ED 223 051 EC 150 353

AUTHOR Evans, Ellis
TITLE Program Evaluation in Early Childhood/Special Education: A Self-Help Guide for Practitioners. WESTAR Series Paper #13.
INSTITUTION Western States Technical Assistance Resource, Monmouth, Oreg.
SPONS AGENCY Special Education Programs (ED/OSERS), Washington, DC.
PUB DATE 82
CONTRACT 300-80-0753
NOTE 30p.
PUB TYPE Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Data Analysis; *Data Collection; *Disabilities; Early Childhood Education; *Evaluation Methods; *Program Evaluation

ABSTRACT

Intended for practitioners in early childhood special education, the document offers guidelines for conducting a program evaluation. Information is organized around seven questions: what is the purpose of the evaluation? what information is needed, and from what sources can it be obtained? when and under what conditions will information be gathered? by what means can information be obtained? how will the information be analyzed? how will the evaluation plan be accomplished, and what are the constraints on this plan? and how and to whom will the evaluation results be reported? Five purposes of evaluation are addressed: to make decisions about program installation; to make decisions about program development and modification; to make decisions about continuation, expansion, certification, or termination of a program; to marshall evidence for support of or opposition to a program; and to advance the understanding of basic psychological, educational, social, and other processes. The process for determining what information is needed and the sources for obtaining that data are considered. Types of analytic designs (including true experimental and quasi-experimental designs) are pointed out, and charts of some basic evaluation designs are offered. The types of measures for obtaining data and considerations for selecting appropriate measures are pointed out. Defined are differences in kinds of statistics, level of measurement, and analysis techniques. Some practical matters concerning evaluation roles and responsibilities, resource review, cost estimates, and ethical guidelines are mentioned. Finally, the impact of the final evaluation report is reviewed. Appended are a glossary of key terms for educational measurement and program evaluation and a list of information sources about tests and measurement. (SW)

# Program Evaluation in Early Childhood/Special Education:

## A Self-Help Guide for Practitioners

WESTAR SERIES PAPER #13

Ellis Evans, Ed.D.

Prepared and distributed by:

WESTAR
Western States Technical Assistance Resource
345 N. Monmouth Avenue
Monmouth, Oregon 97361
503/838-1220, ext. 391

WESTAR Coordinator:
Gabrielle du Verglas

Editing by
Ruth Pelz and
Arnold Waldstein

Word Processing by
Valerie Woods

Reproduced and distributed under contract number 300-80-0753 from Special Education Programs, U.S. Department of Education. The opinions expressed herein do not necessarily reflect the position or policy of the U.S. Department of Education, and no official endorsement should be inferred.

Printed in the United States of America

1982

# Program Evaluation in Early Childhood/Special Education: A Self-Help Guide for Practitioners

Like it or not, contemporary educators live in the age of an evaluation imperative. Outcomes of general educational and special intervention programs, in particular, are subject to unprecedented scrutiny by legislators, funding agencies and sundry taxpayers. Controversial flags of accountability and cost-analysis wave briskly, if not menacingly, in the social-political tradewinds. It seems that the entire nation now claims a Missourian heritage: "Show me!"

Yet the meaning and importance of program evaluation in educational settings can hardly be overemphasized. Evaluation can help us to understand what we are doing (or have failed to do), examine the results of these actions and determine how worthwhile are these results. Systematic evaluation should enable more informed and deliberate decision-making about program installation, improvement, continuation, or termination--since it will often answer key, interrelated questions about standards of quality, program components that are linked variously to program outcomes, and the progress of individual participants, including children, parents and staff.

As the emphasis on systematic program evaluation has increased, so has the professional literature on evaluation philosophy and methodology. For persons who have been nursed by this literature, weaned on principles of measurement and statistics, and sustained by a controlled regimen of technical evaluation reports, program evaluation can be a savory experience. Others require heavy doses of conceptual castor oil for healthy program evaluation. And still others may even suffer from chronic evaluation skills malnutrition. Such a plight is understandable. Field workers are under great pressure to produce adequate programs, to say nothing of program evaluation.

Consultants. For those who lack the skills themselves, some salvation is available from a growing horde of evaluation experts, technical consultants and data systems specialists. These experts serve up a variety of evaluation meals, some palatable, others strangely exotic, and still others of uncertain value. At the least, these specialists can provide temporary relief from the evaluation methodology blues. At best, they can be likened to miracle-workers, solving a host of perplexing problems about context-specific program evaluation. Either way, if we are lucky, an expert can be called in at the last minute to help shoulder the blame for program disasters!

Evaluation consultants should be chosen carefully. Ideally, a competent consultant will excel in the following areas: knowledge of innovation in education, research design analysis, educational measurement, data processing, evaluation administration, communications and public relations, and so on (see Millman, 1975). Local program evaluation needs may require highly specific job competencies and should figure strongly in the selection of a consultant. Above all, program directors and staff should reach a consensus about the consultant's role definition and performance. For example, expectations about evaluation as an aid to decision-making or an assessment of merit by an external authority should be clarified. Whatever the role definition, evaluation is most likely to succeed when program staff work cooperatively with evaluation specialists from the outset of the program.

Although it is sensible for practitioners to make selective use of evaluation consultants, local program evaluation is probably best thought of as a "do it yourself" thing. This serves the interests of authority and responsibility for control of educational resource deployment. One important step toward such control is achieving a full perspective on program evaluation needs, the problems to be solved and relevant sources of information about evaluation.

Hence the reason for the present treatise. (The subject is not roses. Rather the subject is self-help for program evaluation.) It is based upon the assumptions that for a given setting, evaluation is feasible, can be conducted in a capable manner, and will influence program decisions. Readers who currently belong to ONOE (Organization for No Organized Evaluation) are requested to drop, or at least suspend, their membership for the duration of this reading.

## THE SEVEN DEADLY QUESTIONS

Evaluation, like sin, may appear deadly to most well intentioned educators. But the following seven questions--far from being sinful--will, in fact, serve as a simple and effective guide through imagined evaluation perils. By answering each of the questions in turn, an evaluation plan can be logically and fairly easily developed which meets the needs of your particular program.

### Question 1. What is the Purpose of My Evaluation?

The key to successful, meaningful and relatively painless evaluation is planning. This should begin with planning the educational program in general and should take into account program needs and justification, goals and objectives, resource availability, service delivery and so on. As with any plan, the first question to be answered must be that of purpose. The purpose of your evaluation will affect all other work in this area.

To begin developing your answer, the following five major purposes should be considered (Anderson & Ball, 1978); a specific plan will generally involve one or more of them. All are independent of specific contexts and personnel. Though interrelated, each purpose is sufficiently distinct to call for somewhat different emphases on evaluation strategy or method.

| Evaluation Purpose | Primary Emphasis |
|---|---|
| 1. To make decisions about program installation. | Activities to determine the need and demand for a program, to assess program concept and integrity, and to appraise the adequacy of resources for inaugurating the program. |
| 2. To make decisions about program development and modification. | Activities to improve a program during the formative stages of its development; to determine the extent to which a program is implemented as intended or is congruent with desired standards of quality; to determine the merit of instructional materials and suitability of characteristics of the physical environment; and to assess program staff competencies. |

| | | |
|---|---|---|
| 3. | To make decisions about continuation, expansion, certification, or termination of a program. | Activities to assess overall program impact and effectiveness in relation to stated objectives, costs, and/or competing programs; or to determine continuing need for programs, and to examine possible side effects, some of which may be unintended and even negative. |
| 4. | To marshall evidence for support of or opposition to a program. | Activities similar to (3) above, but in a context of political realities, mindful of funding priorities and/or vital ethical issues. |
| 5. | To advance the understanding of basic psychological, educational, social and other processes. | Activities to yield basic knowledge, insights for theory, and generalizations that apply within various scientific disciplines. |

Clearly, these purposes are not mutually exclusive. But they do help us to distinguish strategic directions for our evaluation efforts. Further guidelines for determining evaluation objectives and priorities are cogently presented by Rossi and McLaughlin (1979).

