EAC West, New Mexico Highlands University, December 1995.

EVALUATION HANDBOOK

Judith Wilde, PhD
Suzanne Sockey, PhD

Evaluation Assistance Center-Western Region
New Mexico Highlands University
Albuquerque, NM

December, 1995

Table of Contents

NOTE: ***Appendices are not included in the electronic version of this publication.***

I: Overview

> *Increasingly since the mid 1960s, funds for large programs in the public interest have been allocated with the stipulation that the programs be evaluated. Increasingly since the early 1970s, the nature of those mandated evaluations has been prescribed and regulated.*
>
> *Anderson & Ball (1978, p 212)*

Evaluation is the process of systematically aggregating and synthesizing various types and forms of data for the purpose of showing the value of a particular program. More specifically, Walberg and Haertel (1990) define evaluation as a careful, rigorous examination of an educational curriculum, program, institution, organizational variable, or policy. The primary purpose of this examination is to learn about the particular entity studied. ... The focus is on understanding and improving the thing evaluated (formative evaluation), on summarizing, describing, or judging its planned and unplanned outcomes (summative evaluation), or both. Evaluation pertains to the study of programs, practices, or materials that are ongoing or in current use. (p. xvii)

Because much of the evaluation process is based on testing, test scores, and analyzing data, many people

fear the very word *evaluation*. In fact, many programs hire evaluators specifically so that they will not have to worry about the technical nature of evaluation. This, however, also leads to a view of the evaluation as strictly for the use of others--send the evaluation report to the appropriate agency (e.g., the funding agency, the local Board of Education), ensure that copies are available for others, and continue on with the program. This approach denies the utility of evaluation and the necessity of modifying the educational program based on its current strengths and limitations.

Some authors (e.g., Popham, 1990) feel that the field of educational evaluation really came into its own as a formal specialty with the passage of the *Elementary and Secondary Education Act* (ESEA) in 1965. The purpose of this Act was to provide financial support from the federal government for the improvement of education. A great deal of money was offered to local school districts, but only on the condition that an evaluation was completed each year. Local school district administrators who had not been aware of evaluation suddenly became interested, discovered what evaluation was (or who evaluators were), and evaluated programs. Indeed, the contents of this *Handbook* are an outcome of ESEA funding through the *Bilingual Education Act*. Although now reauthorized and refocused as the *Improving America's Schools Act of 1994* (IASA), evaluation still is a requirement of IASA programs, and most other specially-funded educational programs.

Especially since the signing of IASA and *Goals 2000*, education is emerging as a major priority of our government. Leaders of school systems are being challenged to examine their educational environments and to restructure for true improvement of their educational systems. As a result, ongoing strategies for building positive educational opportunities are being explored among communities, educators, administrators, parents, and students. All shareholders are expected to redesign their schools with the purpose of enhancing teaching in order to impact the learning experiences of all students. "All students" now includes ethnically and linguistically diverse populations across the country: English learners, migrant students, American Indian students, students living In poverty, and students who are neglected or delinquent. These are students who may need unique provisions within educational settings to meet content and student performance standards as part of the educational reforms.

To maximize the change process, school leaders will be reshaping their priorities to meet the challenge. The necessary processes will include planning, implementing, assessing, and evaluating programs in accordance with IASA criteria. Along with this, they will need to show progress in program improvements. We hope that this *Handbook* will help with this process.

Previewing the *Handbook* may help some readers to identify the portion(s) that are most appropriate for their needs. The purpose of this document is to (1) offer some suggestions to these reforming administrators in the "how to" of a good evaluation, (2) alleviate some of the fear and mystery of evaluation, and (3) provide guidelines for evaluation. It is divided into five sections. Each section has a specific purpose that is described below. Each section has its own appendix at the end of the *Handbook* that includes stand-alone materials that can be shared with staff, an evaluator, or others. In most cases, these materials provide more detail than found in the text -- the text provides a brief explanation, with the stand-alone material providing greater detail. In cases some cases the materials in the appendix may provide the same information as the text -- the stand-alone materials are provided here as a briefer version of the text that can be shared quickly with others. Materials within the appendices may be photocopied for not-for-profit purposes as long as the credit line at the bottom is preserved. While directors of most programs defined within the IASA will find the information helpful, the *Handbook* is more specifically aimed toward the directors, staff, and evaluators of IASA's Title VII bilingual programs. Where possible, standards for Title I programs also are mentioned.

This first section of the *Handbook* merely provides an overview of the *Handbook* along with some definitions that might be appropriate. Included in Appendix I are the nationally accepted *Standards* that have been developed for educational evaluations (Joint Committee on Standards for Educational Evaluation, 1981). Evaluators and program staff should be aware of these standards and ensure that their own evaluations meet them. Indeed, those writing grant applications may want to read this section before beginning the writing process.

The second section, "Thinking About the Evaluation," provides background information about, and definitions of, evaluation, assessment, and analytic techniques. Various types of evaluations are described and guidelines for managing an evaluation are suggested. Working with an evaluator also is addressed. These all are topics with which program directors and evaluators should be familiar before planning an evaluation; e.g., those writing grant applications may want to read this section.

The third section provides information about planning the evaluation: how to write and modify objectives that are measurable; create management timelines, select assessments that will measure learner success in a manner sensitive to their language, culture, and gender -- as well as to the needs of the educational program; and how to select scoring methods. All of these topics pertain to the early phases of an evaluation. This section may be of greatest interest to program director and evaluators who are involved in the early phases of a funded program; some of the materials may be appropriate for staff members as well.

The next section deals with the actual implementation of an evaluation. Ensuring that timelines are met, training staff to assist in the evaluation, collecting data, analyzing data, and the specific needs of Title I and Title VII are addressed in this section. Because the program director has ultimate responsibility for all aspects of the program, including evaluation, s/he will want to read all of this section; the evaluator may be especially interested in the portion about analyzing data.

The fifth section deals with the report itself, providing guidelines for interpreting the analyses, presenting the results, making recommendations, and writing a complete and accurate report. While report writing frequently is considered the domain of the evaluator, the entire staff must understand the report and must support the results. This section is especially important for evaluators, but program directors also will need to be familiar with the information.

A well-planned, well-implemented evaluation can provide a wealth of information about the program and about the students in the program. It can determine program effectiveness, monitor the implementation of the program, motivate students, and meet funding-agency requirements. A poor evaluation can misconstrue and misinterpret student skills and knowledge, as well as staff skills and knowledge. (See Appendix I for the document "About Evaluations.") However, we must note that a good evaluation alone is not enough to ensure a good, and improving, program. In order to be successful, the school team must be involved with strategic planning, quality management, benchmarks documenting program improvement and assessing effectiveness, and evaluation which utilizes tools relevant to the total program and its composites.

IASA evaluations do have some specific requirements as stated in the Statutes and in the Education Department General Administrative Regulations (EDGAR). The Title VII guidelines will be referred to throughout the *Handbook*; some of the major requirements for Title I also will be mentioned. However, our purpose is not just to provide a set of guidelines for preparing the Title VII evaluation or the Title I evaluation. Rather, the information provided herein should be appropriate for virtually any evaluation of an educational program funded by virtually any funding agency or foundation, as long as agency-specific requirements and regulations are followed.

To assist those who are not familiar with the federal government's language and the number of "alphabet soup" terms that might be utilized within the *Handbook*, the Appendix I document "KEYS TO ... Understanding 'Title VII-ese'" is provided.

Finally, we hope that this *Handbook* will meet the needs of a diverse audience: those experienced with evaluation and those who are just beginning their first evaluation experience. The constructs presented in this handbook are both new and old -- they are based on management premises, evaluation theory, innovations and practices in the field, and approaches which show both strengths and weaknesses of actions. The arena is expansive, yet includes meaningful real-world applications that have been field-tested by exemplary schools, leaders, and practitioners.

Program directors should consider sharing this *Handbook* with staff members so that they will understand all elements of the program are planned and the importance of record keeping and sharing of ideas and results. When staff have a greater understanding, their ownership and "buy-in" will be increased and the evaluation process will become less threatening to all concerned. Good luck with your evaluation experience.

(table of contents)

II: Thinking About Evaluation

> *People are always evaluating. We do it every* day. We *buy clothing, a car, or refrigerator. We select a movie or subscribe to a magazine. All these decisions require data-based judgements.*
>
> *Payne (1994, p 1)*

Evaluation must be carefully planned from the beginning of the project in order to be useful. The question then may be either "Useful to whom?" or "Useful for what?" In order to answer either question, the purpose of evaluation must be considered. Some authors (e.g., Nevo, 1990) distinguish between what *evaluation is* and what *evaluation's function is*.

Evaluation is a determination of the worth of a thing. Program evaluation, the purpose of this document, consists of the activities undertaken to judge the worth or utility of an educational program. Usually this program is undertaken as a means of improving some aspect of an educational system. For instance, the purpose of bilingual education in the United States is to ensure that students learn English and content-area skills, and perhaps to promote bilingualism. Anderson and Ball (1978) describe six more specific capabilities of program evaluation. These capabilities are not mutually exclusive and need not be important for every evaluation undertaken:

- to contribute to decisions about program development and implementation,
- to contribute to decisions about program continuation, expansion, or "certification,"
- to contribute to decisions about program modification,
- to obtain evidence to rally support for a program,
- to obtain evidence to rally opposition to a program, and

- to contribute to the understanding of basic psychological, social, and other processes.

In general, these evaluation concepts are not new; they are agreed upon. Nevo (1983) reviewed the evaluation literature, finding that they all tend to focus on ten key issues. For a summary of his research, see the document "Conceptualizing Educational Evaluation" in Appendix II.

Various approaches can be used to satisfy the purposes of evaluation. Most typically used within education is the *Objectives-oriented* (sometimes also defined as *goals-oriented*) evaluation. For this approach, the program staff creates broad, generally stated goals. Within each goal, the program staff then must have concrete, behaviorally-defined objectives. The program's success is determined by measuring whether the specific objectives have been met. The major limitation of this type of approach is that the evaluation generally does not measure outcomes that were not anticipated, and stated as objectives, at the beginning of the program. A more major dilemma philosophically is that objectives-oriented evaluation does not attempt to measure the utility or worth of the goals and objectives set for the program. As one humorous example of this, in the 1970s and 1980s, Senator William Proxmire created the "Golden Fleece" award for research projects that were federally funded, but which did not serve a real purpose for the general population. One year, the award was presented for a study on the sex life of the bumblebee. The research may have been good, and did meet its objective of describing and understanding the sex life of the bumblebee, but was this a project worthy of funding with tax dollars?

Other approaches to evaluation can be used effectively, and approaches do not have to be mutually exclusive. For further details on specific types of evaluation that can be used for program evaluation, see Appendix II for the document "Current Frameworks for Program Evaluation."

Regardless of the approach used for the evaluation, there are several functions that the evaluation can serve. Scriven (1967) coined the terms used for two of the functions evaluation most frequently serves: *formative* and *summative*. Formative evaluation is used for the improvement and development of an ongoing program. Based on the outcome(s) of the formative evaluation, the program can be modified to ameliorate problems or bypass potential pitfalls. This does not mean that formative evaluation is done once or twice during a program, it is, as described by Beyer, "*ongoing* in that it occurs repeatedly, at various stages throughout the development process" (1995, p7; original emphasis). Summative evaluation usually serves an accountability function. At the end of the program, a summative evaluation is completed to describe the overall successes of the program and to determine whether the program should be continued. The summative evaluation should include information from the formative evaluations as well as from the final overall product.

The other two functions of evaluation generally are not seen within, or utilized to examine, educational programs. One of these is the administrative function -- to exercise authority. In many organizations, a higher-level administrator will evaluate the performance of subordinates. This is sometimes accomplished in order to demonstrate authority. The fourth type, which Chronbach refers to as the "psychological" or "sociopolitical" function, is utilized to increase awareness of special programs, to motivate desired behavior, or promote public relations. This *Handbook* focuses on the design and implementation of formative and summative evaluations of educational programs.

