DOCUMENT RESUME

ED 384 779 CE 069 465

AUTHOR	Polzella, Donald J.; Fine, Mark A.
TITLE	A Manual for Conducting Even Start Program Evaluations.
SPONS AGENCY	Ohio State Dept. of Education, Columbus. Div. of Vocational and Adult Education.
PUB DATE	9 Aug 94
NOTE	54p.
PUB TYPE	Guides - Non-Classroom Use (055)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Adult Basic Education; *Adult Literacy; Citations (References); Early Childhood Education; *Educational Assessment; Evaluation Methods; Family Programs; Federal Programs; Integrated Services; *Literacy Education; Program Evaluation; Program Implementation; Program Improvement; Resources; Student Evaluation; *Tests
IDENTIFIERS	*Even Start; Family Literacy

ABSTRACT

Project Even Start is an intergenerational literacy project that promotes adult literacy, trains parents to support the educational growth of their children, and prepares children for school. The program is designed to facilitate joint participation by parents and children. It includes home-based instruction and child care, and it is integrated within a network of relevant support services such as Head Start, volunteer literacy programs, and legislation such as the Adult Education Act, the Education of the Handicapped Act, and the Job Training Partnership Act. This manual was developed to train Even Start administrators and staff in assessing the effectiveness of their programs. The manual addresses the major components of the evaluation process: (1) definition of literacy and program evaluation; (2) general evaluation methods, concepts, and guidelines; (3) important procedures to follow when conducting a program evaluation; (4) tests and measurement instruments that can be used; (5) general guidelines for recording, storing, and analyzing program effectiveness data; (6) guidelines for documenting the results of the program evaluation, including constructing graphs and tables and writing program evaluation reports; (7) a description of how evaluation results can inform program planning; and (8) an annotated list of resources relating to family literacy program evaluation including 29 books and articles, 9 publishers of tests and measurement instruments, and 11 literacy organizations. (KC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 384 779

A Manual for Conducting Even Start Program Evaluations

Donald J. Polzella, Ph.D.
Mark A. Fine, Ph.D.

August 9, 1994

Even Start project funded by the Ohio Department of Education,
Division of Vocational and Adult Education

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

□ Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

CE 069 465

BEST COPY AVAILABLE

Table of Contents

Table of Contents (continued)

Table of Contents (continued)

A Manual for Conducting Even Start Program Evaluations

Rationale

Purpose of this Manual

The primary purpose of this manual is to train Even Start administrators and staff how to assess the effectiveness of their program and help ensure that it successfully integrates early childhood and adult education into a unified literacy training program. The manual addresses all major components of the evaluation process including:

(1) definitions of literacy and program evaluation;

(2) general evaluation methods, concepts, and guidelines;

(3) important procedures to follow when conducting a program evaluation;

(4) tests and measurement instruments that can be used;

(5) general guidelines for recording, storing, and analyzing program effectiveness data;

(6) guidelines for documenting the results of the program evaluation, including the construction of graphs and tables, and the writing of a program evaluation report;

(7) a description of how evaluation results can inform program planning; and

(8) an annotated list of some available resources relating to family literacy program evaluation including books and articles, publishers of tests and measurement instruments, and literacy organizations.

Definitions, Perspectives, and Problems

In simplistic terms literacy may be said to refer to the ability to read. Yet in today's complex technological world, it means much more. A useful "functional" definition has been provided by Princeton University psychologist George Miller (1988):

> Today any purposeful use of written language (reading, writing, or comprehending print, including an appreciation of tables, maps, diagrams, or mathematical symbols and formulas) may be taken as a criterion defining literacy. (p. 1293)

A similar definition, and one particularly relevant to Even Start programs, was provided by The National Literacy Act of 1991. In this Act, literacy is defined as "an individual's ability to read, write, and speak in English, and compute and solve problems at levels of proficiency necessary to function on the job and in society, to achieve one's goals, and develop one's knowledge and potential."

Achieving this level of literacy is obviously a challenge, and so it is not surprising that semiliteracy is a way of life for millions of Americans. Research has shown that many semiliterate individuals cannot locate information in a newspaper article, follow directions to travel from one location to another using a map, use a bus or train schedule, determine correct change in a restaurant, calculate and total the costs of items in a catalogue, or answer a help-wanted advertisement.

The Even Start family literacy program was created to address the problems of semiliteracy and illiteracy in our country. In passing the National Literacy Act of 1991, the United States Congress has recognized that a comprehensive approach for improving the literacy and basic skill levels of adults is needed. The legislation calls for coordination, integration, and investment in adult and family literacy programs, such as Even Start, at the federal, state, and local levels. The legislation facilitates research and program delivery by helping to support the wide range of organizations-- public, community-based, volunteer, business, and industry--that are involved in providing literacy services.

Even Start is one of the most comprehensive and ambitious of the national programs. Its principal elements are innovative instructional strategies that:

(1) promote adult literacy,

(2) train parents to support the educational growth of their children, and

(3) adequately prepare children for regular school programs.

Because the program is designed to facilitate joint participation by parents and children, there are provisions for home-based instruction and child care. In addition, the program is integrated within a network of relevant support services (e.g., Head Start, volunteer literacy programs) and legislation (e.g., Adult Education Act, Education of the Handicapped Act, Job Training Partnership Act).

What is "Program Evaluation" and Why is it Important?

Program evaluation is a process that measures the effectiveness of a program like Even Start. It involves four general components:

(1) articulating desired outcomes or objectives for the program,

(2) collecting data and other information relevant to those objectives,

(3) analyzing the information to determine the degree to which the outcomes were achieved, and

(4) using evaluation results to make modifications in the program.

The program evaluation process differs on several dimensions, including whether the information is subjectively or objectively gathered, whether the evaluation methodology examines process or outcome goals, and whether results are used only at the local level or are aggregated across local programs.

Type of data. Information pertaining to program effectiveness can either be subjectively or objectively gathered. Subjectively gathered data are derived from the impressions of program staff, individuals who know program participants, or the program participants themselves regarding how much the participants in the program have improved. Objectively gathered data are derived from standardized assessment instruments, such as tests. As we discuss further later in this manual, we believe that a comprehensive program evaluation needs to include both types of data. However, of the two types of evaluation data, administrators of public funding agencies, whose support for programs like Even Start is critical, and legislators, who appropriate funds for these programs, are particularly likely to value objective evidence for program effectiveness. We do not intend to claim that objective data are superior to subjective data; however, practical concerns dictate that at least some objectively derived evidence be gathered to satisfy important external constituencies, particularly for demonstration projects, such as Even Start.

Type of goals. Another important distinction in program evaluation methodologies needs to be made--between process and outcome goals. Process, sometimes referred to as formative, evaluation attempts to evaluate the extent to which program staff and participants are engaging in the activities that will ultimately lead to improvement in literacy skills. For example, with respect to the program, an evaluation of process may include an assessment of the extent to which program staff successfully recruit new participants, whether staff members engage in the interventions that they are supposed to, and whether there are sufficient numbers of staff to meet the participant demand. With respect to participants, process evaluation may consider such elements as how often participants attend their scheduled classes or home intervention sessions, staff impressions of participants' motivation to engage in literacy training activities, and how often participants seek additional help from program staff.

Outcome, sometimes referred to as summative, evaluation is more traditionally recognized by program staff and evaluators than is process evaluation. Outcome evaluation examines the extent to which program participants improve in the areas that are directly targeted by the program. In the case of Even Start, outcome evaluation serves to assess the extent to which participants have improved on any dimension judged to be directly related to Even Start goals (i.e., development of literacy skills, the training of parents to support the educational growth of children, and the preparation of children for regular school programs). Although outcomes are typically assessed at the completion of a program, they can, and should be, measured at selected intervals throughout the course of the program.

Notice that outcome evaluation can use either subjectively or objectively derived data. An outcome evaluation is not necessarily restricted to results from standardized tests; impressions of program staff and the participants themselves can provide useful information relevant to outcome evaluation.

Use of results.. A final distinction that needs to be made is that between evaluation results that pertain to a specific program and those that pertain to aggregated results across programs at the national level. Evaluation results at the local level will help administrators target particular strengths, weaknesses, and possible modifications of their particular program packages. If used appropriately, local evaluation results lead to changes in the specific program, but not necessarily at the state or national level.

By contrast, aggregated results across local programs can provide important information on the extent to which the general program package is effective, on which variants of the general package are particularly effective (e.g., center-based, home-based, vs. both), on which subgroups of individuals appear to be assisted most effectively, and perhaps on which program components are most effective for program participants. Aggregated results can have substantial implications, including the discontinuation of the program, a revamping of its general treatment package, or a modification tailored to particular groups of high risk individuals.

How to Locate and Work with an Evaluator

Evaluating the effectiveness of programs as comprehensive as Even Start generally requires special expertise in evaluation methodologies and a considerable investment of time and resources. For these reasons, it is advisable to access the consultative services of a professional evaluator. To locate an evaluator, we recommend contacting either or both of two sources likely to employ individuals with expertise in evaluation: local colleges/universities or private research firms. At a college or university, we recommend contacting the chairpersons of the departments of psychology, sociology, or education. The chairpersons should be able to identify faculty with evaluation expertise. Private research firms are also potentially fruitful sources. Many of these firms specialize in program evaluation and, thus, their staff have considerable experience in conducting evaluations and presenting the results in the format required by funding sources.