*Formative and Summative Evaluation*. Readers of this paper are most likely to be concerned with Purposes (2) and (3). Roughly speaking, these purposes are summarized by the terms *formative* and *summative* evaluation, respectively (see Glossary--Appendix A). Formative evaluation occurs during a program's developmental stages, while changes can be made to finely tune program processes before taking the measure of overall impact. Any meaningful formative evaluation requires that we know as much as possible about a program's structure, intent, content and so on. For a beginning, there is no substitute for a detailed prose description of all program components. Thereafter, careful thought must be given to procedures for monitoring and judging the adequacy with which all components of a program are in place and functioning as intended.

A second type of evaluation is called for at the program's completion. This will involve taking measures to sum up program impact and determine the merit of cumulative accomplishments. It is commonly called summative evaluation. Meaningful summative evaluation requires that we determine program fidelity and seek clear evidence of accomplishments vis-a-vis program goals and objectives. Summative evaluation need not be confined to stated program objectives, however. It can be instructive to examine possible unintended or unanticipated outcomes of program implementation, not all of which are necessarily positive. For example, one or another program of intensive, highly structured academic activities conceivably could produce increased anxiety or negative attitudes in young learners even though their academic competence shows progress.

Most professional evaluators now agree that any attempt at comprehensive program evaluation will involve the complementary forces of formative and summative evaluation. Typically, these are concerned with the power of an educational program to produce desired changes in the behavior of program clientele, especially children and their parents (See Fine, 1980, for information about parent education program evaluation). In many cases, however, they may be applied to individual program components, services, and staff competencies and become ends in themselves (Elliot, 1972; Harms & Clifford, 1980; Millman, 1981; Walberg, 1979). Staff evaluation procedures, for example, are often explicitly carried out for making personnel decisions about training needs, promotion or advancement, merit pay increases and termination. And program implementers often seek to create and maintain certain standards of quality irrespective of any direct linkage to outcomes. The relationship of formative and summative evaluation is summarized by Chart I.

## Question II. What Information Do I Need, and From What Sources Do I Obtain It?

The second step in evaluation planning is to determine what information is needed for program merit assessment and where to get this information. A clear response to the first question, "What are we evaluating and why?" makes the answer to Question II fairly straightforward. If service delivery is the program's focus, then program efficiency and effectiveness are at issue; consequently, both should be evaluated. In other words, both process and product evaluation are usually called for (see Glossary). Cost information may be important as well. In general, program objectives will suggest, if not dictate, an answer to the question about needed information. (This alone is sufficient reason for clearly stated objectives which are supported in principle by program staff.)

Goal Sampling. We have noted that not all evaluation is necessarily confined to stated program objectives. That is, program evaluation may be extended to a scrutiny of unanticipated or unintended outcomes. But in the more typical case, where evaluation is based on program goals, we often face a decision about the feasibility of uniform, across-the-board goal assessment on the one hand, or some method for goal-sampling on the other.

Depending upon evaluation objectives, we may have little choice in this matter, especially if our primary audience is a funding agency demanding full accountability. The stringent condition of full accountability can be met if (1) all program objectives are subject to sound, practical measurement and (2) sufficient resources of time, money and skilled evaluation personnel are available to support comprehensive evaluation.

Otherwise, some system for goal sampling and subsequent merit assessment is required. Two alternatives for this purpose quickly surface. In the unlikely event that all program objectives are equally important, but are too numerous too measure, a random sampling may be necessary. No inviolable rule of thumb exists for this purpose, although something in excess of a twenty percent minimum would seem advisable. A more likely alternative requires a system for determining goal priorities, whereby only the highest priority outcomes are targeted for major assessment. Stake's (1972) procedure for judging the importance of individual objectives is helpful for this purpose as are other systems based upon some version of the Delphi technique (Straus & Zeigler, 1975). This technique seeks to document priorities by consensus (and divergence of opinion) from among program personnel and/or clientele. Occasionally, expert judgment (either internal or external and including the summative evaluator) may be introduced for priority-setting. Of course, programs built upon an explicit theory of development and learning or an a priori value system will already have a basis for defining priorities.

Sources. Information about what data are needed for program evaluation normally will alert us to the second part of Step II in our planning sequence: determining sources of needed information. Restricting ourselves again to typical cases of formative and summative evaluation, at least four important sources of information emerge: clientele (e.g., children and their parents), staff (e.g., teachers, administrators and auxilary personnel), instructional

## CHART I

### Some Distinctions Between Formative and Summative Evaluation

| Feature | Formative Evaluation | Summative Evaluation |
|---|---|---|
| Principal Purpose | Developmental improvement of a program or product | Judgment of the overall merit of a program or product |
| Schedule of use | Continual--data are fed back into developmental cycle | Normally at program completion, when the product is finished, or at a crucial decision point; such as continue or terminate, funding or no funding |
| Evaluative Style | Rigorous, systematic and diagnostic | Rigorous and systematic, with emphasis on comparisons against absolute standards or competing programs |
| Normal evaluators | Internal staff or supportive consultants hired by program or product developers | Preferably, but not always, external and disinterested personnel |
| Consumers of evaluation results | Program designers and staff, product developers, and other "insiders" | Market consumer, funding agencies and other "outsiders" |

Reproduced by permission of Jossey-Bass publishers, from Goodwin, W.L., & Driscoll, L.A. Handbook for measurement and evaluation in early childhood education. San Francisco: Jossey-Bass, 1980.

materials and techniques (e.g., workbooks, manipulables, cueing and reinforcement strategies), and the physical environment (e.g., sanitation, heating, lighting and quality and arrangement of physical equipment in space). The first source--clientele behavior--is mandatory for determining program impact. The remaining three are important sources of information about program processes, i.e., those conditions or events that presumably contribute to changes in clientele behavior.

Program planners and evaluators usually find it helpful, if not essential, to use some sort of classificatory scheme or taxonomy for thinking about program impact on clientele. No standard or universally applicable taxonomy exists for early childhood education, and most require modification for specific adaptations to special education settings. Alternatives abound, however, and are well worth a review by program planners (see Bloom, Engelhardt, Furst, Hill, & Krathwol, 1956; Gagne, 1978; Hoepfner, Stern, & Nummedal, 1971; Kamii & DeVries, 1977; Steinaker & Bell, 1979). Messick and Barrows (1972) suggest ample domains of measurement regarding both program impact and process, from children's cognitive, personal social and health status, through parental and family variables, to aspects of classroom, school and community.

Question III. When and Under What Conditions Will I Gather Needed Information?

An answer to Question III requires what is typically called an evaluation design. A good design will enable us to chart the conditions, timing and method(s) of data collection (Popham, 1975). Again, the choice of evaluation design or investigatory method will depend largely upon evaluation purpose. Existing designs vary considerably in their rigor, scope and applicability to specific field settings.

Types of analytic designs. A number of analytic evaluation designs are available. All seek clear answers to questions about program effectiveness, including causal links between program components and program outcomes. Experimental and quasi-experimental designs based upon some version of "treatment group" and "control group" comparisons are commonly used for these purposes. These include experimental and applied behavioral analysis procedures usually well-known to early childhood special educators.

True experimental design requires random assignment of clientele to one or another program group or condition. Such a design is often difficult, if not impossible, to realize in a field setting. Enter quasi-experimental designs, so-called because they cannot satisfy the criteria of random assignment and strict control over when and to whom a given treatment is applied. Quasi-experimental designs do, however, permit control over decisions about when and from whom to collect data. Workable quasi-experimental designs include time-series designs, pretest-posttest, nonequivalent-group designs, the one group "before and after" design and methods based upon regression analysis (Cook & Campbell, 1979).