These are two preliminary steps in designing an appropriate evaluation: defining the function of the evaluation (summative or formative) and determining the approach to be used (objectives-oriented or another, or a combination of approaches). Once these are agreed upon, the general type of assessment to be used can be considered.

Assessment should be considered separately from evaluation, although the two are related. Assessment includes such activities as grading, examining, determining achievement in a particular course or measuring an individual attitude about an activity, group, or job. In general, assessment is the use of various written and oral measures and tests to determine the progress of students toward reaching the program objectives. To be informative, assessment must be done in a systematic manner, including ensuring consistency within measures (from one assessment period to the next with the same instrument) and across measures (similar results achieved with different instruments). Evaluation is the summarization and presentation of these results for the purpose of determining the overall effectiveness of the program, the worth of the program, in order to evaluate the program.

These definitions are provided in Appendix II, the document "Uses for Evaluation Data." With this basic knowledge, we now can turn to the steps in designing an evaluation. After describing the general steps to an evaluation plan, the specific requirements of the Title VII bilingual education evaluation will be addressed.

Evaluation design has one purpose: to provide a framework for planning and conducting the study. Benson and Michael (1990) suggest that there are two major components of evaluation design: (1) defining the criteria by specifying exactly what information is needed to answer substantive questions regarding the effectiveness of the program and (2) selecting the method by determining an optimal strategy or plan through which to obtain descriptive, exploratory, or explanatory information that will permit accurate inferences concerning the relationship between the program implemented and the outcomes observed. The evaluation should be designed so that it meets the needs of the program. Unfortunately, some evaluators are more "method-bound" than "problem-oriented." The former often have one particular type of evaluation design that they use, and they continue to use it whether it is appropriate in a particular situation or not. The problem-oriented evaluator considers the specific problem, and the specific program, then determines the type of evaluation design that is most appropriate.

Evaluation designs generally fall into one of four types: (1) experimental, (2) quasi-experimental, (3) survey, or (4) naturalistic. Each of these is described briefly below, with the description focused on application to bilingual education programs. In addition, resources that describe each of these in detail include Anderson and Ball (1978), Campbell and Stanley (1967), Fitz-Gibbon and Morris (1978b), and Walberg and Haertel (1990); Guba and Lincoln (1981) focus primarily on naturalistic evaluation.

Experimental *and* ***Quasi-experimental*** designs are quite similar. The true experimental design is used to study cause-and-effect relationships; that is, did the bilingual program *cause* students to learn English and increase their academic achievement? This is the most powerful design, but is restricted by two requirements: (1) that students are selected randomly and then assigned randomly to the program being studied rather than to the regular education program, and (2) that the program being studied is carefully controlled with no other students receiving its benefits. An experimental approach is considered one of the strongest methods because it does allow a clear determination of whether the program under consideration caused the students to improve in some way. However, the first condition is especially difficult for bilingual education programs -- it is not possible to randomly assign students since the very existence of the program is based on a demonstrated need of students for the program. In fact, it would be illegal to deny students access to the bilingual program once they have been identified as needing the program.

The quasi-experimental design is somewhat less restrictive. The design is similar to the experimental design except that learners are neither randomly selected from the regular school program nor randomly assigned to the bilingual program. These designs offer greater flexibility and greater potential for generalization to a "real" educational setting. It is still desirable to control as many other elements that may impact the program

as possible.

Both experimental and quasi-experimental designs require some type of pretest (a test taken before the program begins) followed by a posttest (a test taken after the program ends) to determine whether students have increased their knowledge and skills. It often is desirable to have a control group (students who were not in the bilingual program) of some type so that the evaluator can say (1) students in the program increased their knowledge and skills and (2) students in the program increased their knowledge and skills at a greater rate than did students not in the program. How are "control" groups selected? Some funding agencies require a comparison of project students against another group of students. Title VII evaluations, under IASA, require "data comparing children and youth of limited-English proficiency with nonlimited English proficient children and youth with regard to school retention, academic achievement, and gains in English (and, where applicable, native language) proficiency" (IASA Title VII, 7123 [c][1]). Title I does not have such a statement at the present time. Three types of nonproject comparison groups are possible; each is appropriate in different situations.

True control group(s) are students who are randomly selected from the school and randomly assigned to the control group. In the case of Title VII, these students are just like the students in the bilingual program, except that they are receiving the traditional education program (probably English only or a type of English-as-a-Second Language curriculum) rather than the bilingual curriculum. This type of control group is essential for a true experimental design.

Nonproject comparison group(s) are students who are similar to those in the educational program, but are not identical to them; they have not been randomly selected or assigned. For Title VII, these may be students who have similar backgrounds to the bilingual program students, but who are attending another school that does not offer bilingual education; students whose parents did not want them enrolled in a bilingual program; students who speak a language not included in the bilingual program; or students who attend the same school but are English speakers. Nonproject comparison groups usually are used with quasi-experimental designs.

Norm group comparisons are not really "live" students who are in a particular educational program. These students are (1) the norm group from a norm-referenced test or (2) a test score such as the school district average or state average used to represent the norm group. When considering Title VII, these students may be more or less similar to the bilingual program students and generally do not attend the same school as the bilingual program students. Frequently, no students actually are involved in this comparison group: since 50 NCEs (normal curve equivalents, a type of score on standardized norm-referenced tests) always is the national average, this score can be used as the norm group comparison. Again, this type of control group is often seen in a quasi-experimental design. Evaluating educational programs by comparing program students with a norm group is appropriate if the purpose is to show that the program students are becoming more similar to mainstream, predominantly English speaking, students. This type of comparison often is used in evaluation procedures such as the gap reduction technique (see IV: Implementing the Evaluation, pages 73-76).

Survey designs are especially useful when collecting descriptive data; e.g., the characteristics of learners and their families, staff, and administrators; current practices, conditions, or needs; and preliminary information needed to develop goals and objectives. Survey designs follow four steps:

(1) determine the population of interest (for instance, Spanish-speaking students in grades K-5, their families, and their teachers);

(2) develop clear objectives for the survey, develop the questionnaire, and field-test the questionnaire;

(3) if the population is large, identify a sample to be surveyed and administer the survey; and

(4) tabulate the results to provide the descriptive information.

The number of surveys distributed, and the number returned (the "response rate") should be documented. Although surveys are powerful, a limitation on their generalizability and on their worthiness is the response rate -- a low response rate makes interpretation of the results difficult.

Surveys can be highly structured (specific questions with a set group of responses) to unstructured (general questions with the respondent providing whatever responses s/he feels appropriate); surveys can be sent through the mail, completed in-person, or used as an interview. The information gathered is only as good as the questions on the survey instrument. It can be difficult to interpret the results if the questions are open to interpretation or if the possible responses do not allow the respondent a full-range of options. (For instance, consider this question: "Is the program staff sensitive to culture, language, and gender issues?" If the answer is "no," does this mean that they are not sensitive in all three areas, or in one or more of the areas? in which area[s] are they sensitive?) In addition, it will be difficult to design a complete evaluation using only survey methodology. This type of design should be only a part of the total evaluation design.

Naturalistic or pluralistic designs were developed in response to criticisms of the other three design-types: none of them really capture the context of the school and the program. The context, which includes students and their families, teaching staff, school administrators, and various elements of the surrounding community, can interact with the program in unique ways.

Naturalistic techniques are based on ethnographic methodologies developed by anthropologists. They can provide in-depth information about individuals, groups or institutions as they naturally occur. They are regarded as "responsive" because they take into account and value the positions of multiple audiences (Hamilton, 1977). These evaluations tend to be more extensive (not necessarily centered on numerical data), more naturalistic (based on program activity rather than program intent), and more adaptable (not constrained by experimental or preordained designs). In turn, they are likely to be more sensitive to the different values of program participants (Parlett & Hamilton, 1972; Patton, 1975; Stake, 1967). Guba and Lincoln (1981) consider naturalistic evaluation models as highly responsive, offering meaningful and useful approaches to evaluation design.

A major feature of many naturalistic evaluations is the observer who collects, filters, and organizes the information; this person's biases (both for and against the program) can have an impact on the outcome(s) of the evaluation. Naturalistic inquiry differs from surveys and experimental or quasi-experimental designs in that usually a relatively small number of learners are studied in greater depth.

While naturalistic approaches have long been accepted as a method for collecting information for planning an evaluation, for monitoring program implementation, or for giving meaning to statistical data, Lincoln and Guba (1985) suggest that naturalistic information is much more important. They maintain that an entire evaluation can be based on naturalistic methods of information collection. However, few evaluations have been completed and published using naturalistic techniques only. Therefore, we suggest that naturalistic approaches should be part of a complete evaluation design, but not the sole technique used.

Mixed-method designs are described by Payne (1994) as involving both qualitative and quantitative

techniques about equally in one evaluation. He states that mixed-method designs in which 'the evaluation team consists of both qualitative and quantitative evaluators committed to their inquiry paradigm and philosophy is a particularly strong design" (p 127). This method allows for triangulation, defined as "the combination of methodologies in the study of the same phenomenon" (p 125). Four types of triangulation can be described:

(1) using several different evaluators, with different orientations (e.g., qualitative and quantitative);

(2) using several data sources (e.g., standardized tests, alternative assessments, and interviews);

(3) using several data collection methods (e.g., reviewing students' cum-folders and surveying teachers); and

(4) using different theoretical approaches (e.g., using an evaluator familiar with and supportive of two-way bilingual education and another evaluator familiar with and supportive of transitional-type programs).

Using multiple methods enhances the overall evaluation design because the weaknesses of one particular design can be off-set by the strengths of another design. Using triangulation should result in corroborative evidence across sites, methods, and data sources. As Miles and Huberman point out (1984, cited in Payne, 1994), triangulation should "support a finding by showing that independent measures of it agree with or, at least, don't contradict it" (p 127).

Another way to look at the combination of quantitative and qualitative techniques is to recognize the frequently quantitative data can show what is happening while qualitative data can show why it is happening. For instance, the quantitative data may show that the bilingual education program is not working (a statistical result). The qualitative data then may reveal that the bilingual education program has not been implemented as planned, leading to its lack of success.

Most funding agencies have specific requirements for evaluation -- many of which serve an accountability purpose. Survey and naturalistic designs can provide invaluable information about the program, but by themselves will not meet the regulations of many funding agencies. By integrating the best models of evaluation, school programs should have a strong evaluation providing information about their effectiveness and improvements. This combination of qualitative and quantitative methods and data analysis will benefit the program greatly.

Assessment systems are key to a good evaluation. The overall purpose of an assessment system is to initiate and maintain discussion about how the program addresses the needs of all participants. As part of this, the program staff must be prepared to assess their own effectiveness as well as participant needs and outcomes. In general, an assessment system should lead directly to the evaluation by ensuring measurement at three times throughout the program:

A needs assessment will determine the current status of participants' (and potential participants') expertise and knowledge. A needs assessment allows program planners to determine the needs, desires, and goals of the potential participants and/or their parents, teachers, and other stakeholders. The basic questions are, "Where are we now? What do we know about what these students need, what areas are lacking, and what should we address first?"

On-going measures of progress will determine the successful features of the program, the shortcomings of

the program, and whether program implementation and the participants are progressing in the expected manner. Measures of progress allow staff to determine whether the program is working and allow participants to see their own growth. The basic questions are "How much change has there been from the beginning of the program until now? At this rate of change, will we meet our objectives and goals by the end of the program period? What else is 'going on' about which we should be aware?"

Outcome measures will determine whether the objectives of the educational program have been met. These measures make it possible to summarize the progress made by the participants across the entire program. The basic questions are, "How much change did we effect this year? What do participants know now? Do they know what we had planned for them to know?"