Once a program evaluator is selected, it is important that program staff maintain a close relationship with the evaluator. Program staff may mistakenly believe that the evaluator can evaluate the program without input from program staff and, thus, may be tempted to allow the evaluator to function independently. We cannot emphasize strongly enough the importance of maintaining regular contact with the evaluator. Regular contact is important for several reasons: (a) it allows the evaluator to gain and maintain an overall sense of the program, (b) it allows the evaluator to acquire the perspective and impressions of program staff in addition to results from standardized test scores, (c) it reinforces to program staff the importance of evaluation, (d) it provides opportunities to prevent minor problems (e.g., data are being collected inappropriately, an instrument is not being administered at the planned time) from becoming major problems, and (e) it provides program staff with regular feedback on the results of the evaluation activities.

It is also helpful if program staff, or at least a representative person or two, are familiar with some of the evaluation terminology and strategies. Of course, it is unreasonable to expect program staff to have the same level of expertise as does the evaluator; however, if the program representative has some degree of expertise, necessary communication is facilitated. One of the purposes of this manual is to familiarize Even Start program staff with the basic terminology and concepts of program evaluation.

Evaluation Methodology

In this section, we examine evaluation methodology, which refers to how we determine the extent to which programs are meeting their goals. (See Appendix B for a list of sources.) There are four primary issues to address in conducting a sound program evaluation: (1) the desired program outcomes, (2) the evaluation design, (3) measurement, and (4) ethical and practical considerations. Each is considered below.

The Relation Between Program Outcomes and Evaluation

The first step in conducting a sound program evaluation is to articulate clearly a set of broad program goals. These goals are then translated into objectives or outcomes. It is particularly important that the objectives be stated in such a way that they can be evaluated. For example, consider the following goal: "program participants should increase their literacy skills." While a reasonable goal, it is not specific enough to dictate how progress toward this goal will be assessed. For example, an increase in literacy skills could be assessed by an examination of changes in objective test scores, by the impressions of program staff, or by accounts from the participants themselves.

When we specify the manner in which literacy skills will be assessed, we are translating program goals into objectives. Desired program objectives need to be operationalized (i.e., defined in concrete terms). In the example above, appropriate objectives might be "program participants will show an increase in their test scores after being in the program for three months," "parents will read more books to their children each month after three months of the program than they did before they began in the program," and "participants will report that they are more confident about their abilities to read and write than they were before the program." These objectives are specific enough that they can fairly easily be evaluated.

It is also important that program outcomes be reasonable in number. Program evaluators should realize that for every stated objective there will need to be at least one means of assessment. A large number of objectives may place an unreasonable burden on the program staff, who must administer the assessments. Moreover, the greater the number of assessments, the less time will be available for helping the participants.

One strategy for lessening the tension between providing services to participants and evaluating their progress is to integrate the assessment process into the educational activities. Although this approach has promise, it also has a potential cost; evaluation information may not be gathered in the same manner from each participant. In the absence of this standardized gathering of information, it is hard to aggregate results across participants to make some inferences about program effectiveness.

Just how many objectives is "reasonable" is not easy to determine. The appropriate number will depend on several factors, including the length of the program, the number of staff, mandates from federal sources, and the degree of effort and time required of the participants. Each of these factors must be carefully weighed when determining program objectives.

Finally, it is important that program objectives be realistic in scope. For example, "program participants will score at the 80th percentile or better on a standardized test of literacy skills" is certainly a measurable objective, but it is clearly unrealistic. This level of performance is more attainable for college-educated students than it is for Even Start participants, who may not have graduated from high school. A realistic objective might be "program participants will a show a statistically significant increase in literacy skills test scores."

Evaluation Designs

Evaluation designs refer to how the evaluators choose to test whether the program has had a beneficial impact on participants. To develop an evaluation design, the evaluators must choose: (a) when to gather the relevant information (e.g., test scores, impressions of program staff), (b) how often to gather this information, and (c) whether to have a comparison group of individuals for whom the same information is gathered but who do not participate in the program.

In the discussion that follows, we often refer to information that can assess program effectiveness. Our definition of information is broad and includes both subjective impressions and objective standardized measures. As noted above, these two sources of information can provide complimentary evidence of the extent to which the program has been successful.

Therefore, when the term information is used, the reader should consider both of these complimentary types of assessment methods.

From an evaluation standpoint, the optimal evaluation design is one that allows evaluators to determine the extent to which the program has had a beneficial impact on participants. There are two key components of this optimal evaluation design: (a) it allows one to determine the extent to which program participants have improved, and (b) it allows one to conclude that, among all the possible reasons for why improvement may occur, the program itself was at least partially responsible for the improvement. Evaluation designs differ in the extent to which both components are present.

There are many designs available to evaluators. Several of the more popular evaluation designs are presented below, with each subsequent design an improvement over the previous one. To prepare the reader for the presentation of these designs, we highlight two features of our review. First, of the evaluation designs that are presented below, the pretest-posttest design is the one that is most likely to be used to evaluate tne effectiveness of Even Start programs. However, the other designs are also briefly reviewed to place the strengths and weaknesses of the pretest-posttest design in its appropriate context. Second, the information used to evaluate program effectiveness in these designs can be either subjective or objective.

Posttest only. The weakest evaluation design is one that only assesses participants after they have completed the program. This design is weak because one cannot determine precisely how much participants have improved. Even if one assumes that the participants had literacy deficiencies before the program and the assessment results show that they are now in the "normal" range, it is not possible to determine exactly how much participants benefitted and whether it was the program participation that led to the participants' progress. This design only allows one to determine the level of literacy (compared to published norms on the assessment instrument, if they are available) exhibited by participants after they have participated in the program.

Pretest-posttest. A research design that is stronger than the previous gathers information from participants before and after they participate in the program. This design allows one to determine precisely the extent to which participants have improved. For example, an evaluator may be able to determine that participants gained an average of 80 points on a measure of literacy.

The limitation of this design is that one cannot determine why participants have improved. They may have improved because they participated in the program, but there are also other plausible possibilities that this design cannot rule out. For example, participants may have engaged in some outside activities that led to improved performance (e.g., were involved in an adult education class that was independent of the literacy program), and/or they may have performed better because they were familiar with the assessment instrument the second time they completed it (the "practice effect"). Thus, an improved research design is needed to allow one to specify the reason for the improved performance.

Pretest-posttest with comparison group. In this design, two groups of individuals are evaluated--those who participate in the program ("program" group; also called a "treatment" group in the evaluation literature) and a similar group of individuals who do not participate in the program ("no program" group; also called a "comparison" or "control" group). These two groups are assessed before and after those in the "program" group are served.

If those in the "program" group improve to a greater degree than those in the "no program" group, the evaluators can conclude that the program was at least partially responsible for the improved performance among those in the "program" group. If these results are obtained, one can rule out the possibilities that the mere passage of time or a practice effect (i.e., the participants' performance improves merely because they have already been evaluated and are now familiar with the evaluation procedures) caused the changes.

On the other hand, if individuals in both the "program" and "no program" groups improve a similar amount, then one can conclude that the program was not responsible for the improvement in assessment scores. Rather, some other factor(s)--such as the practice effect--have caused both groups of individuals to perform better.

This design is quite strong in that it allows one to determine the extent to which the program itself has caused the resulting improvements in assessment scores. However, this design does not allow one to rule out that those in the "program" group (more so than those in the "no program" group) engaged in some activities that caused the improved performance. Although there are evaluation designs that are more effective than the pretest-posttest with comparison group design, these are not generally available to Even Start evaluators because of ethical and practical concerns.

The typical Even Start evaluation design. As noted earlier, it is likely that most Even Start programs will use the pretest-posttest design to evaluate the effectiveness of their programs. Although it is generally feasible to obtain information on effectiveness before and after program participation (which means that evaluators can do better than the posttest only design), it is impractical and possibly unethical to form a comparison group of needy individuals and families who do not receive Even Start services. One way that some program evaluators generate such a comparison group is to place some individuals and families on a waiting list, have them complete the assessments at the same times as those who are receiving services, and offer them services at a later point in time. However, this may not be feasible in a program like Even Start, in which program and evaluation funds are scarce, because it is not practical to provide services at two different schedules for families in the treatment and comparison groups.

Measurement

After the design is chosen by considering the factors described above and the evaluator and Even Start staff have considered the ethical and practical issues discussed below, appropriate measures need to be selected. In evaluation research, measure refers to an instrument that is administered to program participants in a standardized manner and that yields a quantitative score that is indicative of a person's literacy proficiency and appropriate means that the measure has good psychometric properties. Psychometric properties include standardization of administration, norms, reliability, and validity. Although the discussion below focuses on objective measures, it can also apply to subjectively gathered information if it is gathered in a standardized manner.

Standardized administration. It is critical that each and every person be administered the testing instrument in exactly the same manner. This allows an evaluator or a practitioner to determine how an individual's scores change over time or to compare persons' scores with others, Thus, a sound measure will have detailed, clear, and specific administration instructions that should be followed each time the instrument is used.

Norms. To interpret an individual's test scores, or a group's collective test performance, it is important that there be a set of norms available for the test instrument. These norms consist of summary statistics--means, standard deviations, and standard errors of measurement--on the test from a large and preferably representative group of individuals. This group is referred to as the "standardization sample."

Some of the well-researched test instruments include subgroup norms based on such variables as age and gender. For example, individuals for whom English is a second language may have different intervention needs than those for whom English is their first language. Thus, it would be desirable to have separate norms for each of these groups.

The presence of norms allows one to compare any given individual or group's test performance with the standardization sample. Without the existence of normative data, one is not able to place the meaning of any given score within an appropriate context. For example, one would not be able to claim that a given score is "high" or that another score is "low." High quality test instruments will have large and representative standardization samples and clearly reported normative data. They would also ideally include relevant subgroup norms.

Reliability. The reliability of a test instrument refers to the extent to which it consistently measures something. A sound test should generate consistent, and hence reliable, measurements of the phenomenon of interest.