Quasi-experimental designs can be vulnerable to certain hazards, called threats to the validity of an experiment or treatment. Alert program evaluators will exercise caution in inferring causal links between program treatment and outcome when such hazards cannot be ruled out. Among the most common threats to the validity of treatment are maturation of the learner, uncontrolled experiences that a learner may have outside the program being evaluated, undetermined biases in selecting learners into a treatment group, and the problem of learner "drop-out", i.e., uncontrolled loss of subjects from a program. In addition, program evaluators can be deceived by apparent changes in learner behavior that may be attributable more to novelty or mere participation in an experiment than to the program itself. These and other hazards, and how best to minimize them through evaluation design, are discussed by Cook and Campbell (1979).

Closely related are models and procedures for determining the extent to which a program (treatment) is implemented as intended (Hall & Loucks, 1977; Leinhardt, 1980). Adequate implementation, of course, should be authenticated for any meaningful test of program impact. For the uninitiated, Eash, Talmadge and Walbes (1974) provide a succinct introduction to design alternatives with practitioners' tastes in mind. More detailed applications of time-series designs are explored by Kratochwill and Levin (1978).

Previous WESTAR publications (e.g., May, 1980) also provide practical guidance in decisions about evaluation design. White (1980) reviews problems of equivalent control groups, regression models and single subject evaluation based upon time-series designs. He argues that the single subject design generally most appropriate for field application is the multiple baseline approach adapted from behavioral analysis research methodology. This approach, in combination with curriculum-referenced checklists (i.e., mastery objectives based upon intended experiences with curriculum content) promises much for balancing our concerns for both individual child and classroom group assessment.

In general, the uses of analytic designs are summative. Several of these designs are also useful for formative evaluation (Fitz-Gibbon & Morris, 1978), but formative evaluation often calls for additional techniques too numerous to mention here (see Sanders & Cunningham, 1979).

In sum, these various analytic methods are particularly well suited for Purposes 3, 4 and 5 as listed under Question I, and sometimes useful for Purpose 2. The approaches share a common concern for the documentation of change over time. Data collection points vary in frequency, but usually involve measurement both prior and subsequent to a treatment. They often include repeated assessments during the course of treatment as well. Some basic evaluation designs consistent with the analytical method are presented in Chart II. Chart III illustrates the assessment strategy feature of basic evaluation designs.

Other Types of Designs. Other, less formal types of evaluation design are also available. Moving from analytic to primarily descriptive approaches, evaluators may resort to correlational methods. Such methods are useful for determining cost-effectiveness relationships, predicting teaching (child) success from selection (placement) criteria, estimating learner response to treatment on the basis of individual personal and home support characteristics, and so on. Surveys of opinions and local resources are useful for program needs assessment and cost estimates, among other things. Expert judgments and clinical or case studies can also be useful for program decision-making. Such methods can figure, for example, into an evaluation design for analyzing suitability of program content and methodology and for investigating unanticipated outcomes, respectively.

Further, sample survey designs can be helpful in gathering information from clients and program staff in situations where explanations of either program failure or program successes are sought for purposes of future planning. For example, an evaluator of several competing pilot programs may wish to ferret out the program elements that distinguish the more successful treatments from the less successful ones. Survey methods focused upon program philosophy (including goals and objectives), physical structure (including classroom organization and materials), human resources (administration and teaching staff), clientele (make-up of the group receiving services), and the community context in which a program occurs can offer important insights to the evaluator (Bryk & Light, 1981).

# CHART II

## Some Basic Evaluation Designs

Below are highlighted some basic evaluation research designs for assessing program impact. True experimental designs are distinguished by the characteristic of random assignment of subjects to treatment and "non-treatment" groups. The other designs are considered quasi-experimental, but vary in their control for threats to the validity of a given treatment. Generally, designs are selected on the basis of information about who will be measured (group or individual) at what points in time. This information, in turn, relates to evaluation purpose. Measurement often includes, but is not restricted to testing. See Cook and Campbell (1979), Eash et al (1974), and Fitz-Gibbon and Morris (1978) for details.

### Evaluation Designs

| | |
|---|---|
| Design A: | True Experimental - Control Group Design, Post-Measurement Only |
| Design B: | True Experimental - Control Group Design, Pre-Post Measurement |
| Design C: | Nonequivalent Control Group Design, Pre-Post Measurement |
| Design D: | Time Series Design, Single Group or Individual, Repeated Measures |
| Design E: | Time Series Design, Non-Equivalent Control Group, Repeated Measures |
| Design F: | No Comparison Group (Single Group) Design, Pre-Post Measurement |

### Design Problems

| Problem: | Design Options: |
|---|---|
| When only one treatment group is available for study, or when questions about program are focused on a single program without concern for comparative evaluation | Design D<br>Design F |
| When two or more groups are available for study and evaluation questions are focused upon the impact of a program vs. no program or alternative programs. | Designs A and B (Highly recommended)<br>Designs C and E (Good alternatives, short of randomization)<br>A post-measurement only strategy is not recommended when non-equivalent control group designs are used. Though unusual, a variation of Design E with true control group can be utilized to good purpose. |

# CHART III

## Basic Evaluation Designs Diagrammed

| | | |
|---|---|---|
| Design A: | True Experimental Control Group Design, Post Measurement Only (rare) | R X O<br>R O |
| Design B: | True Experimental Control Group Design, Pre-Post Measurement | R O X O<br>R O X O |
| Design C: | Nonequivalent Control Group Design, Pre-Post Measurement | O X O<br>-----------<br>O X O |
| Design D: | Time Series Design, One Group or Single Individual, Repeated Measures | Specified Time Periods (e.g., weeks, months)<br>(Design<br>O O O X O O O X O O O<br>Analysis) |
| Design E: | Time Series Design, Nonequivalent Control Group, Repeated Measures | O O O X O O X O O O<br>---------------------------------<br>O O O O O X O O O |
| Design F: | No Comparison Group (Single Group or Individual) Design, Pre-Post Measurement | O X O |

| | | | |
|---|---|---|---|
| Code: | R | = | Random Assignment |
| | -- | = | Nonequivalent groups |
| | O | = | Measurement point |
| | X | = | Program or treatment to be evaluated |

Most of us are well aware of the current bias in program evaluation, namely, a premium on coldly empirical, objective and analytic procedure. ("Only the facts, ma'am. Only the facts!") But more informal methods can offer insights into program operations that sheer pre-post change measures may not reveal. In fact, a growing trend in program evaluation involves a greater reliance upon more differentiated methods of naturalistic investigation (see, for example, Cronbach et al., 1980, and Guba & Lincoln, 1981). This trend reflects heightened sensitivity to the inquiry process itself as a means to fuller understanding of programs. This understanding can extend to issues of longer-term program effects and the transfer or generalizability of program benefits beyond the specific program setting. (All too often, educators have limited themselves or have been easily satisfied with information about immediate, short-term outcomes.)

Question IV. By What Means Can I Obtain Needed Information?

Question IV flows logically from a clear answer to Question III. In fact, some evaluation authorities include data collection techniques under the evaluation design umbrella. In any case, our task here is sharply defined: How best to measure the processes and outcomes that give identity to our program. Two related steps must be taken. First, we need to acquaint ourselves with the range of available and pertinent measurement alternatives. Second, we need to select from among these alternatives those measurements best suited to our stated purpose(s). Acquaintance with measures can be accomplished through a variety of information sources about existing tools or techniques (see Appendix B). For the most part, the sources listed in the appendix represent one or another of two types. Those of the first type describe measures available through commercial sources. The second type concern measures, usually of a more experimental nature, that appear in the professional research journal literature.