An assessment system that includes all three of these key features, and leads directly to the evaluation, will provide useful information for a variety of purposes, in a variety of modes, about a variety of participants. In other words, such a system will include multiple measures that provide information regardless of the participant's culture, gender, or language. Of course, it is assumed that the educational program will include valuable, worthwhile, and frequent opportunities to learn. Without the opportunity to learn meaningful material in a meaningful manner, an assessment system has little value. (As an example of a complete system of assessment, see Holt, 1994.)

Various types of assessments can, and should, be used within an appropriate assessment system. Each must be carefully thought out and be related to the others in some manner. As a first layer of definition, an assessment may be norm referenced, criterion referenced, or may be an alternative assessment that describes current levels of knowledge, attitudes, and proficiencies. Some of the most frequently used are defined in Del Vecchio, et al., 1994.

Interviews *and* ***focus groups*** can provide in-depth information. In a structured interview, responses to a set of prepared questions can be recorded by the interviewer who can ask clarifying questions. Focus groups can include small groups of individuals and a facilitator to discuss a specific topic. Generally, scores are not developed; the data is qualitative in nature. It will be important to identify key individuals to interview (teachers, administrators, students, family members, and others in the community); it also will be important to create good questions to ask.

Surveys usually list a series of questions to be answered orally or in writing by the respondent. The responses can be forced choice, where the answers are provided (e.g., Are you pleased with the expertise of the staff facilitating the training sessions? yes/no), or may be scored on a rating scale (4 to 7 response options such as "very pleased with expertise" to "not at all pleased with expertise"). Scores can be developed by assigning point values to the responses (e.g., Yes=1, No=0) and summing these values. The responses also can be open-ended, where the individual provides an answer (e.g., What pleases you most about the expertise of the staff?). As with interviews, scores generally are not developed for open-ended surveys.

Observation checklists can be used to determine whether particular behavioral, physical, or environmental characteristics are present. Typically, desirable behaviors are described briefly and an observer checks () whether each behavior is observed during a particular period of time (e.g., the first week of the program). Scores can be developed by counting the number of checks. When the same checklist is used periodically throughout the program, it can be used to demonstrate progress by showing more behaviors being observed (checked) across time. In addition, observational rating scales can be developed. To provide useful information, observational rating scales should be tied directly to the objectives and instructional activities of the program and conducted on a regular basis. By linking the descriptors and progression of ratings to

instructional priorities, staff can obtain valuable data for assessing learners' ongoing progress and for improving the instructional program.

Alternative assessments are types of measures that fit a contextualized measurement approach. They can be easily incorporated into the training session routines and learning activities. Their results are indicative of the participant's performance on the skill or subject of interest. Observation measures are an example of an alternative assessment. As used within this document, "alternative assessment" subsumes authentic assessment, performance-based assessment, informal assessment, ecological assessment, curriculum-based measurement, and other similar forms that actively involve the participant.

For many types of alternative assessments, different scoring methods can be used. Three typically used methods are holistic scoring, which provides a general, overall score, primary trait scoring, which defines particular features (or traits) of a performance and then provides separate scores for each trait, and analytic scoring, which assigns a weight based on the importance of each trait (e.g., the use of inclusive language might be weighted more than correct grammar).

Criterion-referenced tests (CRTs) are sometimes considered as a type of alternative assessment. CRTs measure whether specific knowledge has been gained; that knowledge being the criterion against which the participant's current knowledge is measured. Answers can be marked as correct or incorrect for scoring purposes. A score of 80% correct usually is considered as mastery of the knowledge.

Standardized tests can be used to measure participant skills. They are so named because their administration, format, content, language, and scoring procedures are the same for all participants -- these features have been "standardized." Locally developed and commercially available standardized tests have been created for most achievement areas and for some aspects of language proficiency. When considering the definition of "standardized test," it is clear that all high-stakes tests should be standardized, whether they are commercially available tests or locally developed alternative assessments.

When referring to standardized tests, most people think of **norm-referenced tests** (NRTs). NRTs typically are used to sort people into groups based on their assumed skills in a particular area. They are useful when selecting participants for a particular program because they are designed to differentiate among test-takers. In addition, NRTs can provide general information that will help to match classrooms for overall achievement levels before assigning them to a particular program.

Portfolio does not refer to a specific type of assessment, but is an approach to organizing the information about an individual or a class/program. Portfolios can serve as a repository for "best" works or for all work on a particular project, from first notes to final draft. The portfolio can contain projects, assignments, various alternative assessments, and/or results from NRTs. The portfolio also can be used as a record of achievement that can be used to demonstrate expertise in a particular area.

Meaningful assessment is essential. To ensure that an assessment is meaningful, two factors must be considered: reliability and validity. While psychometricians still argue about the relative importance of each of these concepts and what constitutes "good" reliability and validity, some general explanatory statements can help to clarify these test qualities.

Reliability is the stability or consistency of the assessment. For instance, two assessments of a participant, performed at the same time, should show similar results; two reviews of a teacher's qualifications should result in similar conclusions. An instrument must be reliable if it is to be used to make decisions about how

well a participant is performing or how well a staff development program is succeeding. As a general rule, the more items on an assessment, the greater the reliability. A test with 50 items usually will be more reliable than an assessment with 10 items; however, an assessment with 300 items may fatigue the test-takers and be very unreliable. Most psychometricians agree that at least 10 items for each area tested are needed to have a reliable instrument. (For instance, on a math test covering addition and subtraction, there should be a minimum of 10 items in each of these areas. The more areas covered on a test, the longer the test will be.)

For a brief but in-depth discussion of reliability, including statistical formulae for calculating reliability, see Thorndike (1990).

Inter-rater reliability is a specific type of reliability that is important when assessing students with alternative assessments. Inter-rater reliability indicates the agreement between two or more people who use the same assessment to determine the skills of the same student. This is important in order to ensure that the scoring criteria are understood the same way by all scorers, and that the scoring criteria are being utilized in the same way by all scorers. To determine the inter-rater reliability, determine the number of times that the scoring of two persons matches; an 80% match is desirable and should not be difficult with a well-designed instrument, with well described scoring criteria. When utilizing alternative assessments, teachers should be trained using video-taped vignettes or play-acted situations. Training and practice scoring should continue until at least 80% match among raters is reached. Periodic retraining should be utilized to ensure that the match continues to be this high.

Validity is more difficult to describe, in part because psychometricians are changing their own views of validity. The newer view of validity is that it asks whether the interpretation, uses, and actions based on assessment results are appropriate (c.f., Messick, 1988). The Joint Committee of the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education adds that "validity ... refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" or assessment results (Joint Committee, 1985). Durán (1985) suggests that it is particularly important to consider the communicative competence of learners when creating a valid test. For a traditional view of validity, see Zeller (1990). For an in-depth discussion of the newer picture of validity, see Messick (1985).

If different assessments (or the same assessment scored by different individuals) provide similar information about the skills of a student, and if that information seems trustworthy, important, and can be generalized to other situations, then the instrument probably is valid and reliable. An instrument is reliable and valid only when it is used in the manner for which it was developed and for the purpose for which it was designed (including, of course, the participants for whom it was designed).

Other factors must be considered when selecting an assessment. Some of the more important are listed below.

- Time--how long does the assessment take to administer and score? is the time appropriate for this program?
- Cost--how much does the assessment cost to copy, administer, and score? is this cost acceptable?
- Personnel--who will administer the assessment? is special training needed to ensure that the instrument is administered in the correct manner?
- Scores--are the scores appropriate? do they provide useful information?
- Evaluation--can the assessments be aggregated to form a viable evaluation of the participant and/or of

the program itself?

Features of the school program that should be assessed include the context, implementation, and student outcomes. While we tend to focus on student outcomes (i.e., language proficiency and content achievement for bilingual programs), other features of the school are equally important. As described by Del Vecchio, et al.

- program context indicators describe the ethos, management, and resources that permeate and influence the attitudes of school staff, students, and parents in culturally and linguistically diverse communities;
- school implementation indicators target features in bilingual education schools including curriculum and instruction, staff development, the responsibilities of administrators, and the role of parents; and
- student outcome indicators identify the skills and strategies required of limited English proficient students to succeed ... and to attain the performance standards outlined in Goals 2000 (1994, p 1).

Thus a complete assessment system will take time and energy to design. It must assess the impact of the program on various aspects of student life and it must assess the impact of various school components on the program. All of this must be done in a cost-efficient and timely manner.

Quantitative analyses are required for experimental and quasi-experimental designs; they might be used with some other design types, but generally this is not the case. The statistics involved can be very sophisticated, or they can be relatively simple. The key is to use the statistics that are (1) appropriate for the study, (2) comfortable for program staff, and that (3) can be explained in simple language. A basic evaluation analysis can be completed by program staff, more complicated procedures may require an evaluation specialist -- the results of either should be succinct, clear statements about the overall outcomes of the project. A brief overview of some commonly used statistics follows. More details are provided in Appendix II.

Statistics can be categorized into two types: (1) descriptive and (2) inferential. Descriptive statistics are those used to describe the population -- numbers, percentages, and averages. Inferential statistics are used when the evaluator wants to make a generalized statement about the importance of differences or similarities among groups. In statistics, "importance" has specific meanings. In general, something is considered important if it probably did not happen by chance; this is referred to as "significance." If the students in the two-way bilingual class had higher year-end test scores than students in the ESL class, and if those differences could happen by chance (because students just happened to guess the right answers or the test just happened to measure their particular knowledge and skills) no more than 5% of the time, the results are said to be statistically significant. While not every result that is statistically significant is automatically "important," statistical significance is one measure of importance.

Descriptive statistics such as simple tabulations of data are required in most evaluations: how many students are in the program? what languages do they speak? what grades are the students in? what courses are they taking? These questions can be answered by constructing a questionnaire that staff fills out from their classrooms, or by reviewing school records; they can be answered by descriptive statistics. When reporting such information, the data should be broken into the smallest pieces possible. For instance, note the differences in the two examples in Table 1. Each contains the same general information, but one is much more useful than the other.

Another type of descriptive statistics involves calculating average scores and information about how much

these vary from high to low -- the standard deviation (SD). Average scores should be presented with their standard deviation, and a listing of the highest and lowest scores possible as well as the highest and lowest scores actually received. As an example,

> The 50 students completed a program-designed assessment of reading skills. The possible scores ranged from 0 to 80, with the students actually scoring from 45 to 75 (average score: 58.6, standard deviation: 5.4).

The standard deviation is a measure of variance, how much the students' scores differed around the average score. In this example, the average score is 58.6. With a standard deviation of 5.4, the reader knows that about two-thirds of all the scores can be expected to be within 5.4 points of 58.6; about two-thirds of the students scored between 53.2 and 64.0.

Table 1.

Example student background information data

Students by grade	Number	%
K	20	25.6%
1	18	23.1%
2	15	19.2%
3	25	32.1%
Total	78	
Spanish-speakers	60	76,9%
Vietnamese-speakers	15	19.2%
Other languages	3	3.8%
Total	78	

Students by grade/ language		#	%
K	Spanish-speakers	15	19.2
	Vietnamese-speakers	5	6.4
1	Spanish-speakers	16	20.5
	Vietnamese-speakers	1	1.3
	Other languages	1	1.3
2	Spanish-speakers	14	17.9
	Vietnamese-speakers	1	1.3
	Spanish-speakers	15	19.2

3	Vietnamese-speakers	8	10.3
	Other languages	2	2.6
	Total	78	

Some of the scores that can be calculated for standardized norm-referenced tests can be used descriptively. Stanines, percentiles, and grade-equivalents all can be used to describe the general skill level of a group of students, but should not be used in calculations (e.g., do not calculate averages). These types of scores also should not be used in inferential statistics.

Inferential statistics usually are based on analyzing average scores and standard deviations. This allows conclusions to be made so that the evaluator can make inferences about the group of students. Inferential statistics usually require a minimum of 10 students in each group being evaluated (e.g., 10 female Spanish-speakers in the second grade two-way bilingual class, 10 female Spanish-speakers in the third grade two-way bilingual class).