There are several ways that reliability is measured. The most common approaches are test-retest, alternate form, split-half, internal consistency, and interrater. The specific statistic used to evaluate reliability is often the correlation coefficient, which ranges from -1 to 1. A score of 0 indicates no consistency and a score of 1 indicates the highest level of consistency.

Test-retest reliability refers to the degree to which the test yields similar results when individuals take the test at two different time periods. A reliable test should yield similar results at the two testing periods. Satisfactory reliability coefficients are typically .80 or higher.

It should be noted that test-retest reliability is only relevant when one is assessing constructs that are thought to be somewhat stable over time, such as literacy; otherwise, test scores would be expected to change over time. In addition, the longer the time between testing periods, the lower the reliability estimate is likely to be.

Alternate-form reliability refers to the extent to which two alternate, or parallel, forms of the same test yield similar scores. To assess this type of consistency, two alternate forms of the test must be available. Alternate forms of the test should have identical means, standard deviations, and relations to other measures.

If alternate forms are available for a given test, a large group of individuals can take both forms and their scores can be correlated. If individuals receive similar scores on the alternate forms, adequate reliability has been demonstrated. Again, satisfactory reliability coefficients are typically .80 or above.

Split-half reliability refers to the degree to which two halves of the same test yield similar test results. Rather than creating a separate, alternate version of the same test, another option is to attempt to create two alternate versions within a single test. Thus, one divides the test into what are believed to be two separate, but equivalent, halves and scores on these two halves are intercorrelated. If the scores are highly intercorrelated, there is evidence to support the reliability of the test. On ability and achievement tests, one common procedure to assess split-half reliability is to sum the scores on the odd items and correlate this summed score with the sum of the scores on the even items ("odd-even").

Internal consistency reliability refers to the extent to which the test items are measuring a similar phenomenon. If the items are measuring one and only one phenomenon, then the internal consistency reliability coefficient will be quite high (.70 or higher). On the other hand, if the measure of internal consistency is low, one cannot conclude that the items are measuring a similar construct. Rather, one would have to conclude that the items are measuring at least two and maybe more different entities. Because measures with low internal consistency estimates are not measuring one phenomenon, they are not considered to be consistent and, hence, are not reliable. The most commonly used measure of internal consistency is Cronbach's alpha.

Interrater reliability refers to the degree to which different raters agree in their ratings of particular people's behavior. For example, suppose one is measuring how often children hug other children in their school classroom. If the measure of hugging frequency is reliable, then different raters should arrive at similar scores for each child.

This reliability coefficient is computed by correlating the scores of one rater with those of another. If their scores are highly intercorrelated (.80 or above), one can conclude that the raters are consistently measuring the number of hugs, or whatever they happen to be assessing. However, if their scores are not highly intercorrelated, then one must conclude that they are not consistently measuring the phenomenon. If this occurs, the ratings are unreliable and either the instrument's rating system needs to be modified or the raters need additional training in how to use the system.

Validity. Validity refers to the extent to which a test instrument measures the construct that it is supposed to. For example, a literacy test is purported to measure literacy skills. Therefore, if the test is valid, there must be evidence that the test does indeed measure literacy skills and not some other construct(s).

It is important to note that a test can be reliable and yet not valid. That is, the test can consistently measure a given entity, but that entity is not what the test was designed to measure. For example, suppose one considered shoe size to be a measure of intelligence. Although shoe size can be reliably measured (i.e., one would obtain the same shoe size each time it was measured), it is not a valid measure of intelligence.

On the other hand, a test cannot be valid unless it is also sufficiently reliable--a test cannot assess what it is supposed to unless it can consistently measure a given construct. When a test does not have adequate validity, one possible reason may be that the test is insufficiently reliable. If one could improve the reliability of such a test, it is possible that the validity would also improve to a satisfactory level.

There are several ways that validity can be tested. Those most relevant to literacy programs include content, criterion-related, and construct validity. As was the case with reliability, the specific statistic used to evaluate validity is often the correlation coefficient, which ranges from -1 to +1. A score of 0 indicates extremely poor validity and a score of 1 indicates the highest level of validity. When used as a measure of validity, this statistic is called the epitemic correlation.

ERIC Full Text Provided by ERIC

Content validity refers to the extent to which experts believe that the test measures the construct that it purports to measure. In the area of literacy, an instrument would have content validity if a group of experts agreed that the items and tasks on the test do indeed measure different aspects of literacy. If these experts believe that the test items do not measure literacy, or if they cannot agree on what the test is measuring, then the measure has poor content validity. Because content validity is assessed by the subjective judgements of experts, there is seldom a quantitative estimate provided.

Criterion-related validity refers to the extent to which scores on the test correlate with scores on other tests (i.e., criteria) which are also designed to measure the same or a related construct. In the area of literacy, a measure of literacy would have good criterion-related validity if its scores correlate highly with another, already established measure of literacy. Good validity coefficients are typically somewhat lower than good reliability ones. Depending on the measure, the area studied, how similar the two constructs are thought to be, and other factors, correlations of .3 or above can be considered adequate estimates of criterion-related validity.

Construct validity is the most important form of validity and actually subsumes content and criterion-related validity. An instrument has construct validity when a body of research literature develops over the years which suggests that the measure is assessing what it purports to. The construct validity of a literacy measure is established over a multi-year period with the accumulation of numerous studies that have demonstrated that scores on the measure relate in the expected direction to a number of other measures of similar constructs. In this sense, each and every study that demonstrates criterion-related validity contributes to the establishment of construct validity for the measure.

For example, scores on a valid literacy measure should be correlated with scores on a measure of a related construct, such as arithmetic ability. However, because literacy is not thought to relate to friendliness, one should not expect to find a high correlation between the measure of literacy and a test that measures friendliness. Thus, if one finds a high correlation between scores on the literacy measure and the arithmetic ability scale and a very low correlation between the literacy measure and the test of friendliness, the construct validity of the literacy measure is supported.

Ethical and Practical Considerations

Although there are evaluation designs that are methodologically superior to the pretest-posttest design, ethical and practical considerations often do not allow one to choose these options, especially when the research involves human subjects. Some of the most important ethical and practical dilemmas faced by evaluators are discussed below.

Ethical dilemmas in evaluation. The primary ethical dilemma associated with designs that have a comparison group is that it may be unethical to deprive those in the comparison group of needed services. While those in the "program" group are receiving potentially beneficial literacy services, those in the "no program" group are not. A potential solution to this problem is put the individuals who are on a waiting list--who cannot presently receive services because of excessive demand--into the comparison group.

There are other ethical concerns in evaluating family literacy programs. These apply to all aspects of the evaluation process and are not specifically related to methodology. Therefore, they are considered further below.

Practical concerns. Limitations on resources often influence the choice of evaluation methodology. Strong evaluation designs typically are expensive and time-consuming for both staff members and clients. When these resources are scarce, compromises must be made and, consequently, evaluation designs may be chosen that are less than optimal. However, because Even Start is a demonstration project, it is reasonably well-funded and local programs should be sure to include sufficient funds in their budgets to conduct sound program evaluations.

Another practical concern is that evaluators of local programs often have some of their evaluation activities mandated by state or federal agencies. With Even Start, although local programs are expected to conduct their own evaluations, they must also follow the federally mandated evaluation procedures. Even when these mandated strategies are likely to be quite useful, as is the case with Even Start, they nevertheless place some limitations on what local programs can reasonably expect to accomplish with respect to their own unique approach to program evaluation. For example, if administrators of a local program choose to gather information (either objectively or subjectively gathered) in addition to that required by federal mandates, additional evaluation resources are required and even more of program participants' precious time is consumed. Of course, the potential utility of this information must be weighed against these costs.

Administration Issues

In this section, issues related to how program staff can effectively, ethically, and appropriately gather evaluation information will be addressed. The areas that are covered below include preparing the staff members who will actually be gathering the information, the process of gathering the information, and ethical issues.

Preparation of Staff Members

Staff members in literacy programs may have little or no experience in program evaluation. Some may not understand the importance of conducting sound evaluations to determine program effectiveness. Therefore, it is important that staff members who gather evaluation information (e.g., administer test instruments) be properly trained and prepared.

Perhaps the most important point to convey to test administrators is that the data gathering procedure, if done properly, will actually lead to useful information. Many individuals are skeptical of evaluation--believing that the results are not used to make programmatic decisions and/or that evaluation is something that is done only to satisfy bureaucratic mandates. In many cases this skepticism is well-founded. However, the usual alternative is that program staff will identify a select few of the successful program participants they are most familiar with as "proof" that a program is successful. This not only overlooks those who may not have benefitted from the program, but may overemphasize the success of a program. We argue that evaluation results should accurately represent the success of a program, that they should be used to inform program decisions, and that modifications in programs should reflect the results of well-designed evaluation studies. If this feedback is present, it will be easier to convey to those who gather evaluation information that this process can produce beneficial outcomes.

Process of Gathering Evaluation Information

Standardized tests. During the test administration itself, there are several factors that need to be considered. A necessary element of reliable and valid testing is the development of rapport with test-takers. To maximize the chances of achieving useful testing results, the tester must attempt to ease the anxiety of the test-taker. To help the test-taker feel comfortable, the tester should show respect, empathy, and caring for the individual.

A second consideration is that test-takers should be committed to follow as closely as possible the standardized administration procedures. Even slight deviations from these instructions may substantially influence the results. Thus, all efforts should be made to follow these procedures. If this is not possibie, then this should be clearly documented and an individual's resul ig score can be used for programmatic purposes but should not be compared with published norms.

Another critical element of the test administration process is the physical environment. Ideally, testing should occur in quiet, private, and comfortable surroundings that are relatively free of distraction. If the physical environment is lacking in any of these areas, the reliability and validity of the test results may be compromised.