If no existing measure(s) from these or other sources suit our purpose(s), our task is clear. We must construct our own. Constructing new measures from scratch is sometimes necessary, even desirable. But it can also be hazardous and time-consuming. Unquestionably, this task requires its own specialized expertise (see Sax, 1980).

Types of Measures. Inspection of the generic program evaluation literature reveals that tests of one kind or another are the most common or popular measurement technique in use. Both the nature and type of tests in use vary considerably, and practitioners unfamiliar with the essentials of psychological tests, including their construction and use, are recommended to consult Green (1981). Meanwhile, the increasingly important distinction between norm and criterion-referenced testing is noteworthy. A primary difference between the two resides in how meaning is derived from individual and group scores. For norm-referenced measures, score interpretation is dependent upon how any individual performance compares to the performance of other persons in a group. (See Glossary, Appendix A, for a definition of norm.) In contrast, scores from a criterion-referenced measure take on meaning in relation to some absolute standard or predetermined criterion. Accordingly, comparisons among individuals are not relevant for score interpretation. For convenience to the readers, this distinction is elaborated in Chart IV, which was derived from the work of Ebel and Popham (1978).

However widespread is the use of tests for measurement in evaluation, we should beware of equating measurement with testing (see Glossary). Many other measurement techniques are available, and they often prove superior to testing, especially when young children's performance is concerned. These include formal systems of direct observation and other observation techniques that are expressed in checklists, rating scales and unobtrusive measures. Questionnaires, interview methods, projective measures, sociometric techniques,

## CHART IV

### Features of Norm-Referenced and Criterion-Referenced Tests

| | NORM-REFERENCED | CRITERION-REFERENCED |
|---|---|---|
| Basic Purposes | Compare individuals; summarize general level of achievement in some area of learning; make program placement decisions when placement is competitive or restricted to certain numbers of individuals; determine for what individuals a given program is most effective. | Establish how well an individual performs in reference to a criterion or standard; develop a program specifically for the individual; determine the extent to which a program is effective for promoting desired criterion performance. |
| Item Type | Items constructed to discriminate among individuals. Items passed or failed by all subjects are omitted. | Items must be congruent with established standards or criterion levels; it is imperative that items yield precise information about an individual's competencies and failings. |
| Content | Content may or may not correspond to specific program goals; content is often "generic" and constitutes a sample from some larger domain of tasks. | Content must correspond exactly to program objectives preformulated in behavioral terms; standards of performance are established for content levels with specifications of minimum competencies. |
| Scores | Scores must reflect variability or dispersion; scores may conceal full performance capabilities of the individual, but they index relative standing in the group. | Scores must indicate exactly what an individual is or is not able to do in relation to specified competencies. |
| Type of Ranking | Percentiles, age and grade norms, standard scores. | Percentage or frequency of meeting criterion-level standards, pass/fail data on successive items. |

personality or attitude inventories, and measures of physiological function round out a picture of familiar measurement techniques. Further detail about such measures can be found in Evans (1974) and Goodwin and Driscoll (1980). Recent publications especially relevant for early childhood special educators are Larsen (1980) and Neisworth (1981).

Selecting Appropriate Measures. Acquaintance with a full range of potentially useful measures for program evaluation is one thing; actual selection is quite another. Thus our second major task for Question IV is to evaluate the merits of various measures. By merits we mean technical qualities such as validity and reliability, as well as the practicality or feasibility of a given measure (see Glossary). Commercially available measures should be accompanied by a technical manual which presents necessary supporting information. But presence of a manual does not guarantee sound credentials. Many manuals fall short of minimum standards for technical adequacy. The situation is even more problematical for experimental measures, many of which have not been sufficiently field-tested for validation.

One particularly helpful scheme for the evaluation of tests and scales is summarized by the acronym, VENTURE (Hoepfner, 1972). Validity (V) concerns how well a test or scale measures a characteristic, skill, or ability it was designed to measure. Examinee appropriateness (E) concerns the suitability of the instrument for a person undergoing measurement, including comprehensibility of test language, fitness of physical format, and appropriateness of the response required of that person. Normed excellence (N) refers to the reliability of a measure, i.e., the extent to which it yields consistently dependable results. Extent of provision for teaching feedback (T) information involves both norm group comparison data and score conversion methods (in the case of tests), as well as ease of interpretation by personnel other than measurement specialists (including test-takers themselves). Usability (U) concerns various administrative aspects of a measure, such as financial cost, time demands, and training requirements for administration. Retest potential (R) concerns the number of alternate forms for an instrument. And finally, ethical propriety (E) encompasses moral and human rights-considerations in administration, content and use of results.

A scheme such as VENTURE can point us toward more systematic and comprehensive evaluation of existing measures. It can also guide us in the construction of new measures. Hoepfner (1972) should be consulted for the full story about the VENTURE system. This reference also shows an application of the system to a variety of existing tests. Factors peculiar to the evaluation and selection of criterion-referenced tests are examined by Kosecoff (1976). Equally systematic schemes for the evaluation and construction of observation systems are also available (See Herbert & Attridge, 1975). Cole and Nitko (1981) provide for solving key practical and technical problems of choosing program outcome measures and measuring program treatment.

## Question V. How Will I Analyze My Information?

In search of answers to questions about information or data analysis, we may find ourselves lured, willingly or otherwise, into the parlor of statistical method. To avoid falling prey to the conceptual spiders therein, we must fortify ourselves with basic knowledge about statistics as applied to evaluation. Two main fields of statistics are involved: descriptive and inferential (Anderson, Ball, & Murphy, 1975). Descriptive statistics provide us with the means for summarizing data in order to interpret them. Two descriptive properties of interest to most evaluators are central tendencies (points in a distribution around which scores tend to cluster or center) and dispersion (ways to represent the spread and variation of scores). Specific statistical indices (e.g., means or averages and standard deviations) are computed from raw data to provide more or less precise estimates of central tendencies and dispersion. Inferential statistics, in turn, provide us with the capability of testing hypotheses about program effects and determining the accuracy of estimates represented by descriptive statistics. Statistical inference allows us to make further estimates about the probability that a given treatment (program), or component thereof, produces an effect beyond that which might occur by chance, extraneous forces, or bias unrelated to the treatment per se. To the extent that these factors can be ruled out of the picture, we can speak more confidently about real or statistically significant differences that are consequent to treatment.

The essential story of statistics cannot be simply or shortly told. There is no substitute for biting the statistical bullet long enough to master the essentials--including an understanding of scaling and analytical techniques based upon inferential statistics. This is because our methods for data analysis are often dictated by the type of raw data collected or level of measurement represented by information from Question IV.

Levels of Measurement. Levels of measurement take the form of different scales, or graduations of the properties of events, objects and persons. These scales vary in their sophistication, the simplest being nominal scales. A nominal scale involves merely classifying or categorizing properties on the basis of a qualitative distinction. Numerical designations are arbitrary and do not indicate amount of a characteristic or variable. Thus individuals in a program evaluation may be classified as 1 (normal) or 2 (exceptional); instructional methods may be classified from 1 to 3 depending upon their distinctive qualities (e.g., verbal expository, guided discovery and inductive); and numbers can be assigned to different classes of well-defined behavior, such as antisocial or prosocial behavior. Nominal scale numbers cannot be treated meaningfully in standard arithmetic operations. But they are very useful for sorting purposes and can be subjected to chi square analysis.

Ordinal scales are a modest advance from nominal scales. They involve ordering the objects or events being measured from least to most, smallest to largest, shortest to longest, and so on. Thus the essence of an ordinal scale is the concept of "greater than." Examples are teacher's rankings of children's popularity or school readiness. Like nominal scales, ordinal scales pose limitations for mathematical operations in the strictest technical sense. But ordinal scales are amenable to correlational analysis, including changes in rank order that may result from an intervention program. And in practice, ordinal scale data are often treated like interval scales.

