model.	a. Develop a manual which describes the model. b. Distribute manual to 100 relevant agencies. c. Conduct three workshops to train teachers in the use of the model.

As can be seen in Figure 2, four broad goals were specified for the program (model development, implementation, evaluation, and dissemination). Each goal was then translated into several specific measurable objectives which described how the goal was to be achieved.

In addition to specifying goals and objectives, program planning also involves the development of a management plan that describes the tasks to be completed, the people responsible for carrying out the tasks, and the timelines for task completion. Systematic charting techniques, such as PERT or GANT, can be used to graphically display the management plan. However, for most adult education programs, a standard timeline such as the one in Figure 3 can be used. This timeline corresponds to the adult education program whose goals and objectives are listed in the previous figure.

Figure 3

Program Timeline

Activities	Objectives	Monthly Timeline 1	2	3	4	5	6	7	8	9	10	11	12
Goal 4. To document and disseminate the model													
1. Systematically record the instructional skills sequences, the instructional objectives, the assessment instruments (survey and mastery tests), record-keeping proceudres, and operational accomplishments.	a.				———	———	———	———	———	———			
2. Write the complete manual describing model content and process	a.						———	———	———	———			
3. Identify 100 relevant agencies and distribute manual	b.									———	———		
4. Design a teacher training workshop with accompanying materials	c.								———	———	———		
5. Specify training sites and conduct training	c.									———	———	———	
6. Design a workshop evaluation instrument.	c.									———	———		
7. Administer instrument.	c.											———	
8. Analyze and report evaluation results.	c.												———

Another important part of program planning is the development of an evaluation plan. This plan specifies the evaluation design, the instrumentation, and the analyses that will be used to assess program effectiveness. The individual components of the evaluation plan (i.e., evaluation design, instrumentation, and data analyses) are discussed in subsequent sections of this guide.

Program planning should be reflected in a written plan that includes: (1) a statement of the needs to be addressed, (2) a review of literature and relevant research, (3) a statement of goals and objectives, (4) the management plan (who does what, when, and to whom), (5) the evaluation plan (design, instrumentation, analyses, report timelines), (6) personnel, (7) facilities, and (8) budget. For programs that represent continuations of previous program efforts, the needs should reflect feedback on prior years' program performance and effectiveness. That is, new goals and objectives should reflect priorities in improving the program or specific aspects of the program. For more detailed information on program planning, consult: Klein & Alkin, 1971.

Process Evaluation

The conduct of a process evaluation generally consists of two major tasks: (1) describing and documenting program activities in an ongoing manner, and (2) assessing the extent to which intended activities or procedures have been implemented as planned. Other terms frequently used to address process evaluation concepts are implementation evaluation, progress evaluation, and program monitoring. All of these are included within the general concept of process evaluation.

In terms of documentation, process evaluation provides a formal, lasting description of what the program looks like in actual operation. A process evaluation report should provide an accurate account of the program in sufficient detail so that others who may want to replicate or adapt the program can use the program description as a basis for planning. The program description should address all of the key elements that were implemented as well as other variations that might have been followed. Factors such as program context, students served, and intervention strategies should be included.

In terms of program context, the process evaluator should describe the program setting, including historical antecedents, organizational structure, and physical facilities. Student descriptions should include participant background and characteristics as well as recruitment and selection procedures. Likewise, staff characteristics and selection criteria should be described. Descriptions of intervention strategies utilized should focus on learning resources and materials, learning environments, instructional content, types of instructional activities, and duration and intensity of program services. Depending on the specific nature of the program, there may be several other important program components which bear focusing upon during the process evaluation.

The overall description should also indicate implementation problems, components that are easiest and most difficult to implement, length of time required to achieve demonstrated impact, implementation demands, program costs, and all other factors involved in replicating the program. This detailed description of the program can also be used to determine the

extent to which proposed activities have actually been carried out at the original site or a replication site.

Process evaluation can also provide information on whether or not the program is being implemented according to plan. Although it is often neglected in program evaluation plans, process evaluation is just as important as outcome evaluation in this regard. Program staff with the day-to-day responsibility for running a project can especially benefit from process evaluation findings. By providing process evaluation information on a continuing basis, the evaluator gives the program staff rapid and direct feedback concerning implementation of program activities. This information can be used as a reliable decision-making tool for the initiation of mid-course operational corrections.

Process evaluation likewise provides program planners and administrators with a monitoring system that can be used as the basis for making conceptual and operational programmatic changes. These decisions are different from the day-to-day decisions or changes that operational staff might make on the basis of the on-going process documentation system. Administrators need to determine if the program is evolving as they had anticipated. It is particularly important for innovative demonstration programs that often, in practice, do not look as they were originally described in a concept paper or in a grant proposal.

The ability to make such conceptual or programmatic changes based on process evaluation information may be of particular value to programs involving training of teachers as a prerequisite step in reaching certain eventual student outcomes. In this case, the training workshop must effectively convey the information necessary to later providing the program

to students. The training must be effective. Evaluation of the training sessions will yield clues to the project staff as to what aspects of training will need to be revised or modified in order to enhance its effectiveness. Samples of types of instruments that can be used for this purpose may be found in the appendix to this guide.

Measuring program implementation is also important for the purpose of accountability. Program sponsors and administrators need to know that proposed activities are being carried out. In other words, program staff are accountable in showing where the program resources are spent. Comprehensive documentation of ongoing operations provides this assurance and also lends backup support for any deviations in proposed plans of activities. For example, staff can show evidence to the sponsors concerning why a planned activity had to be changed or deleted if there is adequate process documentation.

Monitoring of process objectives may sometimes be an end in itself. For example, an outcome of a small adult literacy project in the dissemination/diffusion stage of its development might be to create program awareness for educators throughout the state. A process objective, designed to move educators toward achievement of this outcome, might be to conduct a comprehensive statewide publicity campaign. While program staff might be held accountable for accomplishing the publicity campaign activity, it would not be practical or feasible to administer a state-wide survey to document the intended awareness outcome. Detailed documentation of this publicity campaign would be a more efficient way to certify that the activity had been effectively carried out.

Finally, process evaluation is important in that it provides a context for understanding the extent to which intended program outcomes may or may not occur and why. Frequently, actual implementation of a program does not directly coincide with the project description included in a program plan or proposal. If, for example, a project has failed to carry out several essential activities, it is less likely that the intended results of the program will be realized. In such a case, if process evaluation does not accompany the outcome evaluation, one might mistakenly conclude that the programmatic approach has failed, whereas in reality it was never really tested, for it was not implemented as intended. Thus, an understanding of how specific measured outcomes are related to program procedures and activities greatly enriches the usefulness of an outcome evaluation.