The surest way to insure that the physical environment is conducive to testing is to assess individuals in a center-based room. However, this may not be possible. If testing in the home must be done, it is incumbent upon the tester to approximate as closely as possible desirable conditions.

A common complication that arises when testing individuals at home is that there are multiple family members who must be attended to. The presence of a number of individuals, particularly children, may complicate the testing process. In such situations, we recommend that several staff members participate in the evaluation session. One of these staff members can conduct the test evaluations, while one or more of the others can either provide services to other family members (if appropriate) or watch children.

Subjective impressions. As we noted earlier, valuable evaluation information can be gathered from the impressions of program staff and the participants themselves. Some of the considerations discussed earlier with respect to standardized tests, such as the importance of handling multiple family members in the home when forming impressions, also apply to collecting subjective data. However, there are also considerations that are specific to this mode of information gathering.

One consideration that parallels the need for standardized administration of tests is that, to what ever extent possible, the impressions should be gathered in a systematic and consistent way over time. For example, if program staff provide a global judgement of the client's progress every month, it is desirable for the impressions to be gathered in approximately the same environment each time (e.g., after a home visit without the children present), that the evaluation to be made on the same rating scale (e.g., 1 = no progress to 5 = excellent progress), and that the ratings be made at predetermined time intervals (e.g., every six months).

Another important consideration is that those who provide their impressions should attempt to be as objective as possible. Particularly when one provides subjective impressions about progress in a program, there are pressures to believe and to report that substantial improvement has occurred. These pressures, which apply both to program staff and to participants, may lead to social desirability responding--responding in a more positive light than is warranted. Because of these compelling pressures, it is particularly critical that individuals who give their subjective impressions be aware of the likelihood of social desirability responding and try as hard as possible to give objective ratings. Only if these ratings are valid can they be useful in evaluating the program.

Ethical Issues in Gathering Evaluation Information

Conducting the evaluation in an ethical manner is very important. In general terms, an evaluation can be conducted in an ethical manner when the information is gathered from program participants in a respectful and sensitive way. In addition, there are some other specific ethical considerations that should be attended to.

One important consideration relates to the doctrine of informed consent. Acquiring informed consent means that the individual participating in the program has been provided with sufficient information (about the program, about the evaluation process, about how the evaluation results will be used, who will have access to the information, and about how he or she will receive feedback on the results) to decide whether he or she wishes to take part in the evaluation. It should not be assumed that individuals will automatically consent to having information gathered about them because "it is mandated" that they do so. Rather, they have the ethical right to make this choice and by signing an informed consent form they express this choice. Program administrators must decide (early in the program planning process)

ERIC Full Text Provided by ERIC

whether individuals can still receive program services if they choose not to participate in the evaluation.

Another ethical issue that must be addressed is the need to maintain confidentiality. Program participants should be assured that their evaluation data will remain confidential (i.e., known only to program staff), will not be released to outside parties without their consent, and only summary data (e.g., means for program participants on a measure of literacy) will be used in reports. Program evaluators should make sure that all possible steps are taken to maintain confidentiality. For example, completed tests or rating scales should be safely stored so they are not accessible to either other program participants or other individuals.

When participants are administered testing instruments or have their progress rated by program staff, we maintain that they deserve some feedback on their performance and on how the information can help them improve their literacy skills. This feedback should be conducted in a structured way (e.g., during a regularly scheduled program session) and in language that program participants can readily understand. Given the relatively low level oi literacy skills characteristic of those in literacy programs, this feedback must be very simply and clearly formulated.

Available Measurement Instruments

As noted earlier in this manual, both subjective impressions and standardized instruments are useful in assessing the effectiveness of literacy programs. In this section, each of these approaches to measurement is discussed.

Subjective Impressions

The program staff members who are in regular contact with participants have a unique perspective on how the program is affecting clients. For example, program staff are often aware of daily mood fluctuations, level of motivation to learn literacy skills, attitudes towards the program and program staff, relative strengths and weaknesses in reading and writ g, and the environmental and familial influences on the participant.

From our discussions with Even Start program staff, the following were identified as indicators that parents and children were making good progress: (a) parents and children interacted more frequently; (b) parents used more effective means of discipline (e.g., reasoning with the child rather than spanking); (c) there were signs of enhanced self-esteem, including that participants were more assertive; (d) children were less shy and more sociable; (e) personal grooming improved for both parents and their children; (f) the participants seemed to be gaining an expanding view of their environment, including a greater interest in local, regional, and national news stories; (g) increased enthusiasm for learning; (h) an increased desire to try new things and to add variety to their lives; (i) parents and children report that they are changing in positive ways, and (j) the participants engaged in more effective and deliberate decision-making processes. All of these indicators provide useful evaluation information and some lend themselves to systematic rating at regular intervals.

Program staff have identified the following as signs that services were not effectively helping participants: (a) participants seemed to not be motivated; (b) participants cancel or miss appointments; (c) parents are motivated for the child to improve, but not for themselves; and (d) parents respond reactively and passively, rather than proactively, to the opportunities provided by the program.

We recommend that staff members routinely record their observations in these and other areas. The records can take the form of general notes that are made on a weekly or biweekly basis, or can be more formal (e.g., rating the client's level of motivation on a 1 (very low) to 5 (very high) scale after every program contact). Not only is this information of value in itself, but these subjective impressions may help program evaluators interpret the results of the standardized test instruments (see below). To aid in the collection of this important information, it is critical that program staff, along with program evaluators, come to some mutual agreement on which observations will be systematically recorded.

Standardized Instruments

Following are a selection of standardized tests that can be used to assess literacy aptitude (i.e., an individual's potential) and/or achievement (i.e., the individual's actual level of accomplishment). The tests span a wide developmental range from preschool through adult. All the tests include norms for respective ages/grade-levels. In most cases "specimen" test packages can be obtained for evaluation by contacting the publisher (see Appendix C).

NAME: **Basic Achievement Skills Individual Screener (BASIS)**
AGES/LEVEL: Grades 1 through 12 and post-high school
DESCRIPTION: This individually administered test assesses reading, mathematical, and spelling skills. Reading is assessed through comprehension of passages. At the lower levels comprehension is assessed through word or sentence reading or through letter identification. The mathematics test focusses on computation and problem-solving.
TIMING: untimed, usually less than one hour
PUBLISHER: The Psychological Corporation (1983)

NAME: **Test of Written Language-2 (TOWL-2)**
AGES/LEVEL: ages 7 1/2 through 17
DESCRIPTION: This test uses both essay analysis and traditional formats to assess thematic expression, vocabulary, syntax, spelling, and style.
TIMING: about 65 minutes
PUBLISHER: The Psychological Corporation (1988)

NAME: **Metropolitan Readiness Tests, Fifth Edition (MRT)**
AGES/LEVEL: Preschool through Grade 1
DESCRIPTION: These tests assess skills that are important for early school learning, particularly reading, mathematics, and language development. Integral parts include the "Early School Inventory-Preliteracy," a checklist that assesses a child's progress in acquiring skills needed to learn to read and write.
TIMING: 80-100 minutes depending on level
PUBLISHER: The Psychological Corporation (1986)

NAME: **Preschool Language Scale-3 (PLS-3)**
AGES/LEVEL: Birth to 7 years
DESCRIPTION: This test assesses a broad range of receptive and expressive language skills including syntax, morphology, vocabulary, concept development, and cognitive skills.
TIMING: 20-30 minutes
PUBLISHER: The Psychological Corporation (1992)

NAME: **Boehm Test of Basic Concepts - Preschool Version**
AGES/LEVEL: 3 to 5 years
DESCRIPTION: This test assesses a child's understanding of 26 basic concepts that are considered necessary for success in the beginning years of school. The concepts refer to relational characteristics of persons and objects such as size, direction, position in space, quantity, and time.
TIMING: 10-15 minutes
PUBLISHER: The Psychological Corporation (1986)

NAME: **The Educational Testing Service (ETS) Tests of Applied Literacy Skills**

AGES/LEVEL: Adult

DESCRIPTION: These sophisticated tests of functional literacy measure an individual's facility in understanding and using printed and written materials. The tests were refined from those used by the U.S. Departments of Labor and Education to assess adult literacy on a national level. The tests measure prose literacy (examples of text), document literacy (understanding forms, tables, schedules, maps, etc.), and quantitative literacy (various everyday arithmetic operations, such as balancing a checkbook, completing an order form, or determining the amount of interest on a loan).

TIMING: 45 minutes for each test

PUBLISHER: Simon & Schuster Workplace Resources (1991)

NAME: **Peabody Individual Achievement Test (PIAT)**

AGES/LEVEL: age 5 through adult

DESCRIPTION: This individually administered screening test provides an overview of scholastic achievement. The 402 items cover mathematics, reading recognition and comprehension, spelling, and general information.

TIMING: 30-50 minutes

PUBLISHER: American Guidance Service

NAME: **Comprehensive Adult Student Assessment System (CASAS)**

AGES/LEVEL: Adult

DESCRIPTION: The CASAS has been designated for use by agencies participating in the national Even Start program evaluation project. The tests are designed to assess various competencies for all levels of adult basic education, English as a second language, and adult special education. The tests focus on functional literacy, measuring reading and understanding of forms, charts, graphs, maps, paragraphs, sentences and directions, labels, and advertisements.

TIMING: untimed, usually less than one hour

PUBLISHER: Foundation for Educational Achievement (1991)

Program Evaluation Data and Analysis

The following information on data and how to analyze them will be familiar to most program evaluators. Further, a program evaluator will have the expertise to use the data analytic techniques reviewed in this section. However, as we have noted earlier, it is important that Even Start staff have some familiarity with these concepts so that they can communicate with the program evaluator.