Interval scales are so-called because their distinguishing characteristic is that adjacent units on the scale are equidistant from one another, irrespective of their position on the scale. We prefer to think that most well-constructed measures of cognitive-intellectual skills, academic achievement and language proficiency qualify as interval scales. Certainly they are treated as such in most data analysis systems. That is, scores derived from interval scales are routinely added, subtracted and fed into calculators and computer programs for complex manipulation.

So also are scores based upon ratio scales, the most advanced form of scaling in use. Ratio scales are distinguished by an absolute zero point, i.e., a zero on this scale signifies the absence of an attribute being measured. Weight, age and engaged time-on-task are examples of variables measured on a ratio scale. All basic arithmetic operations can be legitimately and meaningfully performed upon numbers from ratio scale measurement. Thus a child who spends 30 full minutes on a learning task can be said to spend twice as much time as a peer who spends only 15 minutes on the same task. In contrast, IQ score comparison, at best representating interval measurements, is a different matter. It is incorrect to claim that a child whose IQ is 100 is exactly twice as intelligent as one whose IQ is 50. This is because a "zero point" in intelligence measurement is arbitrary.

Analysis Techniques Commonly used techniques for statistical analysis include chi-square, t-test, analysis of variance, analysis of covariance, and multiple regression (see Chart V). Further information about these techniques can be obtained in any basic statistics textbook for psychologists and educators (see, for example, McCall, 1980; Sax, 1980). Data analysis procedures appropriate for single case experimental designs are reviewed by Hersen and Barlow (1976). Among the most helpful resources for data analysis is a handbook for decision-making prepared by Andrews, Klem, Davidson, O'Malley, and Rodgers (1981). Nearly 150 currently used statistical techniques and applications are concisely reviewed. This handbook is notable for its sequential description outline of decisions that a researcher or evaluator might follow in choosing a particular data analysis procedure. The format for this is a "decision tree," i.e., a branching structure of sequential questions and answers that lead the analyst to a suitable technique. Some common techniques are summarized in Chart V.

Short of carefully structured self-study, field workers are recommended to an informed consultant for statistical assistance. With luck, an enlightened consultant may help us to understand statistics without really understanding statistics.

Question VI. How Will My Evaluation Plan Be Accomplished, and What Are The Constraints On This Plan?

We now confront some very practical matters concerning evaluation roles and responsibilities, resource review, cost estimates, ethical guidelines and the like. It is convenient to think about such matters in categorical terms. (Owens & Evans, 1977; Juarez, 1980). First, personnel role specifications will be necessary. This means a division of labor to accommodate the overall management and coordination of evaluation activities, design decisions, selection of measures, data collection and analysis, and evaluation report writing. Resource allocation also requires our attention, especially in relation to funds available to support the evaluation activity. Preparation of an itemized budget is usually called for, with an eye to any needed consultant assistance, materials and supplies, physical facilities, clinical support, and computer time. Scheduling decisions are paramount as well. It can be helpful to establish a timeline on which are entered key dates for completing the overall evaluation plan, selection and/or construction of evaluation instruments, data collection and analysis, progress report and final report writing, and so on. Periodic monitoring of evaluation activity is particularly important during the formative evaluation phase. This may require regular staff meetings and individual contact with personnel who are in a position to provide valuable corrective feedback to the program manager and chief evaluator. Where they exist, funding agency requirements for evaluation usually provide explicit guidance in these matters. And in public school contexts, conformity to district-wide and individual school policies about evaluation research is imperative.

Ethics. Such policies nowadays should include safeguards for the protection of human rights. In fact, evaluation ethics have become so prominent a feature in education and psychology that we can ill afford to underestimate their significance. The complete evaluation plan should be examined to see that persons in a program or comparison group are treated fairly. Prior to any data collection, it is desirable, often necessary, to obtain appropriate permissions or consent from subjects (or their parents), to guarantee confidentiality of information, and to ensure rights to privacy. And it is usually a good idea to confine data collection to areas that are explicitly germane to a program's evaluation needs (Anderson & Ball, 1978). Evaluation instruments and their underlying rationale should not contain material that is offensive, or implicitly demeaning, to persons being assessed. Data storage procedures should secure confidentiality and mitigate against any possible use of dated or obsolete information. And

# CHART V
## KEY FUNCTIONS OF PREVALENT STATISTICAL ANALYSIS TECHNIQUES

| STATISTICAL TECHNIQUE | FUNCTIONS | EXAMPLES |
|---|---|---|
| Chi Square ($X^2$) | To determine the significance of differences between (a) expected frequencies and observed frequencies; or (b) two sets of observed frequencies. Appropriate use depends upon a sample size of 5 or more cases in the observed frequency category. | -Do males significantly outnumber females in their incidence of, and special class placement for, specific learning abilities?<br>-Is there a significantly larger proportion of on-task behavior among children who are under controlled medication than among those who are not? |
| t-test (t) | To determine whether the differences between two means group averages is statistically significant. Applicable to the difference between two groups or pre- and posttest measures of the same group. Can be performed on small sample sizes (less than 10 cases per group) but power of the test increases as sample size increases. | -Do preschool children who experience a formal sensory discrimination training program score significantly higher on a measure of pre-academic skills than those undergo informal or incidental training?<br>-Do parents who participate in a special child development course show significant change in knowledge and attitudes about handicapped children? |
| Analysis of Variance (ANOVA) | To determine the significance of differences among more than two groups and/or among two or more variables per group. Again, this test can be performed on small samples, but its power decreases as sample size decreases. More commonly used in large group studies (15 or more per group) | -Which of three strategies for teaching sound-letter combinations is most beneficial for visually impaired learners?<br>-Do the duration and intensity of a memory training program make a significant difference in the memory skills of normal and handicapped learners? |
| Multiple Analysis of Variance (MANOVA) | To determine the significance of differences between multiple groups simultaneously on two or more variables. Sample size preference similar to ANOVA. | -Do children in an experimental preschool score significantly higher in language skills (as measured by separate scores in vocabulary, sentence production, articulation, and auditory discrimination) than children in a traditional program or no program? |
| Analysis of Covariance (ANCOVA) | To determine the significance of differences between two groups when such groups are not considered equivalent on relevant variables at pretest time. Sample size preference similar to ANOVA. | -Do mildly or moderately handicapped learners make greater gains in different programs for social skills training when measured intelligence is accounted for? |
| Multiple Regression | A combination of two or more predictive measures to forecast achievement gains or other outcomes. A minimum of 10 subjects per prediction variable is generally required for appropriate use of this technique. | -To what extent do age, sex, and pretest scores on a test of basic experiences predict the performance outcomes of hearing-impaired children enrolled in a special intervention program? |

Note. From Owens, T.R. & Evans, W.D. Program evaluation skills for busy administrators. Portland, OR: Northwest Regional Educational Laboratory, 1977. Reprinted by permission.

evaluation report writing should be done in the best spirit of objectivity, free of biased or unwarranted conclusions and any negative implications for persons or groups who were involved in the program. It is unfortunate that ethical aspects of conducting and reporting evaluations have until recently figured quite weakly into evaluation practice. Today we might argue that such aspects should receive foremost attention. When they cannot be accommodated satisfactorily, entire evaluation plans may be scuttled.

Question VII: How and To Whom Will Evaluation Results Be Reported?

As are most components of an evaluation plan, evaluation reporting is determined largely by the original evaluation purposes (Question I). These purposes should help us to specify the primary audiences for the report, as well as the level of presentation and methods for reporting. Typical audiences include funding and administrative agencies, advisory boards, staff and parents. Dissemination may also reach beyond the immediate audiences to the broader professional community and general public. Timing is an important consideration, especially under conditions of formative evaluation and when continued funding decisions are at issue. Such conditions may require briefer, interim reports. Little is served by dallying with comprehensive reporting while important program decisions are being made.