The four steps involved in designing and conducting a process evaluation are discussed below.

Developing Measurable Process Objectives

It is important to stipulate measurable process objectives for all critical elements of the project. The evaluator should make sure that there is at least one process objective that corresponds to and enables accomplishment of each project goal and outcome objective. Process objectives can be stipulated in terms of program context and background as well as salient components and activities. These can be based on the program plan or proposal, expert opinion, or one's own observations. Program staff might find it helpful to state the process objectives in terms of evaluation questions. For example, one question might be "Were adequate and appropriate staff training opportunities effectively

provided?" Wherever possible, intended success criteria should be established.

Preparing Appropriate Measurement Approaches

A variety of information sources can be used to measure process evaluation objectives. Three common approaches are: examination of documentation/records, conduct of systematic observations, and use of self-report measures. Examples of documents/records include: project plans or proposals, policy and procedures manuals, operational files (e.g., student background characteristics, teacher logs, attendance records), curriculum scope and sequence charts, course syllabi, textbooks, descriptive brochures and public relations pieces, periodic research and evaluation reports, and so on. Existing records should be used to the extent possible, but the project's management information system should be thoroughly reviewed to make sure that all essential records are systematically and accurately being collected.

Direct observations of program activities can provide a valuable picture of the context and the dynamics of the implementation process. Observations can be formal (i.e., structured) or informal, but in either case the observer should carefully document the events that occurred during the observation visit. In a structured observation, the observer typically uses a checklist or outline of key program features to focus attention.

Self-report measures of program implementation include questionnaires, checklists, and interviews. For example, questionnaires can be used following a training event or at the end of the program year to assess participants' perceptions of programmatic components and the success of

implementation. Interviews can be used to collect in-depth information on program operations. More detailed information on measurement techniques is provided in the section on methods of data collection.

Collecting and Analyzing Process Data

The process evaluation data collection system should be sufficiently open-ended that unanticipated activities or events can be documented. Multiple data collection strategies should be employed wherever possible in order to insure the accuracy, validity, and credibility of the information.

Process evalution data can be collected periodically during operation of a program or retrospectively. The former is preferred, wherever possible, to insure that valuable data will not be lost, to tap participants' memories while they are fresh, and to provide timely feedback to program staff and managers. Also, it is important to describe the status of participants or activities on a pre-post basis wherever possible. For example, teacher behaviors before and after a training event can be described.

Analysis of process evaluation data is almost always descriptive. In some cases, narrative descriptions will be appropriate for analysis of qualitative data, whereas categories or themes can be derived to transform the information into quantitative data in other cases. Usually frequencies, percents, and simple descriptive statistics (e.g., means and standard deviations) are sufficient for reducing quantitative process data. More detailed information on statistical procedures for analyzing evaluation data is provided in a later section on statistical analysis procedures.

Reporting Process Results

Reporting of process data will depend upon the purposes of the evaluation and the intended audience. For example, program staff may require timely, informal feedback, whereas policy-makers would need detailed summaries of the entire project in order to interpret process data.

Morris and Fitz-Gibbon (1978) provide a more detailed discussion of procedures for process evaluation.

Outcome Evaluation

Outcome evaluation measures how effective a program has been and estimates what effect it may be expected to have in the future. Outcome evaluation data will support decisions to continue, expand, terminate, or modify a program. Outcome evaluation goes hand-in-hand with a thorough process evaluation.

An outcome evaluation can focus on several different types of questions. Frequently, the primary aim is to determine how well a program's goals have been met. Other types of outcome evaluation studies focus on the comparative value of a program in relation to alternative approaches, on its side effects, and/or on its cost effectiveness. Finally, some outcome studies involve the determination of programmatic impact for special groups (i.e., differential effectiveness). Before proceeding with an outcome evaluation, all parties involved must have a clear understanding of the evaluation purpose and the types of decisions that will be made on the basis of outcome evaluation findings.

The specific design and procedures for an outcome evaluation depend on the evaluation purposes. There are six basic steps involved in conducting

any outcome evaluation study. These are briefly described below. Subsequent sections take up several topics in detail.

Specifying Outcome Objectives and Evaluation Questions

The program evaluator should understand the program's goals and activities in order to specify accurately the observable, intended consequences of the program operations. The evaluator must make sure that the list of objectives is complete, reasonable (i.e., outcome objectives should follow from program activities), and measurable. Evaluation plans should indicate how the evaluator will deal with objectives that are not easily measurable (e.g., infer accomplishment by documenting corresponding process objectives). If there are resource or feasibility constraints, the evaluator may have to prioritize objectives within the complete list.

Most outcome objectives will reflect program impacts on participants (i.e., students and staff). Typical outcome areas for adult basic education programs include improvements in cognitive skills (e.g., reading level, writing, staff knowledge, school completion), affective skills (e.g., self-concept, confidence, aspirations, interpersonal), life skills (e.g., consumer, family, personal, civic), or job improvement skills (e.g., employment, promotion, earnings, employability). Each of these areas should be considered with other potential areas of program impact, including unplanned outcomes. "Non-participant" outcomes might include costs, products, policies/procedures/practices, or social benefits.

Evaluation questions are derived from outcome objectives through an exploration of content and implications. Each objective will result in at least one evaluation question. Many objectives, upon exploration, will

require more than one question to represent their intent in the evaluation. One helpful way to go from objectives to specific questions is to determine what kind of information would be accepted as convincing evidence of the program's merit with regard to the objective. This kind of information can be easily transformed into questions about how well the program's goals were achieved. In addition, this process will suggest standards or criteria that can be used to judge the evidence and, thus, provide answers for the questions.

Constructing an Evaluation Design

Constructing an evaluation design involves the specification of a plan that indicates how information will be collected, analyzed, and reported during the course of the evaluation. Evaluation designs are discussed in more detail in a later section on evaluation design.

Planning Information Collection

Planning information collection involves the specification of efficient procedures for answering the evaluation questions, including the designation of evaluation measures or instruments and the logistics for data collection. The correspondence between evaluation questions and the selected measurement instruments is critical. Too often an inappropriate measure is administered, and it should not be surprising in such a case if results are not positive. Following are some basic principles for preparing an effective data collection system.

- Make sure that the measurement instrument is directly linked to the evaluation question for which it is being used.
- Consider alternative and/or multiple measurement strategies.