Compilation of Measures - the Student Portfolio

Program evaluation data are derived from the particular measurement indicators that are used. In the case of Even Start, it is likely that there will be both subjective and standardized indicators. One useful approach to organizing this diverse information is to compile the measures in a student "portfolio," which is basically a collection of student work done over time. Portfolio contents are likely to be varied and may include samples of the student's assignments, test scores, staff impressions, and self-evaluations.

Once primarily associated with the fine arts, portfolios are increasingly accepted in many educational settings as a valid means of compiling evaluation data. The critical feature of portfolios is that their contents may vary depending on the student's needs, accomplishments, and circumstances. Because of this flexibility, portfolios can provide unique insights into a student's progress.

Portfolios are well-suited to a program like Even Start, in which a variety of indices of student performance and development are available. Some suggested contents for an Even Start portfolio are:

(1) performance record on a standardized test battery,

(2) excerpts from a parent's diary describing his/her experiences with shared reading,

(3) written self-evaluations and evaluations of the Even Start Program,

(4) staff members' written impressions of a student's progress,

(5) staff members' notes following a program session,

(6) a journal of book reviews,

(7) a scrapbook of movie reviews,

(8) completed crossword puzzles, and

(9) pieces of original short fiction or poetry.

Collection of Data

There are two types of data that are used in an evaluation of a program like Even Start. The first type includes numbers that indicate amount, quantity, or degree of something. Examples might be household income, age of child enrolled in the program, score on a standardized test of literacy skill, a teacher's quantitative appraisal of a student's portfolio, or the number of books or magazines read each month.

The second type of data includes classifications that may be useful as a basis for comparison. Examples might be whether the child is male or female, whether or not English is a second-language of the participant, or the types of reading materials that are kept in the home.

Whatever data are ultimately collected, their accuracy is of critical importance to a successful program evaluation. Any conclusions regarding the effectiveness of a program can only be accurate if the data and information that go into the program evaluation are themselves accurate. If accuracy is to be assured, the program evaluator must maintain a high level of care and vigilance in the recording of data. Even an error of a few points can lead to an inaccurate assessment, distort the results, and contribute to invalid conclusions.

Storage of Data

In addition to accuracy, it is also important that program evaluation data be carefully archived or stored. Moreover, the data should be stored in a form that can be accessed by computer, which is the most powerful tool for data analysis. Because the various available computer programs may have unique input requirements, we provide here only general guidelines for the preparation of data for analysis on a computer. For this reason it is wise to have at least one member of the program evaluation team who is technically knowledgeable about how to use computers for data analysis.

In the most usual case, data are entered into the computer via the keyboard and stored on some magnetic storage device suc'. as a hard drive and/or floppy diskette. Data are saved and stored as a "file," which is a list of outcome-relevant numbers and other information. It is critical that this file be rigidly organized, which means that the various data must be assigned unique locations within the file. For example, if age of the child is to be recorded, it must be located in the same column(s) for each child. To continue the example, one might record age of child in columns 1-2, parents' household income in columns 3-7, years of formal education of mother and father in columns 8-9 and 10-11, respectively, number of books in the home in columns 12-14, mother's score on a standardized literacy test in columns 15-17, and so on, until all the relevant data for each family are recorded.

Because of the possibility of human error or computer malfunction, it is prudent to guard against lost or contaminated data. The best way to achieve this is to maintain "back-up" copies of the data file, one version on the hard drive and a copy on floppy diskette, for example. If the data are inadvertently erased or there is a hard drive malfunction, the program evaluator can restore the file using the back-up copy.

Statistical Techniques for Analysis of Program Effectiveness

There are various statistical procedures that can be used to test whether program outcomes have been achieved. However, choosing an appropriate procedure is a complicated decision. For one thing, the appropriate choice depends on the question that is to be answered, e.g., did the child's literacy level change over the first year of the program? Second, the appropriate choice depends on the kinds of data that are available. Certain procedures require that the data consist of test scores, for example. Other procedures can accommodate qualitative information, such as the various categories of reading materials found in the home.

In general, there are two classes of statistical procedures, descriptive and inferential. Descriptive statistics are used to summarize data and include various "averages" and measures of variability, i.e., the degree to which the data differ across individuals. Descriptive statistics are frequently shown in tables or graphs to provide a concise summary of the findings contained in the data file. Descriptive statistics also include powerful correlational procedures, which yield measures of relatedness among groups of data or variables. For example, a correlation statistic can be computed that reflects the relationship between number of hours spent reading to children and the childrens' scores on a literacy test.

The second class of statistical procedures, inferential, permits the program evaluator to test specific hypotheses concerning outcome. By testing hypotheses, inferential statistical procedures allow the investigator to describe not only the specific sample studied, but also to draw inferences about whether other participants who may enroll in similar programs are likely to benefit.

Most inferential statistical procedures provide information that enables the evaluator to decide whether an observed outcome was likely due to chance or whether it was due to a feature of the program. For example, an inferential statistical procedure could be used to test the relative effectiveness of two forms of tutoring, in-home and on-site, on literacy test scores. The procedure would yield the following determination: Is it reasonable to assume that the observed difference in test scores was indeed due to the relative effectiveness of the two forms of tutoring (i.e., the result was statistically significant), or was the difference too small to preclude chance as an explanation (i.e., the result was nonsignificant)? If it can be concluded that the observed difference in test scores was statistically significant, then one can infer that one form of tutoring is not only more effective in this particular sample, but also will be for other samples of similar participants.

Handling large data files and executing statistical procedures on those files can be complex and time-consuming. For this reason, a computer and software are frequently used to perform the necessary calculations. Statistical software programs, such as the powerful, though expensive, Statistical Package for the Social Sciences (SPSS-X, Inc., Chicago, IL) and Statistical Analysis System (SAS), are widely available for micro-, desk-, or lap-top computers, as well as for larger main-frame machines. Most of these software packages contain a sample of the most useful descriptive and inferential procedures, of which there are hundreds.

The two major components in the analysis of program effectiveness--choosing a statistical procedure that addresses an outcome-relevant question and executing that procedure on the data file to answer the question--are complicated. It is recommended that a program evaluation team include the skills of someone with training in both statistics and computer operations.

Reporting Evaluation Results

There are several issues to consider when deciding how to present the results of evaluation studies. These include the sections to include in evaluation reports, the relevant content for these sections and tailoring the content for the relevant audience.

Sections to Include in Evaluation Reports

The key sections that should be considered for inclusion in an evaluation report are as follows: introduction, methodology, results, conclusions, references, and appendices. Each is described briefly below.

Introduction. This section provides the context for understanding the need that the program addresses, how this programmatic effort relates to previous work, and some description of the nature of the present program. In the area of family literacy, this section would include a description of the literacy problem that requires attention, interventions that have been previously used to address literacy deficits, how the current program operates, and how this program is an improvement over previous ones.

Methodology. This section describes the evaluation process in sufficient detail that other evaluators could replicate (i.e., conduct in exactly the same manner) the evaluation design. A "participants" subsection provides basic demographic and descriptive information on program participants and, if one is used, the comparison group. A "measures" subsection reviews all the instruments used in the evaluation, including their purpose, format, norms, reliability, validity, and appropriate uses. A "procedure" subsection describes how subjects were recruited, how they were tested, what they were told about testing, and any other information relevant to understanding the results. An optional "data analysis" subsection reviews the statistical approaches that were used to analyze the data.

Results. This section details the results of statistical analyses of the data. Depending on the research design, different types of statistical procedures can be employed. At the simplest level, descriptive information (i.e., means, standard deviations) should be provided for how the participants performed on the test instruments at each time period that they were tested. This information can be compared to published norms on the instruments, if they are available.

Assuming that participants were tested on more than one occasion, statistical tests should be conducted to determine whether the changes were statistically significant. If a comparison group of individuals who did not participate in the program was tested, sophisticated procedures are available to determine if those in the "program" group improved to a greater extent than did those in the "no program" group.

The basic purpose of evaluation reports is to present the findings in a clear and understandable manner to the appropriate audience. Rather than presenting the results of statistical analyses in the text, it is helpful to present these in either graphs or tables. If the information in the graphs or tables is central to understanding the text, they should be placed (on separate pages) in appropriate places in the results section. If the information is not central, graphs and tables can be placed in the appendix.

Tables consist of rows and columns of statistical results displayed so that they convey some key aspect of the evaluation results. For example, a common version of a table contains the means and standard deviations on some assessment instrument for the different groups in the study (e.g., "program" vs. "no program"). The table should be constructed so that it is interpretable without reference to the text and it should be cited in the text (e.g., "see Table 1"). An example of such a table, with fictitious results, is presented in Table 1.

Figures display the results of statistical analyses in graphical form. Depending on the nature of the evaluation design, figures can consist of bar graphs, histograms, polygons, and curves. If the data lend themselves to this format and if the graphs are constructed well, figures have the potential to be of great assistance in helping readers understand the results. Examples of such figures, again with fictitious data, are shown in Figures 1 and 2.

Conclusions. This section is perhaps the most important of the report, particularly for those who do not have extensive expertise in evaluation strategies. In fact, some readers of the report may only examine this section. In this section, the results of the evaluation are summarized, the results are related to those from previous studies, some interpretations of these results are presented, strengths and limitations of the evaluation design are identified, and implications for programmatic planning are discussed. Throughout this component of the report, information is presented in a manner that is understandable to those who lack evaluation sophistication.

References. This section provides the full citations for any articles, books, or other sources that were identified in the report.