In any case, reporting should emphasize results that can be used constructively for making decisions and answering priority questions that emanate from interested parties. And dissemination should always be governed by principles of effective communication (Datta, 1981). It is helpful to keep in mind the major lines of communication among evaluators, program directors and staff, program participants, community representatives, funding agencies, and decision-makers (see Anderson & Ball, 1978, Chapter 5).

The communication format should be tailored to fit an intended audience's needs. For most funded program evaluations the task is clear: a full technical report with accompanying executive summary. Program administrators and professional colleagues are likely candidates for a technical professional paper in a form similar to journal articles. Executive summaries normally are fed to management-level and instructional staff and advisory boards or committees. As applicable, the news media and lay public are usually best served by concise, nontechnical reports. Given this diversity of interests, skillful report writing is mandatory--to include prior attention to any social, political, economic and ethical implications of evaluation results.

Impact. From the preceding comments, we can infer that major dissemination problems can inhere in poorly timed and poorly written evaluation reports. Such problems, in time, may result in weak utilization of evaluation results, if utilization even follows. Even when dissemination is adequate, evaluation results are not always utilized in intended ways. Thus looms still another issue: the impact of evaluation results. While poor impact may be related to reporting style, it is also associated with attitudes that may pervade a given organization or audience: general inertia or apathy, resistance to change, fear of the consequences of change, and so on (Weiss, 1972). Flawed evaluation design, including the spurious measurement of program objectives, also contributes to impact poverty and deservedly so. Moreover, evaluators are not the only people to whom decision-makers must listen (Anderson & Ball, 1978). For example, pressure or special interest groups often carry big sticks (whether or not they walk softly). Our point is that meaningful and purposive evaluation must be competently planned, executed and disseminated in full awareness of obstacles to utilization.

The Comprehensive Report. Assuming methodological soundness in program evaluation, a comprehensive written evaluation report is the backbone of dissemination activity.

From this master statement briefer reports can be compiled. The comprehensive report will usually incude the following sections (Owens & Evans, 1977): executive summary (overview of major findings and recommendations), introduction (section on report objectives, intended audiences, and overview of content), program description, focus of evaluation, description of evaluation procedures, presentation and interpretation of findings, conclusion and/or recommendations, and appendices (as needed for technical documentation).

## SUMMING UP

We have explored seven deadly questions for program evaluation. These questions lend themselves to a checklist format for quick reference by vigilant program directors. In short, program evaluation planning can begin with a review such as Chart VI.

From this humble checklist, eager practitioners can opt for more elaborate guidelines to determine the adequacy of evaluation planning and design (e.g., Anderson & Ball, 1978; Sanders & Nafziger, 1976). These more elaborate guidelines also serve as a conduit to detailed, practical, working manuals for program evaluation strategists (e.g., DeRoche, 1981; Fink & Kosecoff, 1980; Morris & Fitz-Gibbon, 1978; Owens & Evans, 1977). If benefit-cost analysis is a criterion for program evaluation, still other sources are timely and pertinent (e.g., Thompson, 1980). Practitioners may also wish to examine some broadly encompassing conceptual blueprints or schemes for program evaluation. Called evaluation models, these conceptual schemes provide a structure for the application of evaluation planning and program monitoring criteria. They also clarify the boundaries of an evaluator's role. Models vary in their relevance to any given evaluation purpose and scope. Goodwin and Driscoll (1980) can be consulted for a succinct introduction to evalution models and their use in early childhood education.

Some final thoughts. Most program planners, implementers, and evaluators committed to the general welfare and personal development of children and families carry a hopeful torch for successful educational intervention. Alas, the successes are often small, and sometimes deeply disappointing to such protagonists. Reasons for this are much debated if less than speculative. But astute evaluation-conscious observers of the early intervention scene offer important clues to more likely success.

Sheehan and Keogh (1981), for example, propose four plausible and manageable strategies in pursuit of improved intervention practice. A first strategy is to sustain the maximum power or, if possible, increase the total power of the intervention program. Program duration and intensity warrant attention here. So does breadth of program, particularly for securing strong home, family and community support. Second, we can augment the power of our program evaluation design and analysis. We begin this by insuring that all program variables are operating as intended. More potent and refined statistical techniques often help us to determine elusive treatment effects as well. Third, we can strive toward more sensitive assessment procedures. Technical standards of quality must be verified. And steps must be taken to insure that personnel are sufficiently skilled in arranging appropriate conditions for assessment. Finally, we can expand the base of our assessment procedures beyond conventional testing to include broader and more diverse information. Multi-method assessment may be called for in combination with more qualitative, ethnographic methods (See Spindler, 1982; Wilson, 1977).

On this note, we end our brief guide to self-help for program evaluation. The guide is intended primarily as a springboard to further study. Accountable personnel who cannot avoid the icy waters of program evaluation will find a potential lifebuoy in the reference list for this guide.

CHART VI

Evaluation Status Checklist

| Decision Points | Resolved | Unresolved | Work Remaining (describe) |
|---|---|---|---|
| 1. EVALUATION PURPOSE(S) | | | |
| Clarity | | | |
| Consensus Achieved | | | |
| 2. INFORMATION NEEDS AND SOURCES | | | |
| Information requirements selected | | | |
| Information sources determined | | | |
| 3. CONDITIONS AND TIMES OF INFORMATION SEEKING | | | |
| Evaluation design selected | | | |
| Data collection points outlined | | | |
| 4. DATA COLLECTION TECHNIQUES | | | |
| Alternative measurements evaluated | | | |
| Measurement instruments with adequate technical qualities selected | | | |
| 5. DATA ANALYSIS | | | |
| Appropriate descriptive and internal statistics determined | | | |
| Means for analyzing data established | | | |
| 6. OVERALL EVALUATION PLANS AND CONSTRAINTS | | | |
| Evaluation roles and responsibilities determined | | | |
| Budgeting and scheduling accomplished | | | |
| Ethical safeguards insured | | | |
| 7. DISSEMINATION AND REPORTING | | | |
| Roles and responsibilities determined | | | |
| Audiences targeted | | | |
| Appropriate report formats created | | | |

# APPENDIX A

## A Brief Glossary of Key Terms for Educational Measurement and Program Evaluation

1. Educational Evaluation and Measurement:
Educational evaluation generally refers to an assessment of the merit of some educational enterprise, i.e., a determination of worth or value. Most systematic evaluation will require one or more forms of measurement which concerns the assessment of status. Measurement involves the assignment of some type of numerical index to a phenomenon, e.g., a description of achievement test performance in numerical terms. Measurement is indicated when, for example, we may describe the average of a group of students on some measure, or rank order a group of individuals who participate in some program on the basis of some behavioral criterion. But if we go further to make some judgment about how good or bad are various measured performances, we engage in evaluation.

2. Formative Evaluation:
Formative evaluation refers to assessments of worth that occur while a program is still open to or capable of being modified or improved. Thus a formative evaluator gathers data and judges the merits of various program features of an educational sequence in order to correct deficiences and suggest needed modifications. This represents, in one sense, a "trouble shooting" or revisionary orientation, literally evaluation in the formative stages of a program.

3. Norms:
A range of values that constitute the usual performance of a given group to which the performance of any given individual may be compared. Norms are typically presented as age, grade, or percentile norms. An age norm concerns the representative performance or developmental status of persons of a given age level for a given measured characteristic. A grade norm is the score, or narrow range of scores, that is typical of the actual performance of the school population for a given grade. A percentile norm is a point on a scale defined by the percentage of scores in the population that falls at or below that point. The concept of standard scores is often important for representing norms.