The most basic of inferential statistics is the **t-test**. The t-test is used to compare two average scores: the average scores of the boys vs the girls, the Spanish-speakers vs the Vietnamese-speakers, the third grade students vs the fourth grade students. Only two average scores can be compared at one time, although it is possible to calculate multiple t-tests during an evaluation. As a general rule, the larger the t-test value (either positive or negative number), the more important the difference between the two groups' average scores.

t-tests often are used in both true experimental and quasi-experimental designs. They can be used to test the difference between the pretest and the posttest (did students score statistically better at the end of the program than they did at the beginning?) and between the students in the program and the control or nonproject comparison group students (did the students in the program score statistically higher than the students not enrolled in the program?).

Other types of statistics can be used for many evaluation designs. Most of these are outgrowths of the t-test. For instance, the t-test only allows two average scores to be analyzed at once. Other types of analyses (e.g., the analysis of variance, ANOVA) allow three or more average scores to be analyzed at one time. These analyses can become quite sophisticated. However, if there are several average scores that need to be analyzed, these more sophisticated analyses are more appropriate than several t-tests. For information on doing statistics, see Hays (1988), Huitema (1980), Kerlinger (1986), Kirk (1982), Pedhazur and Schmelkin (1991), or Popham and Sirotnik (1992).

Statistical packages are available to assist in the quantitative analysis of data. Virtually any statistical package, and most data base packages, will be able to provide descriptive statistics -- simple frequencies, average scores, and so on. Most of these will be able to do basic inferential statistics, such as t-tests, as well. However, before purchasing one of these statistical packages, be sure that it can handle to number of students in the program. Many of these "smaller" statistical packages limit the number of "subjects" (students) and/or the number of "variables" (other interesting groupings such as nonproject comparison group or project group, gender, age, grade level). Especially for a comprehensive schoolwide program, this could be a problem. Also, of course, be sure the program is available for the type of computer that will be used.

For programs that desire more sophisticated analyses, there are fewer statistical packages available. While

many statistical packages claim to be able to do these statistics, fewer actually can do them in an appropriate manner. One of the main problems deals with numbers of students in each grouping (e.g., number of female French-speaking 3rd graders who are fluent-English proficient). It is unlikely that each small grouping will have the same number of students, but this is a requirement of many statistical programs. It will be essential to find a statistical package that deals with "unequal ns" (i.e., unequal numbers of subjects) in an appropriate manner -- only experience and a well written technical manual will provide this information.

Qualitative analyses are essential for naturalistic designs and for mixed-method designs. Qualitative analyses are inductive. Evaluators generally look for information that can be identified across a variety of data sources and methods, and a great deal of rich data. While most qualitative data are in narrative form, some quantitative data might also be included; e.g., frequency counts and averages, generally any of the descriptive statistics described earlier. The expertise of an evaluator may be needed to interpret data, determine the significance of the results, and to draw conclusions.

The evaluator generally will begin by identifying categories or themes in the data, then attempt to establish relationships among the categories. Finally, the evaluator will look for more evidence to support the categories and relationships by returning to the field setting (the school or bilingual classroom) to collect additional data. Payne (1994) suggests that qualitative analyses generally fall into one of four types. Each is described briefly below.

Phenomenological analyses are most often used with interview data, questionnaires, and open-ended surveys. The purpose is to understand the program in its own right, from the view of those participating, rather than from the perspective of the evaluator. The evaluator must suspend his/her own beliefs about the program and allow the beliefs of those involved to emerge from the data as categories that then can be addressed within the evaluation.

Content analysis is a well known method for analyzing documents obtained about the program being evaluated. Documents produced by the program staff can be a good source of information about program implementation. As described by Payne, "evaluator-generated rules for categorization, demonstration of representativeness of categories, relations among categories, and definitions of categories from participant perspectives are important outcomes of content analysis" (1994, p 137).

Analytic induction is utilized when evaluators begin with a theory to test about a program in a particular setting (e.g., the two-way bilingual education program will result in more in-depth learning on students' part than pull-out ESL classes). Rather than beginning with observations and interviews in order to develop a theory, particular types of data from selected individuals is collected and analyzed based on the theory the evaluator already holds. As data is collected without new information being found, the evaluator stops collecting data and presents the evidence already found.

Constant comparative analysis is an approach to analysis that results in grounded theory. Rather than collecting data, then analyzing it, constant comparative analysis suggests that data be analyzed throughout the data collection process. As a theory begins to emerge from the data collected, that theory will indicate what other data should be collected. If the theory holds, the "new" data will continue to provide information to refine the theory.

Some researchers, particularly quantitative researchers, feel that qualitative studies cannot provide the solid, objective, information that numbers provide. However, a well-designed, multi-site qualitative evaluation can enhance the generalizability of the findings. Multi-site evaluations of the same type of program in dissimilar

contexts (e.g., the studies by Kathryn Lindholm of two-way bilingual programs throughout California) provide a great deal of generalizable information. As with Payne (1994), however, we highly recommend an evaluation plan that includes both qualitative and quantitative methods of data collection and analysis.

Packages for computers now are available. Usually these programs will assist in developing categories for qualitative data and will provide counts of the number of categories and the number and type of data that fit into each category. There are not many of these, and generally the same package is not available for both DOS-based and MAC machines. It will be important to work with the evaluator to find a package that fits the specific needs of the program.

Evaluators often are hired to ensure that the evaluation is as valid and reliable as possible. While it is tempting to "turn over" the responsibility of the evaluation to the evaluator, this is one temptation that should be resisted!

The role of evaluators should be to assist the program staff in ensuring the best possible evaluation -- including creating and/or modifying assessment instruments -- and sharing their expertise about evaluation design and statistical analyses. Evaluators may specialize in a particular type of evaluation (e.g., Title I, Title VII), or they may be generalists. The program director should not assume that the evaluator is aware of the specific purpose of a bilingual education program, of a migrant education program, or that they know the various statutory regulations pertaining to the evaluation of specific types of programs. And, since the regulations are modified fairly frequently, even evaluators who are knowledgeable about a specific funding agency's evaluation regulations (e.g.,Title VII) should be given a copy of the regulations under which a particular program falls.

It is the responsibility of the program director to hire an evaluator early in the life of the program. In order to do this, the hiring practices and rules of the local district should be explored. Some districts require an external evaluator (one who is not employed by the school or school district), others require an internal evaluator (one who is employed by the school or school district). There is no requirement within Title VII, or the other IASA titles, that an evaluator be hired. If the program director, or the person who originally wrote the application for funding, can identify on-staff expertise in evaluation, no one else need be hired. However, it is unlikely that this will be the case; rarely are school staff experts in evaluation. In addition, there are some compelling reasons to hire someone specifically to evaluate the program. Probably the best approach is to form a team with both a professional evaluator (internal or external) and staff members. This will allow a group of people who are knowledgeable about the program to share their information, providing the best of both internal and external evaluation techniques, and affording maximum "buy-in" of staff.

It is not uncommon for a professional evaluator to assist in writing the grant requesting monies for a project. Sometimes, there is no charge for the writing assistance, with the understanding that the same person will be hired for the evaluation when/if the grant is funded. The best way to identify a knowledgeable, competent evaluator is by contacting other program directors, asking them for recommendations (and perhaps who to avoid). In addition, newspaper advertisements can be helpful. Any advertisements should be specific about the qualifications desired in the evaluator; references should be requested and should be contacted. If possible, example evaluation reports should be requested -- this will provide examples of the style of writing, type of report, and general evaluation skill of the individual.

When negotiating the contract, specific tasks should be discussed. Many tasks can be accomplished by the program staff, others really should be completed by the evaluator. For instance, there is no reason for the evaluator to take time (and money) to write the background of the project for the report; the program

director and staff know the background and can provide more details, more quickly, than the evaluator. On the other hand, the evaluator probably will need to write the interpretation of the statistical results since that should be her/his area of expertise. The key is to

(1) identify the tasks of the evaluation;

(2) review the capabilities of the staff and the evaluator, and

(3) consider who will be most capable to complete each task.

It often is possible to "trade" tasks between the evaluator and the staff -- this can provide more of the evaluator's skills at less cost.

Finally, the contract for the evaluator's work should be as specific as possible. The number of meetings the evaluator needs to attend; the number, type, and due date for assessment instruments to be selected/created; the number and type of reports to be delivered; and the types and dates of data collection are some of the details that should be included. In addition, a provision that final payment will not be made until the report is edited and approved by the program director is important. Other information on working with an evaluator is included in Appendix II, in the document "Finding an Evaluator."

The roles of the various participants in the implementation of the program (staff, director, and evaluator) are key to a successful evaluation. These roles are described in four documents in Appendix II: "Role of Project Director in Evaluation," "Role of Staff in Evaluation," and "Role of the Evaluator in a Title VII Project." "Working with Your Evaluator on the Final Report" describes the activities and tasks in which staff, director, and evaluator can participate to ensure a complete evaluation report.

Two facets of hiring an evaluator should be emphasized one more time: (1) there is no requirement within IASA that an evaluator (either internal or external, individual or team) be hired -- the evaluator could be someone already on-staff with the program who takes on the evaluation as part of his/her regular program duties and (2) the team approach should be considered very seriously -- the advantage of working with several individuals who are part of the program staff and who are external to the project cannot be underestimated.

Summarizing, this section has described:

- the purpose and function of evaluation within an educational setting,
- four evaluation designs -- experimental, quasi-experimental, survey, and naturalistic -- and defined three types of control groups,
- basic assessment definitions and procedures,
- some differences between qualitative and quantitative analyses, and
- how to hire and utilize an evaluator in an effective manner.

The text has provided general information, definitions, and descriptions. In addition, an appendix to this section includes several pages that define more fully and/or that can be used as handouts for staff training purposes.

Overall, it always should be remembered that evaluation data is of little value unless the project is able to use the information to improve its program. Developing an action plan for using the evaluation results is

critical for ensuring effective implementation of a an educational program. The key value in integrating evaluation with program improvement efforts is that relevant assessment data can be used as a guide for planning the effective program.

(table of contents)

III: Planning an Evaluation

> *A design is a plan which dictates when and from whom measurements will be gathered during the course of an evaluation. The first and obvious reason for using a design is to ensure a well organized evaluation study: all the right people will take part in the evaluation at the right times.*
>
> *Fitz-Gibbon & Morris (1978, p 10)*

Evaluations almost always involve multiple and diverse audiences: those who will use the evaluation to make decisions, individual administrators or legislators, instructional staffs, or the large group of consumers who purchase the goods and services being assessed. Other typical audiences would be the individuals and groups whose work is being studied, those who will be affected by the results, community organizations, and possibly the general public. In order to ensure that the evaluation has utility, all the details must be worked out early in the program -- the earlier the better. All these details are what we refer to here as planning the evaluation.

This section will provide fairly detailed information that builds upon the general overview and definitions from Thinking About Evaluation. In particular, this section will provide a number of handouts and forms that can be used to assist a program as it considers its evaluation. To some extent, the information is provided in a chronological order. That is, the portion(s) of the evaluation that should be considered earlier in the evaluative process are presented earlier in this section. Activities described in this section all are part of planning the evaluation; these activities should be completed early in the life of the program being evaluated. Activities that should be carried out at various times during the life of the program will be described in the next section, Implementing the Evaluation. Four major areas are discussed in this section:

- Managing the evaluation and creating timelines;
- Ensuring that goals and objectives are appropriate, well-defined, and feasible for the project;
- Assessing context, implementation, and student performance; and
- Scoring the assessments.

While other aspects of the program are important, these are the issues of primary importance for planning the evaluation.