- Make sure that each measure is appropriate for its audience.
- Make sure that the instrument is as valid (i.e., measures what it is supposed to measure) and reliable (i.e., consistent) as possible.
- Make sure that data collection is as unobtrusive as possible.
- Make sure that the instrument is sensitive to positive as well as negative outcomes.

Collecting Outcome Evaluation Information

Major techniques for collecting outcome evaluation information include achievement tests (i.e., norm-referenced and criterion-referenced, commercial and teacher-made), performance tests (e.g., work samples, process tasks), questionnaires and surveys, interviews, checklists and rating scales, logs and journals, observations (structured and unstructured), and program records (e.g., attendance, timesheets, student progress). Measurement instruments can be selected, constructed, or adapted. Wherever possible and appropriate, existing measures are recommended. The evaluator may have to exercise a great deal of creativity in assessing certain types of program outcomes. For example, a sample follow-up survey of teacher and student behavior related to use of a life skills program is provided in the appendix to this guide. More detailed information on measurement techniques is provided in a later section on methods of data collection.

Evaluation data should be collected as systematically as possible. Standardized instructions should be prepared and all information collectors should be adequately trained in administration procedures. Potential sources of bias should be minimized. Communications among groups are extremely important, and perceived threats and anxiety should be minimized. All participants, staff, and data collectors should be thoroughly informed of the purpose, nature, and schedule of data collection activities. The evaluator should closely monitor the data collection process to insure the reliability and validity of results.

Analyzing Evaluation Information

Data analysis involves summarizing and synthesizing the information collected to find the answers to evaluation questions. This includes preparing the data for analysis (e.g., coding, tabulating, computer processing), applying the appropriate analytic methods (i.e., qualitative or statistical), and interpreting the results. Analysis issues are discussed in more detail in a later section on statistical analysis procedures.

Reporting Findings

The final step in conducting an outcome evaluation is reporting the evaluation results. The nature, frequency, and format for evaluation reports will depend on the evaluation purpose (i.e., decisions to be made) and audience (i.e., level of detail, technical vs. nontechnical, etc.). Usually, outcome evaluation reports are prepared on an annual basis or at the end of the project. The report should be written so that each specific audience can clearly understand the results of the program and their implications, and so that they will have confidence in the conclusions and

recommendations. Although the style and co·:ent of the report will depend upon the situation, an outline of a typical outcome evaluation report format is described below. Evaluatic· reporting is described in more detail in a later section on reporting evaluation results.

Sample Evaluation Report Outline

- Summary (overview of entire report, presented first but written last)
- Purposes of evaluation (objectives of study, decisions to be made)
- Program description (program context and activities, process evaluation results)
- Evaluation methods and procedures (evaluation questions, design, instruments, data collection, analysis plan, and design limitations)
- Results (summary of "hard" and "soft" data corresponding to each question, presented in tables, figures, and charts, and, wherever possible, statistical tests)
- Discussion (implications of findings, explanations of results, limitations)
- Conclusions and recommendations (final summary of findings and suggestions for future actions/decisions)
- Appendices (supporting documentation, instruments, raw data, and lengthy information that, although pertinent, would interrupt the flow of the body of the report)

This section on outcome evaluation provided a brief overview of issues involved in conducting such an evaluation. Fink & Kosecoff (1977) and Morris & Fitz-Gibbon (1978) provide further details on the outcome evaluation process. Further assistance is recommended in dealing with the more technical details of the evaluation process.

4. EVALUATION DESIGN

An evaluation design is a plan that dictates when information is to be gathered (i.e., the timing of measurement), and from whom (i.e., subject groups), during the course of the evaluation. Typically, evaluation designs are associated with outcome evaluations since process evaluations tend to use descriptive, case study, non-comparison group strategies. Good evaluation designs will provide information that will help decision-makers draw valid conclusions about program outcomes. Through strong designs, the evaluator tries to minimize the effects of non-program related factors or threats (i.e., extraneous influences or biases) that might offer alternative explanations to observed results. Campell and Stanley (1966) list twelve factors that can threaten the validity of evaluation results. Figure 4 summarizes these factors. Figure 5 is based on the JDRP Ideabook, a publication of the Joint Dissemination Review Panel (Tallmadge, 1977). This figure presents other typical evaluation hazards that diminish the potential validity of conclusions.

The evaluator must select a design strategy that controls the most likely threats to validity. Still, it should be recognized that the specific methods selected will also be a function of the type of program being evaluated, available resources, practical constraints, and reasons for the evaluation (i.e., the kinds of decisions for which the information is needed). Three basic categories of evaluation designs, adapted from Campbell and Stanley (1966) and similar handbooks (e.g., Klein & Burry,

Figure 4

Factors Threatening the Validity of Evaluation Data*

Threat	Description
1. History	Historical threats are outside influences in the environment (i.e., not part of the program) that may affect outcomes. For example, teachers enrolled in a staff development program designed to change attitudes toward adult illiteracy may be affected by extensive publicity from a national report.
2. Maturation	Program participants change due to normal physical and psychological growth. For example, a student's improved social adjustment skills may have nothing to do with the ABE program.
3. Testing	Taking a pretest may have subsequent effects on a participant's posttest performance. For example, a mathematics pretest may give students practice and familiarity with test items and improved posttest scores may not be due to the instructional program.
4. Instrumentation	Instrumentation threats are due to changes in instruments, observers, interviewers, raters, or scoring procedures from one time to the next. For example, a student may respond differently to the personal characteristics of two different interviewers on pre and post surveys.
5. Statistical Regression	This phenomenon refers to the tendency of very higher or low scores on a measure to move toward the average on subsequent testing. For example, if a student scored at the 5th percentile on a reading pretest, he/she is likely to score slightly higher on subsequent testing regardless of the ABE program.
6. Selection	There may be inherent differences between groups getting different treatments before the treatments begin. For example, the motivation of participants who voluntarily enrolled in an ABE program may be very different from those who were required to enroll.

*From Campbell & Stanley, 1966.

7.	Experimental Mortality	Participants in a program at the outset may drop out or move away. In programs with variable entry and exit, typical of many ABE programs, it may be very difficult to obtain posttest scores from participants who were pretested. Thus, the evaluation sample with complete sets of test scores may not be representative of the entire program group (e.g., may end up with only the highest or lowest scoring students).
8.	Interaction	Sometimes two or more of the above factors may operate together to produce an effect that neither one could do alone. For example, the chemistry of selection of a particular group of adult learners and their maturation may show changes not necessarily caused by the program.
9.	Reactive Effects of Testing	A pretest may increase or decrease participants' sensitivity or responsiveness to certain elements of a program. For example, teachers in a staff development program might pay particular attention to concepts covered in a pretest.
10.	Selection-Treatment Interactions	The interaction of selection and program treatments may cause biased results. For example, a successful urban program might not be successful in a rural setting.
11.	Reactive Effects of Innovation	This phenomenon is also known as the Hawthorne effect. Participants may perform better simply because they know they are involved in an experimental program and an evaluation study.
12.	Multiple Program Interference	This occurs when students are involved in two or more programs that might have joint effects. For example, a participant enrolled in an ABE class may also be receiving help from a private tutor on the side. It would be difficult for the evaluator to separate the effects of the two treatments.