4. Process Criteria for Evaluation:
Evaluators may often choose to assess the worth or adequacy of conditions, events and materials involved in a program that are thought to affect the quality of program outcomes. They focus in upon what intervenes in the experience of program participants that may contribute to, or somehow be associated with, changes in participant behavior or products. Thus we may "evaluate" a physical environment, legibility of written materials, or the personality characteristics of teachers in light of assumptions about how such factors affect learners.

5. Product Criteria for Evaluation:
Most educational measurements, and subsequent evaluation, focus upon some type of artifact produced by a learner, such as examinations, papers, aesthetic compositions and so on. Thus we can speak of learner-generated products that serve as evidence of program impact, the merit of which can be judged according to standards. Some

evaluators include learner behavior--what a learner can be observed to do/not to do under specified conditions--in this category. Thus we may "evaluate" athletic competence by observing behavior on the playing field, or "popularity" by observing social groupings. Technically speaking, however, product criteria concern some relatively permanent record of behavior performed by a learner.

6. Reliability:
A test or other self-report measure is said to be reliable to the extent that it yields consistent and stable scores for the particular event or attribute being measured. In the case of two or more observers or judges of an event or attribute, reliability is represented by the extent of inter-rater agreement at a given point in time or across time. Reliability takes one or more of several forms, normally indexed by some type of correlation coefficient. A coefficient of internal consistency can be derived by correlating scores on two halves of one administration of a measure. Scores on the same measure separated by a time interval can be transformed into a coefficient of stability. And scores from alternate, but equal forms of the same measure are represented by a coefficient of equivalence. Other forms of reliability exist, but the important point to remember is twofold: First, reliability is a necessary, but insufficient, condition for validity; and second, no measure can be more valid than it is reliable.

7. Standard Scores:
A standard score is a derived score, i.e., a score that is converted from a qualitative or quantitative mark on one scale into the units of another. This is essentially what happens when a score on a final examination is converted into a grade-point-equivalent. In the case of standard scores, however, raw scores are translated into terms that reflect a particular relationship or meaning vis-a-vis mean (average) performance and variation in performance within a group of persons being measured. Common types of standard scores are T Scores, z Scores, Stanine Scores, and the IQ. The advantage of converting is that standard scores on any one type are comparable, even though the raw scores are not. Weighting and averaging of scores can also occur in a more valid and meaningful way. Just about any general textbook in educational or psychological measurement can be consulted for full details on these and other terms associated with testing practices.

8. Summative Evaluation:
Summative evaluation concerns assessing the merit of a completed program, as in the effects of a program in its final form. Data are gathered about program impact at the end of the enterprise to "sum up" the overall worth of the activity, usually to make decisions about continuing, adopting or rejecting a program.

9. Test:
Any set of situations or occasions established for the purpose of eliciting a characteristic way of behaving (responding). The evidence gained is often considered as a sample of behavior which can be used to generalize more broadly about performance in various settings. The occasions for response in testing most often take the form of questions or similar verbal stimuli. Many distinctions among types of tests can be made (see Evans, 1974), and some authorities distinguish tests from scales and inventories designed to measure values, attitudes, personality characteristics, and the like.

10. Validity:
Generally, validity concerns the extent to which a test (or other measure) provides accurate information from which correct inferences can be made for decision-making purposes. Measures are not more or less valid in terms of some inherent quality, rather in relation to measurement purposes. Thus a given test may have high validity for predicting success (or failure) in a special intervention program, but low validity for measuring the specific outcomes of the program. Validity can be thought about in three interrelated categories. To the degree that scores on a measure allow accurate inferences about some underlying human trait or characteristic (e.g., intelligence, creativity, anxiety), we can speak of construct validity. In evaluation studies, the extent to which a measure indicates genuine achievement of program objectives, we speak of content validity. And the degree to which scores on a measure show an accurate relationship to scores on some external standard, we have criterion validity. (This type of validity is critical in selection and placement decisions, as in the use of paper and pencil measures to predict teacher effectiveness as determined by some later measure of actual on-the-job performance.) Whenever possible, validity is indexed statistically, as a correlation coefficient.

# APPENDIX B

## Information Sources About Tests and Measurements

1. Assessment instruments in bilingual education: A descriptive catalogue of 342 oral and written tests. Los Angeles: National Dissemination and Assessment Center, California State University, 1978.

2. Assessment instruments for limited English-speaking students: A needs analysis. Washington, DC: U.S. Department of Health, Education and Welfare (NIE), 1978.

3. Babcock, B. A practical guide to commonly used standardized achievement tests. Austin, TX: Dissemination and Assessment Center for Bilingual Education, 1979.

4. Buros, O.K. (Ed.) The eighth mental measurements yearbook. Highland Park, NJ: The Gryphon Press, 1978 (See also earlier Yearbooks).

5. Coller, A.R. Systems for the observation of classroom behavior in early childhood education. Urbana, IL: ERIC Clearing house on Early Childhood Education, 1972.

6. CSE elementary school test evaluations. Los Angeles: Center for the Study of Evaluation, UCLA Graduate School of Education, 1970.

7. CSE-ERIC preschool-kindergarten test evaluations. Los Angeles: Center for the Study of Evaluation, UCLA Graduate School of Education, 1971.

8. CSE-RBS test evaluations: Tests of higher order cognitive, affective, and interpersonal skills. Los Angeles: Center for the Study of Evaluation, UCLA Graduate School of Education, 1972.

9. Doucette, J., & Freedman, R. Progress tests for the developmentally disabled: An evaluation. Cambridge, MA: Abt Books, 1980.

10. Evaluation instruments for bilingual education: An annotated bibliography. Austin, TX: Dissemination and Assessment Center for Bilingual Education, 1975.

11. Knapp, J. A collection of criterion-referenced tests. TM Report No. 31. Princeton, NJ: ERIC Clearinghouse on Tests, Measurement, and Evaluation, Educational Testing Service, 1974.

12. Johnson, H. Wayne Preschool test descriptions. Springfield, IL: Charles C. Thomas, 1979.

13. Johnson, O.G. Tests and measurements in child development: Handbook II. (Vols. I and II). San Francisco: Jossey-Bass, 1976.

14. Johnson, O.G., & J.W. Bommarito. Tests and measurements in child development: A handbook. San Francisco, CA: Jossey-Bass, 1971.

15. Klein-Walker, D. Socioemotional measures for preschool and kindergarten children. San Francisco, CA: Jossey-Bass, 1973.

16. Mauser, A.J. Assessing the learning disabled: selected instruments. San Raphael, CA: Academic Therapy Publication, 1976. (See also ERIC ED 128 438).

17. Oral language tests for bilingual students: An evaluation of language dominance and proficiency instruments. Portland, OR: Northwest Regional Educational Lab., 1978.

18. Owoe, P., & Johnson, T.J. A critical survey of tests and measurements in early childhood education. St. Louis, MO: Central Midwestern Regional Educational Laboratory, 1974.

19. Pletcher, B.P., Locks, N.A., and others. A guide to assessment instruments for limited-English-speaking students. NY: Santillana Publishing Co., 1978.

20. Simon, A., & Royer, E.G. (Eds.) Mirrors for behavior: An anthology of classroom observation instruments. Philadelphia: Research for Better Schools, 1967-1970 (A 15-volume series).

# REFERENCES

Anderson, S.B., & Ball, S. The profession and practice of program evaluation. San Francisco: Jossey-Bass, 1978.

Anderson, S.B., Ball, S., & Murphy, R.T. Encyclopedia of educational evaluation. San Francisco: Jossey-Bass, 1975.

Andrews, F.M., Klem, L., Davidson, T.N., O'Malley, P.M., & Rodgers, W.L. A guide for selecting statistical techniques for analyzing social service data. Ann Arbor, MI: University of Michigan Institute for Social Research, 1981.