Managing the evaluation is the responsibility of the program director. The program director should ensure that all staff are trained appropriately, determine whether a formal evaluator is needed, and assign staff members to various tasks. S/he also should monitor the activities of all staff members to ensure that the activities of the program and of the evaluation are implemented as closely as possible in the manner originally intended. This may require numerous staff meetings and training periods, especially at the

beginning of the program (or if possible, in a planning phase that occurs before program implementation begins). Along with this, a Management Time Schedule should be developed. An example Management Time Schedule is included in Appendix III; part of it is duplicated here to demonstrate how it can be completed.

Figure 1.

Example Management Time Schedule

Management Tasks Months	Sept	Oct	Nov	Dec	Jan	Feb
Planning						
Determine need for evaluator; hire if necessary.	xxx**X**					
Meet with evaluator/staff-discuss evaluation plan	x	xxxx	xxxx	xxxx	xxxx	xxxx
Determine feasibility of evaluation plan	x	xxx**X**	xxxx	xxx**X**	xxxx	xxx**X**
Review objectives & assessment instruments	x	xxxx			xxxx	

x indicates approximately one week; **X** indicates the completion of a task or product.

Within this Time Schedule, several pieces of information are evident. First, each month of the program has been indicated; this program begins in September. Various tasks needed for the Planning phase of the program are listed. Each "x" represents one week of work, assuming four weeks in each month. The larger "**X**" indicates a product or the completion of a task. For instance, the evaluator will be hired by the end of September; meeting with the evaluator and/or the program staff to review the evaluation plan is an on-going activity. Reviewing the objectives and assessment instruments is an activity that occurs at various times throughout the project.

The actual tasks listed on a *Management Time Schedule* may differ for various projects. In particular, projects that have a preservice time in which to prepare for the program (hire and train staff; identify, purchase, and become familiar with a new curriculum or texts; select or create assessment instruments) will have a very different set of tasks for the first year, as opposed to the tasks for the actual years this program provides services to students.

Timelines can be the key to implementing a program effectively. A timeline for the evaluation was suggested in Figure 1. Other types of timelines, and other details of the program might be included as well. Also, while the prime purpose of a timeline is to keep the entire project "on-task," centered, and on time with each task, it also may have other purposes. For instance, the timeline can help identify tasks for the evaluator and can be used in contract negotiations with the evaluator. Then again, it can be used to assign responsibility for certain tasks to various staff members. Timelines can be included with other tasks as well; for instance, the document "Planning Goals and Objectives" (located in Appendix IV) allows goals, objectives, activities, assessment measures, responsible individual(s), and the timeline to be indicated on one form. This will ensure that the objectives and activities support the goals, that assessments measure the objective, that someone is "in charge" of each goal, and that the timeline is well-known.

It frequently is helpful to work through the timeline with the evaluator, program staff, and school administrators. In this way, everyone understands what the timeline is, who is responsible for particular

tasks, and why the timeline must be kept as closely as possible. In addition, the data management portion of the timeline must be carefully considered. Many consider data collection to be the center of the entire evaluation plan. For some ideas about data collection, see "KEYS TO ... Planning Data Management" in Appendix III.

The handouts for this section include several timelines. The "Management Plan" has been included with the permission of the evaluator and project personnel who originally developed it; "Example Title VII Management Plan" has been modified from one actually used within a Title VII project; and "Implementation Checklist for Title I Schoolwide Programs" was specifically developed to assist in the implementation and evaluation of a Title I schoolwide program.

Communication among evaluator(s), program director, program staff, school staff, and school/district administrators is essential. For a program to be truly successful, there must be an understanding of the purposes of the program and the accomplishments of the program. For this to happen, the entire staff of the program should read portions of the grant application (e.g., the sections on the purpose of the program, the goals and objectives of the program, and perhaps the evaluation); in addition, a synopsis or executive summary of the grant application should be available for parents, other school staff, and administrators. If the grant application has not been read, how can these individuals understand its purposes and know what is expected of them?

Regular communication among program staff, director, and evaluator(s) should be planned and listed in the timeline for the program. In addition, regular communication between program and school administrators should be planned. Whenever there are major achievements, these should be announced at schoolwide staff meetings, at parent-teacher meetings, and to the media. As businesses have long known, it pays to advertise.

Goals and objectives must be written appropriately to ensure that they are evaluable and feasible. While this should have been done when the project was first developed and funding was applied for, it is not infrequent for good intentions to lead to goals that are too specific, objectives that really cannot be measured or that really are activities, or activities that are poorly described. As stated by Rossi and Freeman (1982), "goal-setting must lead to the operationalization of the desired outcome -- a statement that specifies the condition to be dealt with and establishes a criterion of success. ... Unless the goals are operationalized into specific objectives, it is unlikely that a plan can be implemented to meet them" (p 56).

The number of goals and objectives that a program can attempt to accomplish is limited. A program that attempts to satisfy too many needs will be unsuccessful in many of them -- while frustrating and over-working the staff and students. It will be important to determine the number of goals that are feasible for a project, and then to limit the number of objectives to those that most closely relate to the goals; i.e., the objectives that relate most closely to student needs. Project staff can prioritize or select goals and objectives based on need, feasibility, timeliness, random selection, or another method.

Goals should be broadly declared statements about where the program is headed; what the overall purpose of the program is. Some authors suggest that goals can describe either ends or means (e.g., Morris & Fitz-Gibbon, 1978b). In this vernacular, ends-goals are those that describe an outcome, a measurable end-product for the program; means-goals define the process by which the ends will be met, the means for accomplishing the ends. More frequently, goals refer only to "an intended and prespecified outcome of a planned programme" (Eraut, 1990, p 171); i.e., goals should be stated as end-goals.

The broadly-stated goals should meet four conditions:

(1) their meaning should be clear to the people involved:
(2) they should be agreed upon by program planners and funding agencies;
(3) they should be clearly identifiable as dealing with an end product; and
(4) they should be realistic in terms of time and money available.

As an example,

Students will become proficient in English and Spanish

or

Students will understand the cultures of others.

Goals are written after a needs assessment documents the necessity of these particular goals. If the needs assessment indicates that all students are proficient in English, then a goal such as that above would not be appropriate. Likewise, if the program has no intention of working on proficiency, and there are no objectives further delineating the goal, then such a goal would not be appropriate.

For further details on writing goals, see the documents "Specifying Goals," "Determining Appropriate Goals," and "Methods for Prioritizing Goals" located in Appendix III.

Objectives are more specific statements about the expectations for the program. They describe the outcomes, behaviors, or performances that the program's target group should demonstrate at its conclusion to confirm that the target group learned something. More concisely, an objective is a statement of certain behaviors that, if exhibited by students, indicate that they have some skill, attitude, or knowledge (Morris & Fitz-Gibbon, 1978b). Objectives must be measurable and specific.

As suggested by Tyler (1950), Mager (1962), and others, objectives should identify

(1) the **audience**, who the learner is, the target group;
(2) the **behavior**, what the target performance is;
(3) the **conditions** under which the behavior will be performed; and
(4) the **degree**, the criterion of success.

Following the ABCDs of objective writing will ensure the evaluability and the clarity of the objective. For instance, the objective

Students will learn to read in English

may be admirable, but there is no indication of the time frame for learning to read, how "learning to read" will be measured, or how well students must read before the objective is considered a success. To ensure that the objective is measurable, it should be written

By the end of the project year, students will read and understand grade-appropriate materials as measured by responding with 80% accuracy to the project-developed Reading Assessment Scale.

In this statement, the audience is the "students," the behavior is "reading and understanding grad-appropriate

materials," the conditions are "by the end of the year," and the degree is "80% accuracy on the project-developed" instrument. (Whenever a specific level of accuracy is included, it should be defended. For instance, a note might add that "the state has mandated an 80% accuracy level to indicate mastery." Or, an author who suggests such a level of accuracy might be cited.)

As described by Rossi and Freeman (1982), "four techniques are particularly helpful for writing useful objectives: (1) using strong verbs, (2) stating only one purpose or aim, (3) specifying a single end-product or result, and (4) specifying the expected time for achievement" (p 59). Table 2 presents stronger and weaker verbs. Stronger verbs are "active" while weaker verbs are "vague" and not as easy to measure. When a weaker verb is used, a means for measuring whether the objective was met should be included. For more details on writing objectives, see Appendix III for the document "Reviewing Objectives."

Table 2.

Strong and weak verbs for objectives

Strong Verbs	Weak Verbs
Find Increase Meet Sign Write	Encourage Enhance Promote Understand

Frequently asked questions about creating objectives include "How much should be expected of students? What constitutes a reasonable student achievement level?" Unfortunately, there are no straight-forward answers to these questions. First, if norm-referenced standardized tests are used, be sure to use some type of standard score (e.g., NCEs or scaled scores) when writing objectives (see the portion of this section, "Scores"). The size of the expected gain on this type of test that can be expected will vary depending on several factors, including

(1) whether fall-spring or an annual testing cycle is used,
(2) the grade level of students,
(3) the subject matter being served and tested,
(4) the nature of the program,
(5) student attendance, and
(6) students' test-taking skills.

Given that these elements are controlled for and considered, a change of zero NCEs (i.e., no change) on a standardized NRT indicates that the student has maintained his/her standing in relation to the norm group. That is, the students learned what would be expected for them to learn during the academic year. An NCE gain might be attributed to the impact made by additional instructional assistance offered through the program.

In general, consider the following general interpretations of test results for limited English proficient students and with other students at risk of educational failure.

- A drop in NCEs often reflects the expected patterns. These students are behind their grade-peers, and

continue to fall further behind.
- No change in test scores indicates that the students have made progress at the same rate as their nonlimited English proficient (or not at risk) peers. They are maintaining their level of achievement.
- A gain in scores shows considerable progress. The students are catching up to their grade-level peers.

The first example (a drop In scores) could indicate an ineffective program or other "negative" variables (e.g., an outbreak of chicken pox at a crucial time in the curriculum). The third example, and possible the second, indicate greater than expected achievement. This could be to the program, or could be due to other variables (e.g., staff who are not part of the program who are fluent in the students' home language). Naturalistic or qualitative data can be used to aid in interpreting and reporting such scores.

Activities are another element in communicating the intent of the program. The purpose of the activities is to describe in detail any prerequisites or actions necessary to ensure the achievement of the objectives. "Prerequisites" refer to any conditions and/or criterion in which the objective is to be achieved.

As an example, the objective Students will read a complete novel by the end of the year might be followed by the following activities:

- Define the term "novel."
- Identify the different types of novels.
- Select from one type of novel.
- Read excerpts from at least five novels.
- Select one novel from those reviewed.
- Read the selected novel and complete exercise sheet.

The document "Creating Activities," located in Appendix III, provides further definitions and example activity statements.

Modifying goals and objectives after the educational project has been funded is possible. While the overall focus and purpose of the program cannot be changed, goals and objectives may be modified to make them more in-line with the current needs of students, to recognize that what had been considered as an objective really should be an activity, or to ensure that the results are quantifiable.

Modifying goals or objectives should be attempted only with the permission of an officer representing the funding agency. Each agency will have rules for such modifications. In general, we suggest contacting the agency by telephone to discuss the reasons modifications are needed, following-up the telephone conversation with a letter requesting permission to make the modifications, and documenting the process in any reports that are written (especially the next report and the final report, if required). This is further explained in "Modifying Objectives," located in Appendix III.

Goals and objectives are required by most funding agencies. They can facilitate the work of the evaluator and program staff in "proving" that the educational program was effective. Indeed, the US Department of Education, within the Education Department General Administrative Regulations (EDGAR) -- the basis for most projects funded through Title I, Title IV, Title VII, and others -- specifically states that "a grantee shall evaluate at least annually -- (a) the grantee's progress in achieving the objectives in its approved application" (EDGAR 34 CFR 75.590). However, it should be noted that (1) evaluating the objectives and goals does not evaluate the quality of the objectives and goals and (2) there are several types of evaluation that do not require objectives (see, for instance, goal-free evaluation, naturalistic evaluation -- Guba &

Lincoln, 1981). While these alternative, more naturalistic, forms of evaluation can be powerful, we suggest that they be utilized in conjunction with the goals and objectives approach. This will ensure that the requirements of the funding agency are met as well as the desires of the local school to know quickly and simply whether the program "works" -- looking at other important aspects of the program can be added to the objectives-driven evaluation.