Figure 5

Common Evaluation Hazards*

1. Claiming much, providing evidence of little.
2. Selecting measures not logically related to the program.
3. Use of grade-equivalent scores.
4. Use of a single set of test scores for both selecting and pretesting participants.
5. Use of comparisons with inappropriate test dates for obtaining information.
6. Use of inappropriate levels of tests.
7. Missing data.
8. Use of inappropriate statistical adjustments with nonequivalent control groups.
9. Constructing a matched control group after the treatment group has been selected.
10. Careless collection of data.
11. Use of different instruments for pretesting and posttesting.
12. Use of inappropriate formulas to generate no-treatment expectations.
13. Mistaken attribution of causality.

---

*Based on the JDRP Ideabook.

1971; Fink & Kosecoff, 1977; Morris & Fitz-Gibbon, 1978; Kershner, 1976) are briefly described below. These references can be reviewed for more detailed discussion. The categories are arranged in order from the weakest to the strongest design in terms of validity. The stronger the design, the more confidence can be placed in the conclusions. If the evaluator has a choice, the latter designs (e.g., comparison group) are usually preferred over the former designs (e.g., case designs).

Case Designs

Case designs are used to examine a single coherent group of participants (i.e., the program group). Campbell and Stanley refer to them as "pre-experimental" since they do not involve controlled investigations of effects, yet they can suggest the probable existence of certain outcomes (although they do not confirm the outcomes). These designs can be used for descriptive or exploratory purposes, but should only be used as a last resort for rigorous evaluation studies. Case designs are particularly vulnerable to many of the validity threats noted in Figure 4. Among the most common of these threats are history, maturation, instrumentation, mortality, the Hawthorne effect, and selection bias.

However, in many situations the case designs may represent the only feasible option for the evaluator. If this type of design must be chosen, then as much supporting evidence as possible should be assembled to suggest that results were influenced by the program itself and not by extraneous non-program factors. Consistency of findings, the quantity and quality of information sources, and rational explanations for competing alternatives

will help to build confidence in program impact. Replication of evaluation study results is another powerful strategy for confidently establishing program impact. Extensive process evaluation data should be collected, reported, and related to outcome findings. Correlation analysis techniques (e.g., multiple regression) may be helpful in demonstrating relationships, even though these techniques cannot demonstrate causation.

Three specific types of case designs are discussed below. The first two examples are very weak and are not recommended.

Unassessed Treatment

Actually, this approach is really not a design at all, and it represents the weakest case in the validity continuum. Here, the project administrator intuitively decides whether the program was effective or not. It should be obvious that these kinds of subjective inferences about program merit without objective assessment are dangerous because they are likely to be heavily influenced by personal factors and occurrences rather than by what actually happened.

One-Shot Case Study

This design involves measuring the group's performance at the end of the program. A criterion-referenced test, a performance test, or a self-assessment questionnaire might be administered. Although this approach is better than not measuring at all, there are numerous competing factors (e.g., see the list of threats in Figure 4) that may account for the results and, thus, make it impossible to discover the contribution of the program to the observed outcomes. For example, students might have done just as well on a pretest given before the program.

One Group Pretest-Posttest Design (Before-and-After Design)

This design is a slight improvement over the preceding one because of the addition of a pretest. Like the one-shot case study, it is a relatively weak design and should be used only when stronger designs are not possible. Results are descriptive of what happened, but they may not necessarily be attributed to instruction in the adult education program. As Kent (1974) noted in his national longitudinal study of adult education programs, changes from pretest to posttest in this type of design are "results" in the sense that they "resulted." There may be several plausible explanations for these pre-post changes, only one of which includes participation in the instructional treatment.

Time Series Designs

Time series designs involve collecting data from one or more groups at regular intervals before a program begins, during the program, and after it ends. Usually, at least three pre and three post measures are suggested for effective time series studies. Measures over time should be identical, or at least parallel. Time series studies are used to examine how a group's current performance compares with prior performance. They are also used to determine retention and the durability of program effects. For example, employment and earnings variables lend themselves easily to time series designs. A program participant's employment status, reported at six month intervals, can be compared before and after involvement in the adult education program. However, time series designs are often difficult to implement for out-of-school populations, since it is hard to keep track of

program participants over a long period of time and longitudinal follow-up can be very costly.

Usually, results of time series studies are examined by graphing or plotting the repeated measures over time. Statistical analyses can be very complex and technical.

The most prominent threats to the validity of time series designs are history and, to a lesser degree, instrumentation. Mortality can be a threat if small samples are involved. In addition, reactive effects of repeated testing, selection bias, the Hawthorne effect, and multiple program interference may limit the generalizability of study results using this type of evaluation design.

Comparison Group Designs

Comparison group designs are generally recommended for outcome evaluation studies. In this strategy, outcomes for the program group are compared with outcomes on identical measures for others in alternative program groups or no treatment groups. Comparison group designs are frequently divided into two categories: "quasi-" and "true-" experimental designs. The categories are distinguished by the way that group membership is determined. Both types are briefly described below.

Quasi-Experimental Design

This design typically compares results of two groups which are selected "intact" for the evaluation study. For example, in examining the impact of a tutor-based, volunteer literacy program, test scores of the treatment group might be compared with results of an existing adult

education evening class operated by the local school district. It is essential that pre-post measures be administered in this design. For meaningful comparison, the two groups must be fairly similar with respect to pre-treatment measures and background characteristics. For example, it would be foolish to compare test scores for functionally illiterate adults with those of a freshman college class. The evaluator wants to be as certain as possible that any differences in posttest results are due to program differences and not due to background characteristics. Occasionally, where no intact groups are available, evaluators will attempt to construct an artificial comparison group by matching characteristics of program participants. However, reasonable matching is very difficult and hence, the practice is not usually recommended.

Threats to validity of quasi-experimental designs will depend on the rigor with which they are constructed. Selection and mortality (i.e., differential attrition) are the most likely threats. Other factors such as testing and instrumentation can be problems if the evaluator is careless. Selection bias, the Hawthorne effect, and multiple treatment interference also represent possible threats. To help eliminate these factors, the evaluator must insure that the design is adequately implemented over the course of the study. Results can be analyzed using fairly simple descriptive and inferential statistics.