Biggs, J.B., & Collis, K.F. Evaluating the quality of learning the SOLO taxonomy. New York: Academic Press, 1981.

Bloom,. B.S., Engelhardt, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. Taxonomy of educational objectives: Handbook I Cognitive Domain. New York: David McKay, 1956.

Bryk, A.S., & Light, R.J. Designing evaluations for different program environments. In R.A. Berk (Ed.), Educational evaluation methodology: The state of the art. Baltimore, MD: Johns Hopkins University Press, 1981.

Cole, N.S., & Nitko, A.J. Measuring program effects. In R.A. Berk (Ed.), Educational evaluation methodology: The state of the art. Baltimore, MD: Johns Hopkins University Press, 1981.

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

Cronbach, L.J., Ambron, S.R., Dornbasch, S.M., Hess, R.B., Hornick, R.C., Phillips, D.C., Walker, D.F., & Weiner, S.S. Toward reform of program evaluation. San Francisco: Jossey-Bass, 1980.

Datta, L. Communicating evaluation results for policy design making. In R.A. Berk (Ed.), Educational evaluation methodology: The state of the art. Baltimore, MD: Johns Hopkins University Press, 1981.

De Roche, E.F. An administrator's guide for evaluating programs and personnel. Boston: Allyn and Bacon, 1981.

Ebel, R.L., & Popham, W.J. Debate: The case for norm-referenced measurements and the case for criterion-referenced measurements. Educational Researcher, 1978, 7, 3-9.

Eash, M.J., Talmadge, H., & Walberg, H.J. Evaluation designs for practitioners. Princeton, NJ: ERIC Clearinghouse on Tests, Measurements, and Evaluation, TM Report 35, Educational Testing Service, 1974.

Elliot, D.L. Early childhood education: How to select and evaluate materials. New York: Educational Products Information Exchange Institute, No. 42, 1972.

Evans, E.D. Measurement practices in early childhood education. In R.W. Colvin and E.M. Zaffiro (Eds.), Preschool education: A handbook for the training of early childhood educators. New York: Springer, 1974.

Fine, M.J. (Ed.). Handbook on parent education. New York: Academic Press, 1980.

Fink, A., & Kosecoff, J. An evaluation primer. Beverly Hills, CA: Sage, 1980.

Fitz-Gibbon, C.T., & Morris, L.L. How to design a program evaluation. Los Angeles: Sage, 1978.

Gagne, R.M. The conditions of learning (3rd Edition). New York: Holt, Rinehart and Winston, 1978.

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

Green, B.F. A primer of testing. American Psychologist, 1981, 36, 1001-1011.

Guba, E., & Lincoln, Y.S. Effective evaluation. San Francisco: Jossey-Bass, 1981.

Hall, G.E., & Loucks, S.F. A developmental model for determining whether the treatment is actually implemented. American Educational Research Journal, 1977, 14, 263-276.

Harms, T., & Clifford, R.M. Early childhood environment rating scale. New York: Columbia University, Teachers College Press, 1980.

Herbert, J., & Attridge, C. A guide for developers and users of observation systems and manuals. American Educational Research Journal, 1975, 12, 1-20.

Herson, M., & Barlow, D.H. Single case experimental designs: Strategies for studying behavior change. New York: Pergamon Press, 1976.

Hoepfner, R. (Ed.) CSE-RBS test evaluation. Los Angeles, CA: UCLA Graduate School of Education, 1972.

Hoepfner, R., Stern, C., & Nummedal, S.G. CSE-ECRC preschool-kindergarten test evaluation. Los Angeles, CA: UCLA Graduate School of Education, 1971.

Juarez, T.M. Planning guide for the evaluation of educational programs for young children and their families. Chapel Hill, NC: Technical Assistance Development System, 1980.

Kamii, C., & De Vries, R. Piaget for early education. In M.C. Day & R.K. Parker (Eds.), The preschool in action (2nd Edition). Boston: Allyn and Bacon, 1977.

Kosecoff, J. A system for describing and evaluating criterion-referenced tests. Princeton, NJ: ERIC Clearinghouse on Tests, Measurement, and Evaluation, Educational Testing Source, 1976.

Kratochwill, T.R., & Levin, J.R. What time series designs may have to offer educational researchers. Contemporary Educational Psychology, 1978, 3, 273-329.

Larsen, S. (Ed.) Observation and meanings of difference: Measurement of exceptionality. Exceptional Education Quarterly, 1980, 1(3), 1-101.

Leinhardt, G. Modeling and measuring educational treatment in evaluation. Review of Educational Research, 1980, 50, 393-420.

May, M.J. (Ed.) Evaluating handicapped children's early education programs. Seattle, WA: Western States Technical Assistance Resource (WESTAR), 1980.

McCall, R.B. Fundamental statistics for psychology (3rd Edition). New York: Harcourt, Brace, Jovanovich, 1980.

Messick, S., & Barrows, T.S. Strategies for research and evaluation in early childhood education. In I.J. Gordon (Ed.), Early childhood education. Chicago: University of Chicago Press, 1972, 261-290.

Millman, J. Selecting educational researchers and evaluators. Princeton, NJ: ERIC Clearinghouse on Tests, Measurement, and Evaluation, Educational Testing Service, 1975.

Millman, J. (Ed.) Handbook of teacher evaluation. Beverly Hills, CA: Sage, 1981.

Morris, L.L., & Fitz-Gibbon, C.T. Evaluator's handbook. Beverly Hills, CA: Sage, 1978.

Neisworth, J.T. (Ed.). Assessing the handicapped preschooler. Topics in Early Childhood Special Education, 1981, 1(2), 1-75.

Owens, T.R., & Evans, W.D. Program evaluation skills for busy administrators. Portland, OR: Northwest Regional Educational Laboratory, 1977.

Popham, W.J. Educational evaluation. Englewood Cliffs, NJ: Prentice-Hall, 1975.

Rossi, R.J., & McLaughlin, D.H. Establishing evaluation objectives. Evaluation Quarterly, 1979, 3, 331-346.

Sanders, J.R., & Cunningham, D.J. Techniques and procedures for formative evaluation. Portland, OR: Northwest Regional Educational Laboratory, 1979.

Sanders, J.R., & Nafziger, D.H. A basis for determining the adequacy of evaluation design. Bloomington, IN: Phi Delta Kappa, 1976.

Sax, G. Foundations of educational research. Englewood Cliffs, NJ: Prentice Hall, 1979.

Sax, G. Principles of educational and psychological measurement and evaluation (2nd Edition). Belmont, CA: Wadsworth, 1980.

Sheehan, R., & Keogh, B.K. Strategies for documenting progress of handicapped children in early education programs. Educational Evaluation and Policy Analysis, 1981, 3(6), 59-67.

Spindler, G. (Ed.). Doing the ethnography of schooling: Educational anthropology in action. New York: Holt, Rinehart and Winston, 1982.

Stake, R.E. Priorities planning. Los Angeles, CA: Instructional Objectives Exchange, 1972.

Steinaker, N.W., & Bell, M.R. The experimental taxonomy: A new approach to teaching and learning. New York: Academic Press, 1979.

Strauss, H.J., & Zeigler, L.H. The Delphi technique and its uses in social science research. Journal of Creative Behavior. 1975, 9, 253-259.

Thompson, M.S. Benefit-cost analysis for program evaluation. Beverly Hills, CA: Sage, 1980.

Walberg, H.J. (Ed.) Educational environments and effects: Evaluation policy and productivity. Berkeley, CA: McCutchan, 1979.

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

White, O.R. Practical program evaluation: Many problems and a few solutions. In M.J. May (Ed.) Evaluating handicapped children's early education programs. Seattle, WA: WESTAR, 1980.

Wilson, S. The use of ethnographic techniques in educational research. Review of Educational Research, 1977, 47, 245-265.