Other evaluable factors for an educational program include implementation and context indicators. Both implementation elements (i.e., how the program actually was put in-place) and context elements (i.e., what else is going on in the school and community) can have a major impact on the educational program. While goals and objectives usually are not written in these areas, they will be important to measure and to evaluate. This issue is addressed further in the following sections.

Assessment is an essential element to a useful evaluation. Defined in the previous section were the types of assessments and the technical qualities of assessments. Here we describe how to select, modify, and/or develop an appropriate instrument for your educational program. First, however, consider further the purposes of assessment.

As Roeber (1995) points out, the current effort in assessment is primarily threefold: (1) national, (2) state, and (3) local. Unfortunately, there frequently is little coordination among these assessment "levels," with the presumed hope that they will somehow work together -- the result is "a crazy-quilt of programs and purposes ... [that may result in] too much testing of students and an angry backlash of sentiment from teachers and others at the local level against all of the assessment efforts" (Roeber, 1995, p 1). Before a discussion of the design of a comprehensive, coordinated assessment system, consider the real purposes of assessment. In general, assessment can be used to monitor, inform, improve student performance, allocate resources, select or place students, certify competence, and to evaluate programs. For a further definition of each of these, see Appendix III for the document "Purposes for Assessment." Not all assessments will fulfill all of these purposes. As Roeber points out, "It is virtually impossible to meet these different needs and purposes with a single instrument and to do so in an efficient and effective manner" (1995, p 7). He identifies the ideal assessment system as one which has identified

- the audiences for assessment information,
- the types of information needed by each audience,
- the type of assessment instrument that best meets the assessment need,
- the impact of the use of the instrument on the educational system, and
- the levels for use (national, state, local, student) of the assessment instrument.

While it is not within the purview of this Handbook to describe and define an ideal assessment system in detail, it is appropriate to discuss how to select or develop instruments that might fit into such an assessment system. A particularly complete source for information on various aspects of testing is Robert Linn's Educational Measurement (1989), an edited series of articles by well-known authorities in several fields.

"What to assess?" is a question frequently asked. While we usually focus on the assessment of student outcomes (i.e., language proficiencies and content area achievement), there are features of the school program that also are important to assess. Del Vecchio, et al. (1994) suggest assessing and evaluating school context and program implementation as well as student performance outcomes.

Title VII evaluation guidelines define program context indicators as those that

> describe the relationship of the activities funded under the grant to the overall school program and other Federal, State, or local programs serving children and youth of limited English proficiency. (IASA 7123[c][3])

Title I currently does not mention school or program context. Del Vecchio et al. (1994) suggest that key elements of context are the overall climate of the school, its management, and the equitable use of its resources. The methods for assessing whether these three elements of the school are truly inclusive, flexible, and democratic, and whether they meet the needs of all students and their families include surveys and reviewing the school's existing documents and records.

> Implementation indicators are defined within Title VII as
>
> including data on appropriateness of curriculum in relationship to grade and course requirements, appropriateness of program management, appropriateness of the program's staff professional development, and appropriateness of the language of instruction. (IASA 7123[c][2])

An essential component for a bilingual education program is the effective implementation of an appropriate and sensitive curriculum. In addition, staff must have appropriate knowledge and experience, administrators must understand the purpose of and support all academic programs within the school, and the entire family should be involved in the student's education. These essential elements can be assessed through the use of interviews, surveys, rating scales, self-assessments, and checklists. (For suggestions on the development of such instruments, see Del Vecchio, et al., 1994.) Again, Title I does not currently have specific guidelines for program implementation.

The third portion of educational programs that Title VII programs must evaluate is student outcomes. As defined within Title VII, the evaluation should include

> how students are achieving the State student performance standards, if any, including data comparing children and youth of limited-English proficiency with nonlimited English proficient children and youth with regard to school retention, academic achievement, and gains in English (and, where applicable, native language) proficiency. (IASA 7123[c][1])

These guidelines suggest not only what topics should be evaluated, but also state that a nonproject comparison group of English proficient students must be utilized. Many of the types of assessment discussed in the previous chapter (i.e., alternative assessments, observations, NRTs, and CRTs) can be used for these purposes.

Title I has several regulations pertaining to the assessment of student progress. They are quite complex and are related to State content and performance standards, yearly progress of students, the development of appropriate assessments, and so on. Some of the relevant sections of IASA include 1111(b)(1-7), 1112(b)(1), 1116(a), and others. To ensure that the Title I regulations are met, a careful reading of the entire statute and EDGAR is suggested strongly.

Finally, EDGAR further states that the evaluation must include progress toward achieving the objectives, the effectiveness of the program, and the effect of the program on those being served, including a breakout of data by racial/ethnic group, gender, handicapped status, and elderly (EDGAR, 34 CFR 75.590(a-c).

Selecting instruments that are appropriate to the needs of the program, including the relevant funding agency regulations, is extremely important. While some programs decide to develop all their assessments, this is not really a methodology that can be encouraged -- as will be seen, this is a difficult and time-consuming process. Whenever possible, an assessment instrument that already exists should be selected. This does not mean that the assessment must be a nationally-available norm-referenced test (NRT), but only that already existing instruments will be easier to utilize than one that has to be developed by the program. (In fact, it is important to note that neither Title I nor Title VII require NRTs.)

When beginning the search for an instrument, it will be important to identify the

(1) purpose of the assessment (progress, year-end summary),

(2) content of the assessment (achievement in a content area, language proficiency),

(3) language of assessment (English, L1, or both),

(4) type of assessment (NRT, CRT, alternative assessment),

(5) type of scores needed (a quick or detailed scores), and

(6) comparison group, if one is needed.

While the most important definitive issues in beginning the search for an existing instrument are in the box above, other issues, such as who will administer the assessment and how often it will be given also must be considered. A fuller list of these is included in the document "Issues in Designing an Assessment System," located in Appendix III.

To begin the search for instruments, carefully operationalize exactly what should be "tested." Then, identify existing assessments through one or more of four sources:

- ask other programs serving similar students what they use, how they selected the instrument, and what they feel are its strengths and weaknesses;
- look at tests included with curriculum materials being used;
- consider state-, district-, or funding agency-mandated assessments; and/or
- review published lists of tests.

Any assessments identified through these means should be examined by a local team to ensure that they do meet local needs and should be reviewed in the literature to assure their technical quality. Assessments that appear to be appropriate should be examined further. This examination should consist of two phases: identifying critical reviews of the test and obtaining a copy of the test for on-site inspection.

Books that list tests and books that critique tests frequently are one-and-the-same. For example the *Buros Mental Measurements* Yearbook (10th edition: Conoley & Kramer, 1989) is a periodic listing of new and revised tests. The tests are classified by subject area (i.e., achievement, developmental, education, English, fine arts, foreign languages, intelligence and scholastic aptitude, mathematics, neuropsychological, personality, reading, sensory-motor, social studies, speech and hearing, vocations, and a "miscellaneous" grouping). Tests are reviewed (frequently by more than one person) providing information such as validity, reliability, test construction, and references to studies using them. At the end of the book is a directory of

publishers.

Test Critiques (Keyser & Sweetland, 1984) provides information in four areas: a general overview of the assessment (including brief biographies of the developer[s] and a history of development), practical applications and uses, technical aspects, and a critique; references are listed for each test. A list of publishers is included at the end of each volume of *Test Critiques*. While the information generally is not as comprehensive as in *Buros MMY*, it is written in a straight-forward and easily understood manner. Another book of critiques is *Tests in Education* (Levy & Goldstein, 1984). Information provided for each test includes basic information about the test and test publisher, test content, purpose of the test, item preparation, administration procedures, standardization procedures, reliability and validity, interpretation of test scores, and a "general evaluation." Test are divided into the categories of early development, language, mathematics, composite attainments, general abilities, personality and counseling, and "other topics."

Other books of test critiques and lists include *Major Psychological Assessment Instruments*, volumes 1 and 2 (Newmark, 1985 & 1989) and *How to Measure Performance and Use Tests* (Morris, Fitz-Gibbon, & Lindheim, 1987). In addition, there are several books that provide information for specific content areas; e.g., *Handbook of English Language Proficiency Tests* (Del Vecchio & Guerrero, 1995), *A Guide to Published Tests of Writing Proficiency* (Stiggins, 1981) and *Reviews of English Language Proficiency Tests* (Alderson, Krahnke, & Stansfield, 1987). Finally, the Educational Testing Service has published a catalogue of tests (six volumes, about 1,500 tests listed in each volume) in the areas of achievement, vocational, tests of special populations, cognitive aptitude and intelligence, attitude, and affective and personality. These volumes provide only information about publishers and a brief description of what the test purports to do; there is no critical evaluation of the tests. No test is perfect, but these suggestions should help to find the best test possible for a particular situation.

When reviewing NRTs, it will be especially important to look for any forms of bias that might exist. Three types that are common in standardized tests are cultural bias, linguistic bias, socio-economic bias (FairTest, 1995). The first is based on the fact that most NRTs reflect White, North American middle-class experiences and culture. In addition, NRTs, especially those measuring language proficiency, tend to emphasize discrete components of language rather than assessing how well someone actually communicates in English. Another component of linguistic bias is the need for many language minority students to translate items before they can answer them -- a process that takes longer and can handicap them on timed tests. Finally socio-economic bias comes from the presumption of many tests developers that all test-takers will be familiar with middle-class experiences, activities, and language.

Other types of bias in test items that may cause concern are stereotyping, representational fairness, and content inclusiveness (National Evaluation Systems, 1991). Stereotyping is based on a custom or practice that it isolates and exaggerates. Bias also may occur through the under- or over-representation of particular groups such as women, older persons, persons with disabilities, and so on. National Evaluation Systems suggest specific methods for identifying bias due to representational fairness. Content inclusiveness refers not only to the common concern that the test match the curriculum, but also to a concern that the test materials include the contributions, issues, and concerns of a variety of groups from our society, not just the dominant one or two. The local review panel selecting an NRT for use in a Title I, Title VII, Title IX, or other specially-funded program should ensure that these biases are limited to the greatest extent possible. Some considerations are included in Appendix III in the document "Standards for Testing Bilingual Persons."

Alternative assessments may be more difficult to identify and locate. Some books are beginning to offer

examples of alternative assessments in various areas, but there is little critical information about them. Books on alternative assessment that include full instruments include *Portfolio Assessment in the Reading-Writing Classroom* (Tierney, Carter, & Desai, 1991), *Problem-Solving Techniques Helpful in Mathematics and Science* (Reeves, 1987), *Evaluation: Whole Language Checklists for Evaluating your Children* (Sharp, 1989), *Mathematics Assessment: Alternative* Approaches (videotape and guidebook) (National Council of Teachers of Mathematics, 1992), *Assessing Success in Family Literacy Projects: Alternative Approaches to Assessment and Evaluation* (Holt, 1994), and *The Whole Language Catalog* (Goodman, Bird, & Goodman, 1992). Some schools and school districts have begun developing alternative assessments and are willing to share their work with others. For instance, the Orange County Office of Education (Costa Mesa, CA) and a southern California collaboration among the Los Angeles County Office of Education, Los Angeles Unified School District, ABC School District, Long Beach Unified School District, and Santa Monica-Malibu School District each have developed a series of alternative assessments -- the latter is specifically designed for Spanish-English and Portuguese-English bilingual classrooms. The Curriculum Office of the Juneau (AK) School District has published a portfolio system developed for elementary school children (1994a & b).