Appendices. Appendices contain information that is not central to the body of the text. Candidates for inclusion in the appendices include copies of noncopyrighted test instruments that were administered in the evaluation, tables that provide descriptive

Table 1

Means and Standard Deviations on Evaluation Instruments at Pretest and Posttest by Program Group

Measure		Program		No Program	
		Pretest	Posttest	Pretest	Posttest
X	Mean	76.2	96.7	78.4	83.8
	SD	6.8	7.7	6.9	6.4
Y	Mean	8.2	9.4	7.9	8.3
	SD	1.1	0.9	1.0	1.1
Z	Mean	324.8	362.4	315.1	316.2
	SD	33.2	36.5	29.7	31.6
N		40	36	35	29

Fig. 1

ETS Tests of Applied Literacy Skills

(Mean Scaled Scores)

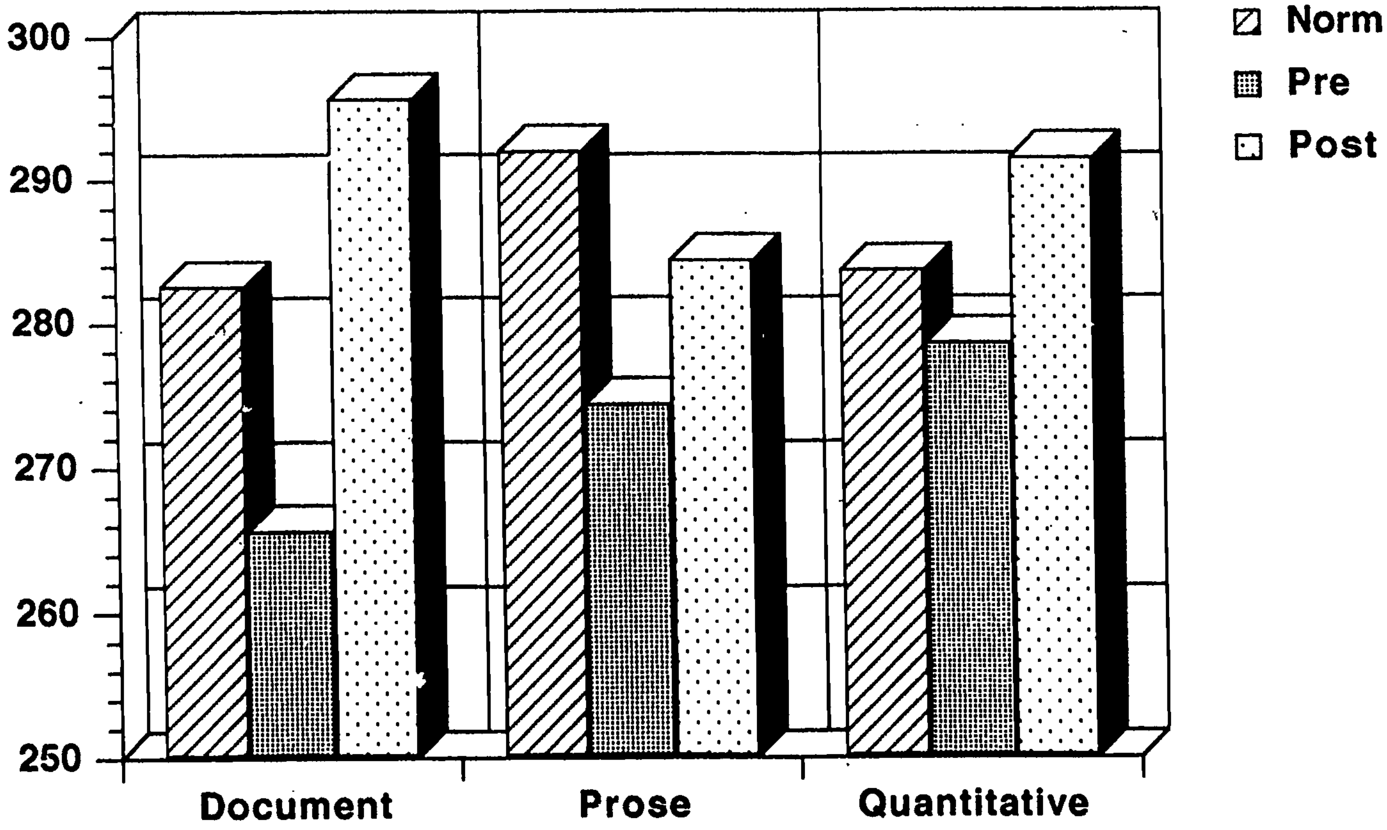
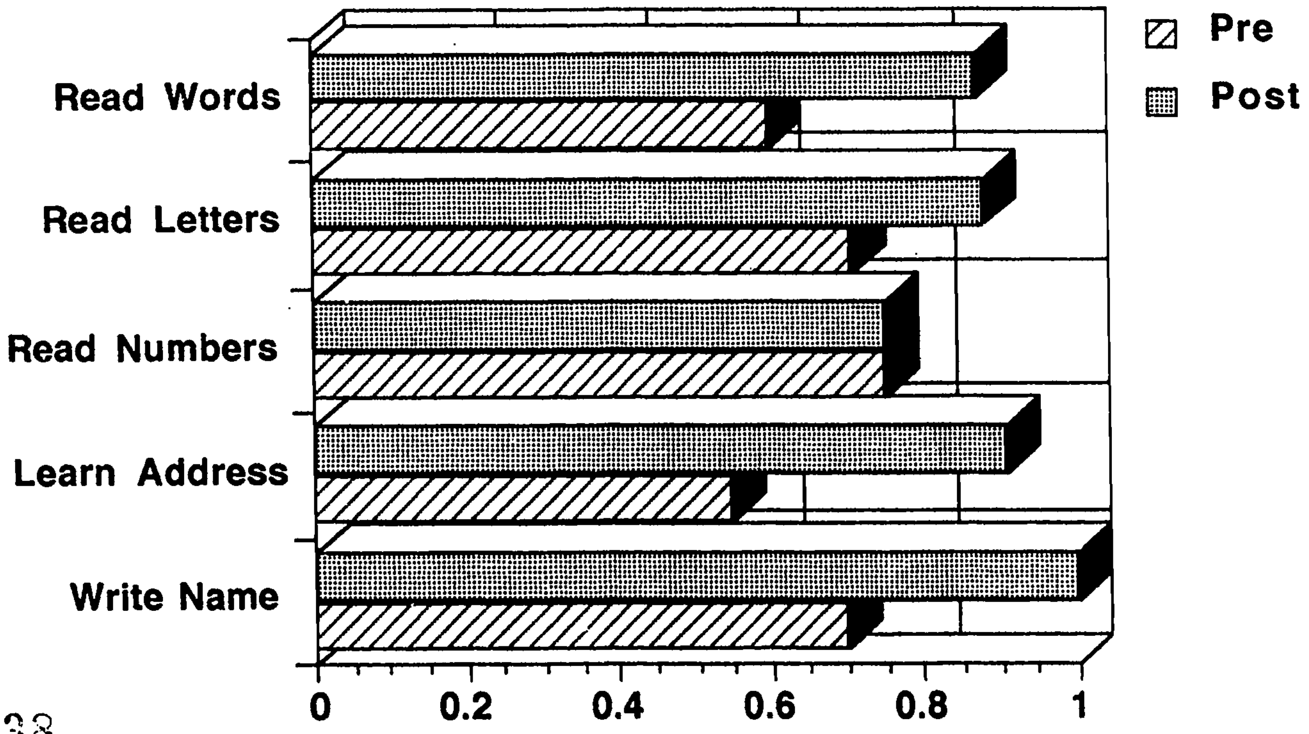


Fig. 2

Proportion of Parents Who Helped Child With Activities



information on the sample and how it performed on the test instruments, and any other documents that may be relevant to the reader.

Tailoring Content to Relevant Audience

Although this message has been communicated elsewhere in this manual, it bears repeating. Evaluation reports must be written in terms that are comprehensible to those in the intended audience. If this practice is not followed, the reports will not be read, their potentially important implications ignored, and program planners will be less likely to consider evaluation important in future studies.

The primary dimension on which different audiences vary is evaluation sophistication. On the one hand, if the intended audience is a group of program evaluators, the report can and, indeed, should present the methodology and results section in detailed form. Members of this particular audience will want to know the intricacies of how the evaluation was conducted and the statistical procedures that were employed. They may be less interested in the background information contained in the introduction.

On the other hand, if the intended audience is a group of program administrators, the conclusions section should be emphasized and written in a manner appropriate for those with expertise and experience in literacy programming. The methodology and results section can be relatively brief and nontechnically presented, with more complete information presented in the appendices. In fact, for such audiences, an "executive summary," which is a condensed report that focuses on highlights and recommendations, may be helpful.

We would also like to emphasize that an evaluation report is seldom by itself sufficient to inform most readers. Before and after completion of a "final" written report, we encourage evaluators to schedule sessions or appointments to present the evaluation results to the appropriate constituency, to highlight key areas, issues, or findings, to respond to questions and comments, and to receive programmatic information that may influence the interpretation of the results. In this way, the completion of the evaluation report is a collaborative effort on the part of program administrators, staff, and evaluators and all constituencies feel that they have had input into the final product. The collaboration and joint ownership increase the chances that the evaluation findings will have their intended impact--to improve the effectiveness of programmatic services.

Using Program Evaluation Results

In the course of sharing evaluation results with program administrators and staff (and certainly afterwards as well), it is critical that the findings be used to inform the ongoing process of program planning. This section of the manual provides a brief overview of how program evaluation results may be used to achieve this end.

If implemented, the evaluation strategies reviewed in this manual can provide an extensive array of data that bear on the question of how effective the program is in enhancing literacy. These data include the results of standardized tests, the subjective impressions of program staff, measures that are thought to promote the development of literacy (e.g., bringing books, magazines, and newspapers into the home), and individualized (i.e., portfolio) information. Based on these data, several specific research questions can and should be asked when planning for possible changes in literacy programs. A sampling of these questions is provided below, along with some examples of how the answers to these questions may inform program decisions.