True Experimental Design

This design is similar to the quasi-experimental design, except that individuals are randomly assigned to program and comparison groups. The comparison group, in this case, is referred to as the control group. Random assignment results in groups that are initially as similar as

possible, and any observed differences can be attributed to participation in the program. Hence, posttest measures are sufficient to determine differential impacts. A pretest-posttest schedule can be used to increase the precision of measurement, and may particularly be helpful if sample attrition is substantial. In this case, the pretests of remaining members of each group can be examined to determine the effect of such attrition on comparability of the groups. This is a very powerful evaluation design and is recommended wherever possible. In practice, however, random assignment is often difficult to achieve.

Threats to validity are minimal with true experimental designs if they are adequately implemented (e.g., low experimental mortality). Selection bias and the Hawthorne effect may be possible limiting factors.

The designs described above have been portrayed in simplest form. The particular nature of the program being evaluated will determine the specific kind of design that is required. Complex programs will require complex designs. These designs are the building blocks that can be adapted for most evaluation studies. Multiple design strategies may need to be employed to address all evaluation questions in an impact study. Selection of an evaluation design involves a number of considerations including time and resources. The cost of planning and implementing evaluation designs can range from 5 percent to 10 percent of a program's budget.

5. METHODS OF DATA COLLECTION

There are four basic criteria for selecting methods of data collection.

1. Is the method agreeable or acceptable to all participants?
2. Will the data collected be relevant, valid, and reliable?
3. What are the resource requirements in terms of people, time, and money?
4. What knowledge and skills are necessary to select or develop instruments, and to collect and analyze data?

Instruments selected or developed to measure progress on program objectives should be both reliable and valid. The term reliability refers to the consistency with which a particular instrument measures. While there are different types of reliability (e.g., test - retest, internal consistency, interrater) they all signify consistency. Instruments are given reliability "coefficients," depending upon how consistent they are, ranging from a low of zero to a high of one. Generally, good instruments have reliability coefficents of .70 or higher. Many standardized tests report reliability coefficients of .90 or better.

The concept of validity is used to reflect how well an instrument truly represents or measures the behavior it is supposed to measure. While there are different types of validity (e.g., face validity, intuitive validity, content validity, concurrent validity, predictive validity), all signify some type of evidence indicating the degree to which the instrument measures what it is supposed to measure. "Face" or "intuitive validity" indicates the degree to which a measure appears to be valid.

"Content validity" expresses the extent to which the items composing a particular measure are a representative sample of content from the domain they are claimed to represent, usually established by a panel of experts. "Concurrent validity" indicates the degree to which the measure in question is similar to other measures of the same behavior or characteristic. "Predictive validity" indicates the relative ability of the measure to predict relevant behavior in the future. As with reliability, validity "coefficents" range from a low of zero to a maximum of one. Respectable validity coefficients generally exceed .50.

While it is often difficult and costly to develop instruments from scratch, available instruments can usually be found which are suitable to meet most program needs. There are a large number of published instruments which have established validity and reliability data already collected on them. Such instruments are regularly referenced and reviewed in compilations of test instruments such as the Mental Measurements Yearbook and Tests in Print.

Of course, along with being both reliable and valid, any particular instrument used for evaluating programs must otherwise be suited to the program evaluation. That is, the instrument must be appropriate for the subjects to which it will be administered (e.g., age, sex), and to the situation or context in which the test will be used (e.g., group administration, individual administration). The test must also be appropriate to the kinds of interpretation one wishes to apply in the scoring. Some common measurement techniques for collecting evaluation data are described below.

Questionnaires

Questionnaires can consist of open-ended or closed items. Open-ended questions can provide more information but are more difficult to score. Closed questions usually take the form of a rating scale (e.g., five-point Likert scale) or a checklist. They are easier to complete and score, but the amount of information obtained from them is limited. Questionnaires can be administered to large groups at relatively low costs. Question development and formatting skills are necessary as is the ability to accurately interpret and categorize open-ended responses. Data from questionnaires are subject to many types of bias. People don't always answer items truthfully and items are often left blank, making follow-up necessary.

Interviews

Interviews can be structured or unstructured, face-to-face or by telephone, individual or group. Interviews can provide more in-depth information but are time-consuming. Interviewers must have special training so that appropriate, consistent, and accurate information is obtained, especially in unstructured types of interviews. The information obtained must be interpreted and classified.

Observations

Observations can provide first hand information, but such information might not be accurate. This is because the people being observed may not behave normally if they know they are being observed. To help alleviate

this problem, observations should be as unobtrusive as possible and should be conducted more than once. Several observers may be required to obtain accurate results. Inter and intra-observer reliability should be obtained. Observation forms must be developed, procedures planned, and observers trained, so that the information obtained is as consistent and accurate as possible.

Tests

Tests can either be developed or selected. Tests should be reliable and valid. They should be appropriate for the subjects being evaluated and to the kinds of interpretation the evaluator wishes to apply in the scoring. For example, in some cases it may be appropriate to use a "norm referenced" instrument and in other cases a "criterion referenced" measure. Scores on a norm referenced test would be interpreted relative to the scores of the group of individuals on which the test was first administered. In this case, scores in terms of "percentiles" or "normal curve equivalents" might be used and interpreted. The evaluator should make sure that the subjects getting the test and the "norm" group have similar characteristics.

In the case of "criterion referenced" testing, scores would be interpreted in relation to absolute performance standards that are pre-established. This type of testing often proves to be an integral part of adult education programs where instruction is designed to meet specifically diagnosed needs.

Other methods of data collection include document review or analysis, diaries or logs, checklists, inventories, and descriptive profiles. Figure 6 presents some commonly used measures of impact for adult education programs. For more information on instrumentation, consult: Covert, 1984; Demaline & Quinn, 1979; Klein & Niedermeyer, 1971; and Kornhauser & Sheatsley, 1976. For compendia of test instruments, consult: Buros, 1974; 1978.