Regardless of the source of information, regardless of the type of assessment, critical reviews of others in the field, even if they are considered "experts," are not sufficient to justify the selection of one test. Besides knowing that "experts" consider the assessment to be good, program personnel must determine whether the assessment is appropriate for this group of students. For instance, do the test items and subtests match the instructional objectives of the program? Has the assessment been used (or normed) with students similar to those in the program? Are scoring procedures appropriate for the needs of the program? For a fuller list of considerations when selecting an existing assessment, see the documents "Selecting Appropriate Achievement and Proficiency Tests" and "Choosing an Assessment Appropriate for YOUR Program" in Appendix III. In addition, *A Guide to Performance Assessment for Linguistically Diverse Students* (Navarrete & Gustke, 1995) provides a detailed discussion of issues that must be considered when contemplating alternative assessments.

Modifying existing assessment instruments may be necessary in order to have an assessment that truly is specific for the program. Any modifications will require further field-testing of the instrument to ensure that new problems have been introduced to the instrument inadvertently.

Modifying an instrument can take one of several approaches:

(1) modifying the actual items,

(2) offering students other response options (e.g., responding in their home language or using a drawing instead of words to show understanding of a concept),

(3) allowing students to utilize aids such as dictionaries,

(4) allowing students more time on a timed-test, or

(5) providing students in extra test-taking skills.

One method of modification that cannot be sanctioned is translating a test from one language to another. This type of modification will necessarily change the technical qualities (i.e., reliability and validity) of the assessment. In addition, translation can introduce other problems, such as how to translate the intent of the

item as well as the words of the item. For instance, in a math item utilizing quarters and dimes, how would these monetary denominations be translated? What is the purpose of the item -- to determine whether students can add or whether they understand the American monetary system?

If the assessment being considered is a nationally available, commercially published one, modification will be difficult. It will be necessary to obtain the publisher's permission to modify the actual items. Their suggestions/thoughts about other types of modifications should be sought as well.

Alternative assessments will be easier to modify because they tend to be less restrictive in their format and purpose. Also, because alternative assessments are planned to be appropriate for various cultural and linguistic groups, translations are not as difficult as with a multiple choice NRT or CRT.

Regardless of the type of assessment (NRT, CRT, alternative -- locally-developed or commercially-published), any modification of items, directions, or response options will result in somewhat different validity and reliability. The program director and evaluator will need to determine actions that will ensure the best possible testing experience for students. This may involve further training of those who will administer the assessment, it may require a field test to ensure that the assessment still "works" as planned, it may necessitate the evaluator calculating reliability or reviewing validity issues.

Creating instruments is something to be avoided if at all possible! The process is not necessarily difficult, but it is time consuming, labor-intensive, and can be expensive. There are several guidelines for developing instruments (e.g., Herman, 1990; Milllman & Greene, 1989; Morris, Fitz-Gibbon, & Lindheim, 1987). Most agree on a series of general steps that should be followed.

In general, several steps are necessary when creating an instrument:

1. Carefully define and operationalize what is to be tested, including the purpose of the assessment and the type of scores needed;
2. Create a team to work on the instrument -- include one more resource teacher, content-area teacher, paraprofessional, administrator, parent, and student (if test is for secondary school area);
3. Write more test items than needed;
4. Review and edit items;
5. Field test instrument, analyze results;
6. Review and edit items, dropping those that perform poorly;
7. Identify panel to review for cultural, linguistic, gender, socio-economic bias;
8. Pilot instrument, analyze results including reliability and validity; and
9. revise items, scoring -- finalize instrument.

Sources such as Herman (1990) and Millman and Greene (1989) provide "rules" for creating multiple choice items. Various other sources such as *How to Evaluate Progress in Problem Solving* (Charles, Lester, & O'Daffer, 1987), *Whole School Bilingual Education Programs: Approachesfor Sound Assessment* (Del Vecchio, et al., 1994), *Designing Tests that Are Integrated with Instruction* (Nitko, 1989), *Assessing Student Outcomes* (Marzano, Pickering, & McTighe, 1993), and Authentic Assessment of the Young Child (Puckett & Black, 1994), among many others, describe the process for creating alternative assessments for classroom use. Two brief guidelines are included in Appendix III: "Guidelines for developing reliable and valid alternative assessments" and "How to develop a holistic assessment." By carefully following such guidelines and rules, an assessment can be developed that is valid, reliable, and specific to the needs of the program.

Although validity and reliability were addressed previously, the attached materials include "Two major assessment issues: Validity and reliability" that provides nontechnical definitions and methods for ensuring these technical qualities are satisfied. "Ensuring Validity and Reliability" lists several other factors that need to be considered when creating an instrument.

These procedures may seem rather extreme if all that is necessary is a quick view of whether students generally are progressing, or have achieved a majority of the information in a given curricular unit. However, the development process is very important for instruments that will be used across several year to determine whether the objectives of a specially-funded project have been met. It is not unreasonable to expect the development process for a really good instrument to take a year or more. Remember, too, that an instrument for evaluative purposes should not be modified once the evaluation of the program is underway (unless, of course, major problems in the instrument are discovered).

State standards are frequently referred to within IASA. For instance,

> The Secretary shall terminate grants ... if the Secretary determines that (A) the program evaluation ... indicates that students in the schoolwide program are not being taught to and are not making adequate progress toward achieving challenging State content standards and challenging State student performance standards. (IASA Title VII 7114[b][2][A])

IASA mandates that states develop content and performance standards that reflect high expectations for all students. In most cases, "performance standards" not only refers to how well students will achieve, but also refers to assessment measures that states should develop. In fact, Title I specifically requires "an aligned set of assessments for all students" (IASA Title I 1111[b][B]).

As of this writing, most states are still in the process of developing standards for both content and performance. While these are in progress, we recommend the following procedures:

> (1) ensure that the national standards developed by various organizations (e.g., the National Council of Teachers of Mathematics, 1989) are met;
>
> (2) refer to any state frameworks, guidelines, or other information on what students should learn in each grade or in various content areas (for instance, California has state frameworks within monographs such as It's Elementary! (Elementary Grades Task Force Report, 1992). Content standards will be based on such works;
>
> (3) demonstrate that the curriculum utilized by the program does support the frameworks or guidelines. If the curriculum matches the guidelines, it ultimately should support the content standards;
>
> (4) indicate how the students will be tested to ensure that the frameworks or guidelines are met. This may be through state-designed assessment instruments, or locally-developed instruments if the state has not yet completed a set of assessments; and
>
> (5) show how the assessment(s) are developed and scored to show that students are meeting preset standards for performance. This should be the state standards, but may be locally-developed if the local standards are more stringent that the state's or if the state has not yet completed this task.

We encourage those working with linguistically and culturally diverse students to contact their state departments of education to offer their assistance in developing state performance and content standards. Representatives of these students often are added to such panels after much of the work is completed, and thus have little input into the process. By becoming proactive participants in the development these standards, they will be more applicable to culturally and linguistically diverse students and may be developed more quickly with the input of qualified educators from diverse areas.

Scoring instruments is nearly as important as selecting/creating a valid and reliable instrument. Tests, particularly standardized tests, can be scored in several different ways. These scores are only as helpful as they are understandable. The interpretation of scores can be confusing and can lead to erroneous conclusions about the students' performances. Some of the basic types of scores are described in this portion of the Handbook.

Raw scores tell the number of items answered correctly. These numbers can be averaged for a particular test to give an idea of how well the class performed on the average, but raw scores cannot be averaged across several tests. Raw scores can be used to assess **mastery** (e.g., 8 of 10 items answered correctly), but usually are meaningless when presented without other information. As an example, stating that "the average score on a test was 35" has little impact; stating that "the possible scores on the test were 0 to 50 -- these students' scores ranged from 28 to 45 with an average of 35" gives a great deal more information. A more useful score often is the **percentage correct**, which provides more information about how well students have done.

Derived scores, rather than raw scores, are usually used (1) to make scores from different tests more comparable by expressing them in the same metric (the same scoring units) and (2) to let us make more meaningful interpretations of test results. All of the scores described below are derived scores.

Percentiles are frequently used scores, yet still are frequently misinterpreted. They range from 1 to 99, indicating the percentage of students scoring at, or lower than, the test score in question. For example, a student scoring at the 70th percentile scored at least as well as 70% of the other students who took the test; s/he scored higher than 69% of the others. The advantage of percentiles: ease of interpretation; the disadvantage: differences between percentile points are **not** equal throughout the scale (e.g., the difference between the 1st and 5th percentiles is not the same as the difference between the 45th and 49th percentiles) -- because of this, percentiles cannot be averaged, summed, or combined in any way. Occasionally **percentile values** are reported. These are the raw scores associated with a particular percentile score.

Some people confuse percentiles with percentage correct; it may help to remember that someone who scores 100% correct on a test will usually be at the 99th percentile. Percentiles can be helpful in describing the scores of the students (e.g., the students scored at the 55th percentile). Be sure to calculate the average score from raw scores, percent correct, or a standardized score of some type, then convert this average score to the percentile score. Do not average percentile scores without this conversion process.

Grade equivalents or "grade placement scores" indicate how well a student is doing relative to other students in the same grade. Grade equivalents are stated in tenths of a school year (assuming 10 months is a school year), so 7.3 indicates the third month of seventh grade. These scores are extrapolated calculations; they only estimate the relationship between grade levels and test scores. More specifically, they are based on the average performance of pupils having that actual placement in school, even though test publishers probably only administered the test two times during a given year and have estimated the scores for other months and for other grade levels. Grade equivalent scores are based on the tenuous assumptions that (1) what is being tested is studied by students consistently from one year to the next, (2) a student's increase in

competence is essentially constant across the years, and (3) tests reasonably sample what is being taught at all of the grade levels for which scores are being reported. This leads to frequent misinterpretation of grade equivalent scores. The advantage of grade equivalents: grade placement is a familiar concept for most people; the disadvantage is similar to percentiles -- inequality of units, thus inability to average, sum, or combine.

Stanines provide a rough approximation of an individual's performance relative to the performance of other students. Originating from the term "standard nine," stanines divide the range of scores on a test into nine equal groupings. The score of 1 stanine represents the lowest of the nine groups and a 9 represents the highest scoring group. Because of the general nature of stanines, many educators prefer to use these gross descriptors in communicating individual test results rather than misrepresent the precision of the data-gathering instruments and forms. Stanines are not designed to be used for describing the average achievement level of a class or a group of students. Also, the breadth of the scores makes it difficult to report information that is very precise.

Standard scores define a whole set of scoring types, each indicating that a raw score has been recalculated to have a predetermine average and standard deviation (measure of how much the scores vary -- a small standard deviation says that the group scored similarly while a large standard deviation says that the group's scores were very heterogeneous). Advantage of standard scores: equal interval scales allow comparison across students and across tests, and scores can be mathematically manipulated; disadvantage: when making comparisons, be sure that the same type of standard score is available for each test.

A particular kind of standard score is the **normal curve equivalent (NCE)**. NCEs have a mean of 50 and a standard deviation of 21.06; they range in value from 1 to 99 and match the percentile curve at 1, 50, and 99. Advantage of NCEs: as a standard score, NCEs can be mathematically manipulated and do allow for comparisons across students and across tests; disadvantage: it is tempting, but incorrect, to interpret NCEs as percentiles.

Another frequently seen standard score is the **scaled score**. Various test publishers have created their own unique scales that cannot be described in great detail here. Suffice it to say that these are appropriate scores for use in an evaluation, but care should be taken when comparing the scaled score of one test to the scaled score of another test -- this cannot usually be done unless the scales used are the same.

Norms can be based on any of the previously described types of test scores. They refer to test data (test scores) that allow the comparison of a particular score with a group of scores on the same test. Norms give a test score meaning by providing a perspective or context. Because test scores don't always give you the information about how well a student has performed on a given test, norms are used to describe how well the test-taker performed in comparison to other persons (i.e., the norm group). While in theory a student should be compared only against others similar to him/herself, this frequently is not the case. Be sure to read the test manual (for standardized tests) to determine the composition of the norm group -- does it match the group on which you plan to use the test? Some other definitions related to norms are provided below.