Is The Program as a Whole Effective?

The most basic question--and one that has been discussed earlier in this manual--is whether the entire program package has a beneficial effect on the development of literacy skills. To answer this important question, one can examine changes over time (relative to a comparison or control group if available) on the chosen measures of literacy development. There are three possible answers to this question: the program shows positive impact on all outcome measures, on some outcome measures, or on no outcome measures.

If there are positive changes on all outcome measures, this is clearly a desirable position for a program administrator to be in. However, this situation does not necessarily mean that no changes should be made in the j rogram. As discussed further below, it is possible that some, but not all, components of the program were responsible for the beneficial impacts that were observed. If this is the case, program planners might consider eliminating the least influential program components.

If there are positive changes on some--but not all--outcome measures, a careful analysis needs to be conducted to determine which measures showed improvement and which did not. The results of this analysis may shed light on what programmatic changes may be helpful, as well as possibly indicating ways in which the goals of the program can be clarified or refined and the evaluation strategy modified. For example, suppose that there were positive changes in staff members' subjective impressions of progress, but no such progress was noted on the standardized test instruments. This may indicate that staff members detected fine-grained changes that the standardized tests were not sensitive to. If this is the case, then one would expect that improvements in standardized test scores would emerge later. As such, an implication for program development might be to extend the length of the intervention. On the other hand, it may be that staff members were overly generous in their impressions of progress, which may indicate a need for further staff development.

No program administrator would like to be in the position of having to defend a program that shows no improvement on any outcome measures. Nevertheless, even when this undesirable situation occurs, it can provide useful information. For example, discussions with program participants and staff may reveal progress in an area untapped by the measures chosen for the program evaluation. The measures may not be sensitive or specific enough to adequately assess participants' progress, or improvement may have been shown had the program been extended. Each of these possibilities has important and useful implications for program refinement.

Are Some Components of the Program More Effective Than Others?

It is quite likely that some program components are more effective than others. In Even Start, fr example, it is possible that a home-based component has a more positive impact on participants than does a center-based component. It is important to systematically assess this possibility. The t:vo key implications are that one influential component (e.g., home-based services) could be expanded and/or extended in time, while a less effective component (e.g., center-based services) could be deemphasized or possibly eliminated.

In making these determinations, one can use the entire range of evaluation data that have been generated. In particular, if a program component is especially helpful for participants, one would expect that there would be greater improvements in outcome measures while participants are involved in this component than in the others. If all program components are similarly helpful, one would expect that improvements would be of approximately the same magnitude throughout all components of the program.

Which Process Variables are Most Strongly Related to Literacy Outcomes?

As noted earlier, it is important to assess processes that are expected to lead to later improvements in literacy skills. Such processes may include how often participants attend their scheduled classes or participate in home intervention sessions, whether they bring literacy materials (e.g., books, magazines) into the home, the staff's impressions of participants' motivation to participate in literacy training activities, and how often participants seek additional help from program staff. By examining how strongly each of these process variables is related to literacy development, one can acquire a sense of which processes should receive greater or lesser emphasis in future program decisions.

For example, suppose that there is a very high association between how often participants are in their homes for scheduled home visits and their scores on standardized literacy measures. Although the meaning of this observed relation is complex, one possible implication is that greater attention should be devoted to ensuring that participants are home during the scheduled sessions. On the other hand, suppose that there is no observed relation between the number of center-based sessions attended and staff impressions of literacy development. A possible implication is that actual attendance at sessions in the center is not necessary for the development of literacy skills. Although these examples chosen to illustrate the possible links between process and outcome are overly simplistic, they illustrate the possible uses that process data can have in program modification.

ERIC

APPENDIX A

ANNOTATED BIBLIOGRAPHY OF SELECTED BOOKS, ARTICLES, AND REPORTS ON LITERACY

Auerbach, E. R. (1989). Toward a social-contextual approach to family literacy. Harvard Educational Review, 59(2), 165-181.

This paper examines family literacy programs that teach parents to assist their children with school activities and assignments. The author argues that the rationale and design of these programs is not based on sound research. She proposes a broader, alternative perspective that focuses on the family's strengths and its socio-cultural milieu.

Baydar, N., Brooks-Gunn, J., & Furstenberg, F. F. (1993). Early warning signs of functional illiteracy: predictors in childhood and adolescence. Child Development, 64, 815-829.

This paper describes a 20-year longitudinal study of 251 black children of teenage mothers to determine the determinants of functional literacy in adulthood. It was found that preschool cognitive and behavioral functioning was highly predictive of literacy in young adulthood. Certain family environmental factors were also found to be predictive (e.g., maternal education, family size, income).

Congress of the United States House Committee on Education and Labor. (1991). National Literacy Act of 1991 (Report House-R-102-23). (ERIC Document Reproduction Service No. ED 340 889).

This document contains the text of the National Literacy Act of 1991, which is designed to improve adult literacy and basic skills. The legislation allows for coordinating, integrating, and investing in adult and family literacy programs at the federal, state, and local levels. It also provides for research and program delivery.

Daisey, P. (1991, March). Intergenerational literacy programs: Rationale, description, and effectiveness. Journal of Clinical Child Psychology, 20(1), 11-17.

This paper presents a rationale for combining programs in adult literacy programs with those of early childhood education. The rationale stresses the importance of the home environment, of parent-child shared reading activities, and of parents' attitudes toward education. Three programs, including Even Start, are described, along with evidence for their effectiveness.

Daiute, C., et al. (1993). Special issue: the development of literacy through social interaction. New Directions for Child Development, 61.

This special issue contains a group of related articles that conclude the following: (a) children become literate in the context of relationships, (b) literacy depends on oral discourse, and (c) literacy should be seen as a set of social functions rather than as a group of linguistic skills.

Ditmars, J. W. (1993). A field guide for literacy: life skills and literacy for adult beginning readers and ESL students. Manual for teachers and tutors. Bethlehem, PA: Northhampton Community College, Adult Literacy Division. (ERIC Document Reproduction Service No. ED 359 820)

This manual provides specific lesson plans for teaching life skills and literacy to adult beginning readers and English-as-a-Second-Language (ESL) students. A total of 95 topics are presented across four curricular sections: language arts, life skills and literacy, holidays and observances, and survival math.

Ehringhaus, C. C. (1990). Functional literacy assessment: Issues of interpretation. Adult Family Quarterly, 40(4), 187-196.

This paper discusses problems and issues in the measurement and interpretation of functional literacy performance. It includes a brief overview of the definition of functional literacy, a synopsis of some of the widely known functional literacy tests, and guidelines for interpreting test performance.

Frager, A. M. (1991). Adult literacy assessment: Existing tools and promising developments. Journal of Reading, 35(3), 256-259.

This brief paper describes several standardized tests of adult literacy, as well as some less-formal alternative techniques. The paper also considers promising developments for the future.

French, J. (1987). Adult literacy: a source book and guide. New York, NY: Garland Publishing, Inc.

This book is designed for use by those involved in the administration and funding of literacy programs and by literacy theoreticians and researchers. The book is divided into two parts. Part I reviews the literature on literacy definitions, models of reading instruction, and models of adult learning. Part II consists of an annotated bibliography covering adult basic literacy,

literacy and the older adult, English as a second language, literacy in the workplace, literacy in postsecondary institutions, and literacy around the world.

Kirsch, I. S., & Jungeblut, A. (1986). Literacy: profiles of America's young adults (Report No. 16-PL-02). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

This report summarizes the results of a survey of the functional literacy skills of a nationally representative sample of approximately 3,600 young aduits (ages 21-25). The survey assessed proficiency on tasks encountered in a variety of settings such as reading and interpreting newspaper articles, magazines, and books, identifying and using information located in forms, tables, and charts, and applying numerical operations to information in menus, checkbooks, and advertisements. The "ETS Tests of Applied Literacy Skills" (see Available Measurement Instruments, above) is a refinement of the instrument used in this survey.

Lobdell, J. E., & Schecter, S. R. (1993). Video resources for the teaching of literacy: an annotated bibliography. Berkeley, CA: Center for the Study of Writing. (ERIC Document Reproduction Service No. ED 362 890)

Responding to the expressed need of classroom teachers and teacher educators for a listing of video resources that can be used for preservice and inservice education for literacy providers, this annotated bibliography identifies video resources that portray literacy teachers and learners in action.

Miller, G. A. (1988). The challenge of universal literacy. Science, 241, 1293-1299.

This theoretical article reviews research by educators and psychologists that has laid a scientific foundation on which new methods for teaching literacy skills can be based. Even with better teaching, however, the hope that all adults can attain the highest levels of skills may be unrealistic.

Moeller, B. (1993). Literacy and technology. Technology in Education, 2, 1-4.

Current efforts to improve literacy teaching and learning are directed at replacing task-oriented approaches to teaching isolated skills with an integrated language arts curriculum, which focusses on cognitive and social processes. Computer-based technologies are an important part of this new approach.

Morrow, L. M., et al. (1993). Family literacy: perspective and practices. Reading Teacher, 47, 194-200.

This paper offers a general perspective on family literacy, defines family literacy, discusses family literacy initiatives, future directions, program planning and initiatives, and dissemination activities.

Padak, N. D. & Padak, G. M. (1991). What works: Adult literacy program evaluation. Journal of Reading, 34(5), 374-379.

This paper reviews criteria that researchers have presented for assessing adult literacy program effectiveness.

Popp, R. (1991). A guide to funding sources for family literacy. (Available from the National Center for Family Literacy, 401 South 4th Avenue, Suite 610, Louisville, KY, 40202-3449 for $5.00). (ERIC Document Reproduction Service No. ED 340 875).