Figure 6

Commonly Used Measures

Measure	Area Assessed	Source
Adult Basic Learning Examination (ABLE)	reading, math	Harcourt, Brace & Jovanovich, Inc. New York, NY
Adult Basic Reading Inventory (ABRI)	reading literacy diagnosis	Scholastic Testing Service, Inc. Bensenville, IL
Adult Performance Level (APL)	occupations knowledge, consumer economics, health, reading, writing, computation, problem solving	American College Testing Program Iowa City, IA
Career Development Inventory	education, occupation groups	Science Research Associates, Inc. Chicago, IL
Everyday Skills Tests	reading, math	CTB/McGraw-Hill Monterey, CA
Hall Occupation Orientation Inventory	affective personality variables	Scholastic Testing Service, Inc. Bensenville, IL
Informal Inventory for Low Literate Adults	diagnostic-placement assessment	Literacy Action, Inc. Washington, D.C.
Kuder Occupational Interest Survey	occupations	Science Research Associates, Inc. Chicago, IL
Kuder Preference Record -Personal -Vocational	work environments	Science Research Associates Inc. Chicago, IL
Life Skills	functional reading, math	Riverside Publishing Co. Boston, MA
Reading Free Vocational Interest Inventory	vocational interests	American Association on Mental Deficiency Washington, D.C.
Self-Directed Search	personality/activity variables	Psychological Assessment Resources, Inc. Odessa, FL

Measure	Area Assessed	Source
SRA Coping Skills	working, consumer economics, household management, health & safety, personal law, government, stress	Science Research Associates, Inc. Chicago, IL
SRA Reading Index SRA Arithmetic Index	functional reading, math	Science Research Associates, Inc. Chicago, IL
Strong-Campbell Interest Inventory	themes, interests, occupations	Consulting Psychologists Press, Inc. Palo Alto, CA
Tests of Adult Basic Education	reading, math, language	CTB/McGraw-Hill Monterey, CA
Work Values Inventory	attitude scales	Houghton-Mifflin Co. Boston, MA

6. STATISTICAL ANALYSIS PROCEDURES

Any evaluation of program effects involves summarizing or analyzing one or more sets of data. Carefully planned analyses facilitate clear interpretation of the evalution outcomes. Three categories of statistical analysis are briefly described below: descriptive statistics, tests of differences, and tests of relationships. This discussion is only intended as an overview of approaches, and statistical textbooks (e.g., Cook & Campbell, 1979; Guilford & Fruchter, 1973; Glass & Stanley, 1970; Hays, 1973; Siegel, 1956) should be consulted for details of computation and interpretation. Technical assistance and consultation may be required to perform complex statistical analyses.

Descriptive Statistics

The first step in the analysis of quantitative data is to "eyeball" the results to see if there are any peculiarities in the scores (e.g., extreme or impossible cases). Next, descriptive statistics should be computed to reduce the data to meaningful indices. Descriptive statistics summarize the distribution of scores (or other quantitative measures) among a particular group of subjects. They include measures of frequency, central tendency, and variability. These are summarized in Figure 7.

The first level of data analysis is always descriptive. Appropriate descriptive statistics should be displayed for each subject group on all available measures. These data depict group characteristics and suggest between-group differences that may need to be considered. Descriptive

Figure 7

Summary of Descriptive Statistics

Statistic	Description	Examples of Use
Frequency and percentage	Tally of number of cases in each data category; often requires specification of score intervals (e.g., scores reported in multiples of ten); percentage indicates proportion of cases in each score category or interval.	Useful in summarizing questionnaire item responses, test score results, attendance, etc.; can be displayed in tabular or graphic format or as histogram.
Mean (central tendency)	Arithmetic average of scores.	Average reading test score, number of instructional hours, hourly wages, etc.
Median (central tendency)	Middle score of a distribution; half of scores above and half of scores below; often used when there are a few extreme scores that might distort the mean.	Median annual income, age of participants, etc.
Mode (central tendency)	Most frequently occurring score.	Useful in reporting free-response data from questionnaires or interviews.
Range (variability)	Arithmetic difference between high and low scores.	Range of number of participant instructional hours.
Standard deviation (variability)	Often considered average distance from mean of scores in a distribution.	Measures of variability can be used for curriculum planning.

statistics may be sufficient to answer several evaluation questions or may suggest further statistical analyses that are required.

Tests of Differences

The evaluator may want to compare performances, based on the descriptive statistics. For example, the evaluator may want to compare the program group's average reading achievement posttest scores with their pretest scores and with the control group's posttest scores. These comparisons are accomplished through the use of inferential statistics. Inferential statistical procedures indicate the probability that an observed difference between performances is real (i.e., program-related) rather than due to chance factors.

Several commonly used statistical tests of differences are briefly described in Figure 8. The choice of a specific test will depend on the evaluation design, the questions being investigated, the assumptions made, and the type of data being collected. More information should be obtained before attempting to use or interpret these analytic techniques.

Tests of Relationships

Correlational techniques are used to describe the associations between variables or sets of variables. In evaluation designs, correlational approaches are valuable as descriptive techniques and may be very useful in understanding the complex nature of the variables that are being examined. Four statistics testing strength of relationships are briefly described in Figure 9. As a final note, however, it should be recognized that such approaches are non-experimental and causation can not be inferred from

Figure 8

Summary of Selected Statistical Tests of Differences

Statistic	Description	Examples of Use
Independent Sample t-Test	Used to compare mean performance of two-samples; tells if difference between groups is significant.	In true experimental design, used to compare reading test performance of program and control group.
Matched Group t-Test	Sometimes called correlated t-test, used to compare performance of same group of subjects measured on two occasions.	Comparison of program group's pretest to posttest change on a self-concept measure.
Analysis of Variance	Can be used to deal with data from two or more groups, or several independent variables (e.g., sex, age).	Analysis of performance test data, comparing program group, no treatment (control group), and alternative program group (comparison group).
Analysis of Covariance	Similar interpretation to analysis of variance, but can be used to control for any between-group differences that exist prior to beginning the evaluation study.	If mathematics pretest scores of control group are initially higher than program group (e.g., due to chance factors), analysis of covariance can be used to statistically adjust posttest scores, allowing a more accurate test of program effects.

Figure 9

Summary of Selected Tests of Relationships

Statistical Technique	Description	Examples of Use
Correlation: Product-Moment Coefficient	Measure of strength and direction of relationship between two variables; shows how variation in one variable is related to variation in another.	Can show strength of association between amount of instructional time and achievement scores.
Correlation: Rank Order	Same as above, except used for ranked (i.e., ordinal) data.	Can show whether student ranks in mathematics are related to ranks in reading.
Chi-Square	Shows relationship between categorical variables (e.g., frequencies); used to determine differences from unexpected proportions or in contingency tables.	Can be used to determine if absenteeism varies according to day of week; also, to determine if there are male/female differences in responses to questionnaire items.
Regression Analysis	Examines relationship between a dependent variable and two or more independent variables.	Can illustrate relationship between reading test score and number of hours of instruction, sex, and highest grade completed.

correlational analysis. For example, a correlational analysis can show that people who complete numerous years of schooling tend to earn the higher salaries, but it can't prove that they earn the higher salaries because of the extensive schooling.

7. THE PROGRAM EVALUATION OUTLINE

The program evaluation outline is a device created to help program administrators and staff sketch out the essential elements of an evaluation design for each objective specified for the program to be developed. A sample program evaluation outline format is shown in Figure 10 below.

This format allows for a listing of measurable "objectives" associated with the needs which the project was developed to address. For each objective, the "measurement techniques or instruments" to be used in collecting data on the objective can be detailed. In addition, the "data collection schedule" or manner in which such instruments will be used and such data will be collected can also be indicated. Finally, the "expected evidence of accomplishment" of specified objectives can be given. This outline, together with an ordinary program timeline, can be used to assess both the extent to which a project is achieving its intended objectives and the degree to which the project is operating according to schedule.

An example of a completed program evaluation outline for the adult basic education program outlined in Figures 2 and 3 is provided in Figure 11. As can be seen in the example, this particular program incorporates both teacher training and instruction of students as objectives. In addition, both process objectives and outcome objectives are mixed together under the objectives column. Many of the process objectives (e.g., development of the model, student instruction using model, teacher training), are measured via documentation of their completion. Outcome objectives (e.g., job placement and retention) are assessed by a form or questionnaire developed to measure the accomplishment of the objective.

Figure 10

Program Evaluation Outline

Page _____

NAME OF PROGRAM: _____

INSTRUCTIONS: Based on the needs identified for your ABE project, list the measurable objectives associated with these needs. For each objective, detail the measurement technique or instrument to be used in collecting data on the objective and the time or times at which such data are to be collected. Also indicate the expected evidence of accomplishment that such data will demonstrate with respect to the objective. Use additional sheets if necessary.

(1)	(2)	(3)	(4)
OBJECTIVE	MEASUREMENT TECHNIQUE OR INSTRUMENT	DATA COLLECTION SCHEDULE (PRE/POST, OTHER)	EXPECTED EVIDENCE OF ACCOMPLISHMENT

Figure 11

Program Evaluation Outline

Page _____

NAME OF PROGRAM: _____

INSTRUCTIONS: Based on the needs identified for your ABE project, list the measurable objectives associated with these needs. For each objective, detail the measurement technique or instrument to be used in collecting data on the objective and the time(s) at which such data are to be collected. Also indicate the expected evidence of accomplishment that such data will demonstrate with respect to the objective. Use additional sheets if necessary.

(1)	(2)	(3)	(4)
OBJECTIVE	MEASUREMENT TECHNIQUE OR INSTRUMENT	DATA COLLECTION SCHEDULE (PRE/POST, OTHER)	EXPECTED EVIDENCE OF ACCOMPLISHMENT
1. Develop the model			
a. Develop a list of occupations b. Identify and sequence skills and specify instructional objectives for each skill c. Select or develop instructional materials d. Develop an assessment system of placement and mastery tests e. Develop a record-keeping system	Project documentation	Continuous	Model completed by 5th month
2. Implement the model			
a. Forty trainees will receive relevant instruction	Project documentation	Continuous	Instruction completed by 12th month
b. Each trainee will achieve mastery on the relevant skill sequences.	Mastery tests	Continuous	Attainment of 85% mastery on 80% of the relevent skill sequences
3. Measure model's effectiveness			
a. Fifty percent of the trainees will be placed in jobs consistent with their occupational goals.	Job placement form	Upon completion of skill sequences	Completed job placement form
b. Eighty percent of the placed participants will maintain jobs for at least three months.	Job retention form	3 months after placement	Completed job retention form
4. Document and disseminate the model			
a. Develop a manual b. Distribute manual c. Conduct and evaluate three workshops to train teachers in the use of the model	Project documentation Project documentation Project documentation Workshop evaluation questionnaire	After 5th month 9th month Begin development 8th months; conduct training and administer 10th and 11th months; analyze questionnaire results 12th month	Manual completed by 9th month Distribution completed by 10th month Workshops completed by 11th month month; 80% positive response on workshop evaluation questionnaire

The criteria, or evidence of accomplishment of the objective, are also specified in the outline (e.g., completion by a specified time) as well as when the data measuring program effectiveness will be collected. The program evaluation outline is a helpful way for program staff to organize their evalution planning.

8. DETERMINING PROGRAM EFFECTIVENESS

Determining program effectiveness answers the question: "Was the program carried out according to plan and did it accomplish the objectives it was expected to accomplish?"

Program effectiveness can be determined in several ways. By referring back to the program evaluation outline, the program timeline, or a similar type of program planning vehicle, the evaluator can examine what program related processes were to be carried out, in what form, and at what time over the program period. The planned implementation of these aspects of the program can then be compared with actual implementation, as indicated from program documentation, in order to determine the extent to which the program was implemented according to plan. This information not only indicates the effectiveness of the management and the implementation of the program by program staff, but also provides a context for the interpretation of any program outcomes.

The objectives of the program presented in the program evalution outline can be compared with the actual results achieved by the program. The extent to which these outcome objectives are achieved is a measure of program effectiveness.

Although a program may be deemed effective with respect to its program evaluation outline, there are various degrees of effectiveness. These different degrees of effectiveness are a result of the rigor of the evaluation design applied to the program, the statistical significance of the outcome evaluation results, and the relative size of the audience to which the program is effective. On this latter dimension, the kind of

evidence needed to prove the effectiveness of a program for expansion on a limited basis is different from the evidence of program effectiveness necessary to justify the expansion of a program throughout the United States. A more rigorous way to measure the effectiveness of the adult education program illustrated in Figure 11 would have been to compare the job placement and retention data between a group of adults who received instruction using the model versus a similar group that did not.

Many evaluation designs are clearly weak, inappropriate, or ineffective in providing an accurate determination of how effective a particular program is with respect to a particular objective. While such designs may be useful at the local level in providing information on which local administrators can make decisions, they are generally not strong enough or rigorous enough to be accepted as evidence of effectiveness for a broader range of audiences.

Statistical significance does not by itself indicate program effectiveness. There is often a difference between statistical significance and practical or educational significance. Once a program's effectiveness has been documented through statistical significance, the educational significance of the program must also be determined.