This guide presents a variety of useful information: (1) the major funding sources (federal, state/local, private) for family literacy programs; (2) funding strategies; (3) guidelines for writing proposals, including sample budgets and checklists; and (4) a list of books and information centers.

Rhodes, L. K. (1992). Anecdotal records: A powerful tool for ongoing literacy assessment. Reading Teacher, 45(7), 502-509.

This article argues that anecdotal records can be an important source of information about the effectiveness of literacy programs. The records can be used to evaluate student progress, guide instructional planning, and suggest new assessment strategies. The author also discusses techniques for collecting and analyzing anecdotal records.

Rickard, P.L. (1991, April). Assessment in adult literacy programs. Paper presented at the Adult Literacy Assessment Workshop, Philadelphia, PA. (ERIC Document Reproduction Service No. ED 337 575).

This paper recommends that adult literacy program evaluations not only monitor student skill level, but also provide information about student needs and placement, and program certification. An example is described--the Comprehensive Adult Student Assessment System (CASAS)--which was developed by the California Department of Education. CASAS is to be used as part of the national evaluation plan for Even Start.

Ryan, K. E. et al. (1991, April). An evaluation framework for family literacy programs. Paper presented at the 72nd Annual Meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 331 029).

This paper presents a five-component framework for family literacy program evaluation: (1) needs assessment; (2) program utilization/accountability; (3) formative/process evaluation; (4) evidence of progress towards objective; and (5) evidence of program impact. It is recommended that evaluation be based on a performance-based assessment "portfolio" rather than a mere listing of standardized test results.

Santopietro, K. & Peyton, J. K. (1991). Assessing the literacy needs of adult learners of ESL (Report No. EDO-LE-91-07). Washington, DC: Office of Educational Research and Improvement. (ERIC Document Reproduction Service N. ED 334 871).

This report focuses on ways to determine what learners want or believe they need to learn. Specific attention is focused on the definition of a needs assessment, the importance of a needs assessment, and assessment tools. In addition, a needs assessment conducted in one adult literacy program, the Adult Literacy Evaluation Project in Philadelphia (PA), is summarized.

Seaman, D. et al. (1991). Follow-up study of the impact of the Kenan trust model for family literacy. (ERIC Document Reproduction Service No. ED 340 479).

Family literacy programs in Indiana, West Virginia, and Kentucky were evaluated using parent interviews, teacher reports, and child academic achievement test scores. The programs were designed around the Kenan Trust Model, a process that is utilized by the National Center for Family Literacy. It was found that parents in the programs developed a positive self-concept, helped their children with homework, attended school functions, increased their understanding of teachers' problems, and read to their children. Teacher reports indicated that about 90 percent of the program students were doing at least as well as other students in terms of attendance, academic achievement, peer interaction, and motivation to learn.

Shermis, M. (Ed.). (1991). Parents and children together. Volume 1, Nos. 1-12. (Available from the Family Literacy Center, Indiana University, 2805 East 10th Street, Suite 150, Bloomington, IN, 47408-2698 for $7.00). (ERIC Document Reproduction Service No. ED 329 942).

This is a useful series of booklets and companion tapes that are meant to be shared by parents and children. The twelve booklets/tapes cover the following topics or themes: (1) family storytelling; (2) motivating your child to learn; (3) self-esteem; (4) reading and writing; (5)

discipline; (6) holidays; (7) science in the home; (8) recreation and health; (9) folktales; (10) mathematics in the home; (11) summertime; and (12) parents as models.

Smith, P. H., Balian, L. R., Brennan, D. E., Gorringe, J. L., Jackson, M. S., & Thone, R. R. (1986). Illiteracy in America: extent, causes, and suggested solutions. Washington, DC: U.S. Government Printing Office.

The report, by the National Advisory Council on Adult Education Literacy Committee, discusses the extent and causes of illiteracy in America and offers suggestions for preventing illiteracy in the future. The suggestions include improved reading instruction, strict evaluation of educational materials, upgrading curricula, the instilling of positive attitudes, objective evaluation of preschool programs, early screening, standardized testing program improvement, and determining the factors that improve student achievement.

St. Pierre, R., et al. (1993). National evaluation of the Even Start Family Literacy Program: report on effectiveness (Report No. ED/OUS-93-47). Cambridge, MA: Abt Associates, Inc. (ERIC Document Reproduction Service No. ED 365 476)

This evaluation report is the third in a series or reports that are part of a 4-year national effort designed to describe the types of Even Start projects that have been funded, the services provided, the collaboration efforts undertaken, and the obstacles to program implementation that have been encountered. The current report provides information about the first two cohorts of Even Start projects, 76 that began in 1989 and 47 that began in 1990.

Stiles, R. E. (1991). Family literacy: An annotated bibliography and selected public library program descriptions. Unpublished master's thesis, University of North Carolina, Chapel Hill, NC. (ERIC Document Reproduction Service No. ED 332 731).

This thesis describes sources of information on family literacy for librarians and other literacy providers. The following are included: (1) a literature review and annotated bibliography; (2) the results of a survey of North Carolina public library systems concerning their involvement in family literacy programs survey; and (3) detailed descriptions of six family literacy programs including recruitment and evaluation methods, instructional materials, and funding sources.

Teale, W. H. (1988). Developmentally appropriate assessment of reading and writing in the early childhood classroom. The Elementary School Journal, 89(2), 173-183.

This article presents guidelines for assessing young children's reading and writing skills. The author contends that observational methods and structured performance assessments are more

ERIC
Full Text Provided by ERIC

appropriate than standardized tests for measuring literacy in young children. Examples of appropriate assessment strategies are illustrated.

United States Department of Education. (1990). Helpful information for literacy programs. (ERIC Document Reproduction Service No. ED 327 195).

This comprehensive report from the Department of Education contains theoretical information on literacy, current statistics, training resources, funding sources, and organizations dealing with adult literacy. Other sections of the report focus on recruitment of students and adult new readers, information regarding volunteers, ways of creating literacy councils, and methodologies used in some literacy programs. The report includes an extensive materials section containing bibliographies, book lists and reviews.

van Kleeck, A. (1990). Emergent literacy: Learning about print before learning to read. Topics in Language Disorders, 10(2), 25-45.

This study examines the effects of exposure to print during the preliterate years on emergent literacy skills. The study also reviews research findings and offers suggestions for assessment and facilitation of these skills.

Venezky, R. L., Kaestle, C. F., & Sum, A. M. (1987). The subtle danger: reflections on the literacy abilities of America's young adults (Report No. 16-CAEP-01). Princeton, NJ: Center for the Assessment of Educational Progress, Educational Testing Service.

This report reviews the results and discusses the implication of a national survey of the literacy abilities of young adults (ages 21-25), which was conducted for the U.S. Department of Education Office for Educational Research and Improvement (see Kirsch and Jungeblut, 1986). The report covers assessment techniques and methodologies, test materials, and comparative abilities of subgroups of individuals.

Winograd, P. et al. (1991). Improving the assessment of illiteracy. Reading Teacher, 45(2), 108-116. (ERIC Document Reproduction Service No. EJ 432 491).

This paper provides a rationale for why literacy assessment strategies need to be improved and presents guidelines for implementing changes in these strategies.

APPENDIX B

EVALUATION METHODOLOGY SOURCES

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

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: design and analysis issues for field settings. Chicago, IL: Rand McNally College Publishing Company.

Hopkins, K. D., Stanley, J. C., & Hopkins, B. R. (1990). Educational and psychological measurement and evaluation (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: an integrated approach. Hillsdale, NJ: Lawrence Erlbaum Associates.

Spodek, B. (1993). Handbook of Research on the education of young children. New York, NY: Macmillan Publishing Co.

APPENDIX C

PUBLISHERS OF ASSESSMENT INSTRUMENTS

American Guidance Service
Publishers' Building
Circle Pines, Minnesota 55014
1-800-328-2560

Comprehensive Adult Student Assessment System
Foundation for Educational Achievement
2725 Congress Street, Suite 1-M
San Diego, California 92110

The Psychological Corporation
Harcourt Brace Jovanovich, Inc.
555 Academic Court
San Antonio, Texas 78204-2498
1-800-228-0752

Simon & Schuster Workplace Resources
15 Columbus Circle
New York, NY 10023-7780
1-800-395-7042

APPENDIX D

ORGANIZATIONS INVOLVED IN FAMILY LITERACY

ADvancE
Pennsylvania Department of Education Resource Center
333 Market Street
Harrisburg, PA 17126-0333
1-800-992-2283

Association for Community Based Education (ACBE)
1806 Vernon Street, NW
Washington, DC 20009
(202)462-6333

Coalition for Literacy
c/o American Library Association
50 East Huron Street
Chicago, Illinois 60611
(312)944-6780

Contact Literacy Center
P.O. Box 81826
Lincoln, Nebraska 68501-1826
1-800-228-8813

ERIC Clearinghouse on Adult, Career, and Vocational Education
1900 Kenny Road
Columbus, OH 43210
1-800-848-4815

Literacy Volunteers of America
5795 Widewaters Parkway
Syracuse, NY 13214
(315)445-8000

Litline
The Adult Literacy Initiative
U.S. Department of Education
400 Maryland Avenue, Room 4145
Washington, DC 20202
(202)732-2959

National Center for Family Literacy
One Riverfront Plaza, Suite 608
Louisville, KY 40202
(502)584-1133

National Center on Adult Literacy (NCAL)
University of Pennsylvania
Philadelphia, Pennsylvania 19104-6216
(215)898-2100

Ohio Appalachian Literacy Project
321 McCracken Hall
Ohio University
Athens, OH 45701
(614)593-4470

Ohio Literacy Resource Center (OLRC)
414 White Hall
Kent State University
Kent, Ohio 44242-0001