Although there is a lack of firm measurement or statistical guidelines to determine educational significance, experts have generally recommended the following rule of thumb. Educational significance may be defined as the mean program to comparison group difference relative to the standard deviation of the scores within groups. A mean difference of about one quarter of a standard deviation would be interpreted as being of small educational significance, a mean difference of about half of a standard

deviation as being of moderate educational significance, and a mean difference of about three quarters or more of a standard deviation being of large educational significance. In addition, other factors such as cost of the program and potential benefits may be considered in making judgments about educational significance. However, there are no commonly accepted rules of thumb for these factors.

Those programs which have applied rigorous evaluation designs and have produced statistically and educationally significant effects may further demonstrate their effectiveness by applying for state or national validation. Validation involves submitting program evaluation results to a non-partial external review panel. The panel verifies that the program effects are reasonable and valid. Approval by such a review panel implies a "validation" of the project's effectiveness. The question of validation status becomes especially important when a completely developed program is to be disseminated to other educators and sites outside the area in which it was originally developed. Validation serves as a kind of quality control mechanism to ensure the effectiveness of programs to be disseminated and diffused to broader audiences.

Most states have developed state validation procedures for validating educational practices. They are usually less rigorous, less demanding, and less time consuming than the process of national validation. It is recommended that state validation be pursued as the initial course of action since it is easier to obtain and can be used as preparation for national validation.

National validation is usually sought after in preparation for national dissemination and diffusion of an educational program. National

validation is achieved through application to, and approval by, the Joint Dissemination Review Panel (JDRP). This is an internal review mechanism established by, and with representatives from, the U.S. Office of Education (USOE) and the National Institute of Education (NIE). This panel was established before the USOE and NIE were merged into the new Department of Education. Its function, however, has not changed. It serves essentially as a quality control mechanism for projects which intend to be disseminated and diffused nationally. It also serves as a gatekeeper for federal education funds for dissemination, since JDRP validation approval is the prerequisite for receiving these dissemination funds.

It is urged that projects make intelligent decisions about the types of evaluation designs and evidence of effectiveness needed to suit their own individual situations. Those projects with no intention of broader dissemination and diffusion should look to less costly evaluation designs, yielding evidence that may produce confidence in the effectiveness of the program locally but not necessarily to a wider audience. However, those programs looking toward eventual national dissemination and diffusion should plan their evaluation to adequately support their eventual validation plans.

9. REPORTING EVALUATION RESULTS

The evaluator must effectively communicate program results to the intended audience. In addition to the full evaluation report, there are several brief but effective ways that program evaluation results can be reported. These reporting strategies are described below.*

- Summaries - These provide brief but comprehensive information about the program. They usually include descriptions of program objectives, the evaluation design and results, and recommendations. Report summaries occur at the end of most final research reports. Executive summaries are separate reports or appear at the beginning of the full research report.

- Memos - These can be used to provide update information on program progress.

- Embedded quotations - These are portions of a report that are set apart within the text. They catch the eye of the reader and provide key information. More comprehensive information is available within the adjacent text. These embedded quotations, when read consecutively, can provide a complete overview of the report.

- Abstracts - These provide very brief but informative summaries of evaluation results.

For more detailed information on how to report evaluation results, consult: Klein with Burry & Churchman, 1971; Morris & Fitz-Gibbon, 1978.

---

* Adapted from Macy, 1982.

REFERENCES

Alkin, M. C. (1969). Evaluation theory development. Evaluation Comment, 2, 2-7.

Buros, O. (1974). Tests in Print. Highland Park, NJ: The Gryphon Press.

Buros, O. (1978). The Mental Measurements Yearbook. Highland Park, NJ: The Gryphon Press.

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

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

Covert, R. W. (1984). A checklist for developing questionnaires. Evaluation News. 5, 74-78

Demaline, R. E., & Quinn, D. W. (1979). Hints for planning and conducting a survey and a bibliography of survey methods. Kalamazoo, MI: Western Michigan University.

Fink, A., & Kosecoff, J. (1977). An evaluation primer. Washington, DC: Capitol Publications.

Glass, G. V., & Stanley, J. C. (1970). Statistical methods in education and psychology. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Guilford, J. P., & Fruchter, B. (1973). Fundamental statistics in psychology and education. New York: McGraw-Hill.

Hays, W. L. (1973). Statistics for the social science. New York: Holt, Rinehart, & Winston, Inc.

Kent, W. P. (1973). A longitudinal evaluation of the adult basic education program. Falls Church, VA: System Development Corp.

Kershner, K. M. (1976). RBS career education evaluation planning manual. Philadelphia: Research for Better Schools, Inc.

Klein, S. P., & Alkin, M. C. (1971). Program planning tools and procedures. In S. P. Klein with J. Burry & D. A. Churchman (Eds.), Evaluation workshop I: An orientation - Participant's notebook. Monterey, CA: CTB/McGraw-Hill.

Klein, S. P. with Burry, J., & Churchman, D. A. (1971). Evaluation workshop I: An orientation - Participant's notebook. Monterey, CA: CTB/McGraw-Hill.

Klein, S. P. & Niedermeyer, F. (1971). Clarifying objectives and planning data collection techniques. In S. P. Klein with J. Burry & D. A. Churchman (Eds.), Evaluation workshop I: An orientation - Participant's notebook. Monterey, CA: CTB/McGraw-Hill.

Kornhauser, A., & Sheatsley, P. B. (1976). Questionnaire construction and interview procedures. In C. Selltiz, L. S. Wrightsman, & S. W. Cook (Eds.), Research methods in social relations. New York: Holt, Rinehart, & Winston, Inc.

Macy D. (1982). Research briefs. In N. L. Smith (Ed.), Communication strategies in evaluation. Beverly Hills, CA: Sage Publications.

Morris, L. L., & Fitz-Gibbon, C. T. (1978). Program evaluation kit. Beverly Hills, CA: Sage Publications.

Provus, M. (1971). Discrepancy evaluation for educational program improvement and assessment. Berkeley, CA: McCutchan Publishing Corp.

Scriven, M., & Roth, J. (1978). Needs assessment: Concept and practice. New directions for program evaluation, 1, 1-11.

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

Spitzer, D. R. (1979). Critical issues in needs assessment. Paper presented at the annual convention of the Association for Educational Communications and Technology. (ERIC Document Reproduction Service No. ED 172 803).

Stufflebeam, D. L., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond, R. L., Merriman, H. O., & Provus M. (1971). Educational evaluation and decision-making. Phi Delta Kappan.

Suchman, E. A. (1967). Evaluative research: Principles and practice in public service and social action programs. New York: Russell Sage Foundation.

Tallmadge, G. K. (1977). The Joint Dissemination Review Panel Ideabook. Washington, DC: U. S. Government Printing Office.