NRTs have been given to a large number of students at specific grade levels. When these tests are normed on a large number of students of the same age or grade level on a nation-wide basis, these norms are referred to as national norms. Test scores that allow the comparison of a student's score with the scores of other students of the same age or grade in the local district are referred to as local (or district) norms. Local norms may be compared with national norms to determine whether local scores are similar to, higher than, or lower than scores nationally. Local norms can be used as the nonproject comparison group in a Title VII

evaluation and usually are more accurate than national norms.

Rubrics are not scores per se, but are a way of creating scores, particularly for alternative assessments. Many alternative assessments are checklists, which require merely that the evaluator count the number of behaviors checked -- this forms the score. However, if the desire is to rate the behavior on some scale of "goodness," and to be able to determine whether and how much students are improving or making progress, then a more precise scale that measures specific aspects of behavior should be used. Rubrics generally begin with a zero-point, indicating no response on the student's part, and can go as high as 10 or above. Generally, something between 0-4 and 0-6 is seen most often. For specific directions on how to construct a rubric, see "Creating Your Own Rubric" in Appendix III.

Gain scores are used to show how much students have progressed. The usual method for calculating gains is to subtract the pretest score from the posttest score. This is problematic because no single assessment is perfectly valid and reliable. When gain scores are created, all of the technical problems in both the pretest and the posttest are contained in the single gain score, thus making it, in essence, doubly unreliable and invalid.

As a summary to considering various scoring options, Table 3 lists each of the scoring types based on whether or not they can be used to describe general performance and/or can be used in computations for an evaluation. For more information on test scores, see a book such as H. Lyman's Test Scores and What They Mean (1978) or John Hill's monograph All of Hill's Handy Hintsabout the interpretation of widely used test scores

Table 3.
Test Scores and their Uses

Type of Score	**Compares students against**	**Can be used to evaluate**	**Can be used to describe**	**Not suggested for any use**
Raw Scores	Nothing	**X**	**X**	
Percent Correct	Standard of 100% correct	**X**	**X**	
Mastery Scores	Mastery/non-mastery of content	**X**	**X**	
Grade Equivalents	Norm group		perhaps	**X**
Standard Scores, including NCEs	Norm group	**X**	**X**	
Stanines	Norm Group			**X**
Rubrics	Criterion performance	**X**	**X**	
Percentiles	Norm group		**X**	
Gain Scores	N/A		Perhaps	

Changing test scores is possible. That is, if test scores have been recorded in the students' files as percentiles, it is possible to change these to more usable NCEs. Most test manuals will provide a conversion

table that includes typically reported scores such as raw scores, grade equivalents, percentiles, stanines, and so on. The table provides equivalencies among the scores. For instance, Table 4 is from a particular test's technical manual. It provides the information just described. By reading across the Grade 4 information, scores can be transformed forma raw score 11 to a stanine of 3, NCE of 30, and percentile of 17; in addition, the final columns indicate that a ascore of 11 is at a grade equivalent of 2.7 and an extended scale score of 430. Note that becuase the raw test scores range from 1 to 45, and percentiles and NCEs range from 1 to 99, some percentile and NCE scores are not on the table (e.g. the jump from 8 NCEs to 13 NCEs or from 33 percentile to 39 percentile); this is a function of the scores not having the same range. Figure 2 further demonstrates the relationship between Stanines, NCEs and percentiles.

Summarizing, planning an evaluation requires expertise and attention to detail. Evaluation should not be seen as an "add-on" to the program, required by those who funded the program, but schoyuld bee seen as a key feature that will lead to program improvement. As the document "Goals - Objectives - Activities - Assessment - Evaluation" shows (see Appendix III), these key features of a well-designed, well-planned evaluation really are tied directly together through the evaluation. If any one of these is a "weak link" the entire evaluation can become an exercise in futility that will lead misinterpreted and misunderstood results for the project. The next section in this handbook deals with the implementation of the planned evaluation.

(table of contents)

Table 4.
Conversion table for standardized test

Figure 2.
Relationships among Stanines, NCE's and Percentiles

NOTE: These figures could not be included in the electronic version of this document.

IV: Implementing an Evaluation

> Major reforms in education have consistently been accompanied by major reforms in methods of evaluation. In the 1930s, 40s, and 50s, the advances in evaluation were mainly in assessing student performance. Starting in the 1960s, however, there were, in addition, many developments related to the assessment of educational programs, projects, and materials.
>
> Joint Committee on Standards for Educational Evaluation (1981, p 2)

Thinking and planning the evaluation are now complete. It is time actually **to do** the evaluation. It will be important to follow the evaluation design that has been developed. However, it also will be important to recognize that there may be problems with the design. "Guidelines for Managing the Evaluation Plan," located in Appendix IV, provides some information about controlling the evaluation and ensuring that it continues to meet its purposes.

This section of the Handbook describes the activities that take place as the evaluation progresses. This includes training the staff, collecting data, and analyzing data. The last section, analyzing data, provides brief overviews of the statistical designs most frequently used within Title VII -- more details are provided in Appendix IV. In this way, staff can become familiar with the concepts of each design while information is available in the appendix that will allow evaluators to implement the design.

Training teachers, evaluators, administrators, or others to administer, score, or interpret assessments will be a key element for any educational program. As stated by Lyman,

> The typical school system has few teachers who are well trained in testing, because most teachers have had little opportunity to take elective courses while in college. Few states require tests and measurements courses. (1978, p 4)

Unfortunately, this has not changed much since 1978.

Administering tests takes some talent. Standardized tests usually have a test administration procedure that should be followed. The instructions are well developed and need only be read and followed. Alternative assessments are more difficult. In this case those administering the assessment may need training to ensure that they do not give different cues to students that might affect students' scores.

Observation measures require a different type of training. In this case, the assessment is not administered to the student but is completed by the teacher or other test administrator. Training will be necessary to ensure that the rubrics are understood (e.g., what is the difference between "frequently" and "often"?) and to ensure that the teacher is reflecting on the student in an appropriate manner (e.g., should playground activities be included, or only classroom activities?). In addition, some proficiency instruments, such as the Student Oral Language Observation Matrix (SOLOM, included in Appendix IV), are to be administered by a native speaker of the language being tested. How can all of these issues be addressed? Training of those who will administer or score the instrument is, once again, key.

> Those administering tests should be trained in:
>
> - reading the directions;
> - explaining allowable modifications
>   - use of dictionaries or word lists,
>   - extended time for responses,
>   - language(s) of responses,
>   - use of alternative forms of response (e.g., drawing a picture);
> - providing assistance -- if allowed, when allowed, etc.;
> - encouraging students who are having problems;
> - scoring the assessment, if approriate; and
> - Interpreting the assessment results for students and families.

In the case of an observation instrument, a series of training events should be considered. Let us use the SOLOM as an example, although the procedures can be generalized to other observation-type instruments as well.

- Create videotapes of students in an appropriate setting (e.g., classroom). Ensure that students of varying English proficiency levels are included in the videotape.

- Ask "experts" to assist in scoring the videotape vignettes. These expert scores will be the standard against which trainees will be measured. Ask the experts not only to score the vignettes, but to explain their scoring.
- Allow trainees to view the videotape several times before attempting to score it. (This is appropriate since teachers presumably would have seen their students on several different occasions before attempting to complete an observation measure about them.)
- Explain the scoring system. Utilize the first 1 or 2 vignettes in demonstrating the scoring procedures to the group.
- Trainees should score the other vignettes on their own during the training session.
- Review the scores of the experts. Any vignettes that trainees score differently should be discussed in depth.
- Work with trainees until at least 80 percent of the scores are the same. (This is inter-rater reliability -- scorers should agree on at least 80% of the subjects they score.)
- Periodically review the training procedures with teachers. This will ensure that their inter-rater reliability remains high and that they are scoring appropriately.

Scoring assessments is another area in which teachers should be trained. This is especially the case if alternative assessments are utilized that have scoring procedures that measure more than "correct" and "incorrect." This standardization process will ensure that all students are scored in the same way. The procedures were described in "Creating your own Rubrics" that appeared in Appendix III.

Data collection procedures must be standardized and begun almost before the program begins. Staff must be trained regarding the importance of record keeping and methods of checking data collection procedures. A technique for ensuring and checking the accuracy of the records should be implemented. Most importantly, the methods for data collection should be as simple and straightforward as possible. The forms themselves should add as little work to the teachers' load as possible. This is an area in which the evaluator will be of great assistance. As an example, Appendix IV contains a "Student Data Sheet for Use in Multi-Year Evaluation." This form meets all the Title VII requirements for data collection, and will meet the requirements of most other programs as well. One of the advantages to this form is that the basic data for one student enrolled in a 5-year project can be collected on the single two-sided sheet. Evaluators may prefer to create their own data collection forms based on the more specific purposes of the program being evaluated and on the data available from the program, school, or school district.

We suggest that more data rather than less data be collected. It is always possible to collapse data into larger categories, but once the data is collected it is difficult to break it into more detailed groupings. Also, remember that some programs require that data be reported for specific groups. For instance, Title I requires that data be

> ... disaggregated within each State, local educational agency, and school by gender, by each major racial and ethnic group, by English proficiency status, by migrant status, by students with disabilities as compared to nondisabled students, and by economically disadvantaged students as compared to students who are not economically disadvantaged. (IASA Title I 1111[b][3][I])

While Title VII does not specifically require this breakdown of data, the data should be collected to allow such an analysis since the two programs (Title I and Title VII) may serve some of the same students.

Formative evaluations are performed to ensure that the program is working as well as possible and to determine whether modifications might be needed. As Beyer says,

> we cannot predict exactly and with confidence how an idea will work in practice. In developing an innovative program ..., we may have a good reason to believe that our innovation will work as intended -- or at least should work -- but we don't know, beyond a reasonable doubt, whether it actually will work. ... How do we know it will work? In the field of curriculum and instructional development, formative evaluation can answer this question. (1995, p 1; original emphasis)

Formative evaluation usually will require the same data as the summative evaluation. The major difference is that full data analyses frequently are not performed for a formative evaluation. Rather, the purpose of the formative evaluation is to ensure that the development of the program is proceeding in a timely manner and that there are no gaps or problems that should be addressed immediately. For instance, a quick review of data, with actually doing analyses, may indicate that older students are responding well, but younger students are not improving their performance or such a review may indicate that one language group's scores are increasing, but not another's.

Data for formative evaluations can be collected through a wide range of procedures and instruments, including

- Annotated analyses of print materials,
- Questionnaires and surveys,
- Quantitative performance or achievement assessments,
- Examination of student- and teacher-produced products,
- Learning and teaching logs,
- Error logs,
- Observations,
- Interviews and focus groups,
- Video and audio recordings,
- Anecdotal records, and
- Open-ended critiques or reports (modified from Beyer, 1995).

Summative evaluations are more formal and specific than formative evaluations. In general, data analyses are required to show that objectives have been met and/or that the students in the program have progressed more rapidly, or in greater depth, than those not enrolled in the program. Data analyses that can be used for evaluative purposes are described in the next section.

Data may be collected for summative evaluations from all the sources listed above as appropriate for formative evaluations. In addition, it will be important to collect data regarding student performance before the program began, and at the end of each year of the program. This will allow a more definitive statement about the progress of students. In addition, if the professional development of staff is important, the information about the skills and proficiencies of the staff before and after the training programs will be important as well. Any of this data may be collected through the use of NRTs, CRTs, and/or alternative assessments. Neither Title I nor Title VII require a specific type of assessment be used.

Data analysis probably is the aspect of evaluation design that sets the most nerves on edge. Remember that this is an area in which the evaluator should be a major player. The evaluator should be an expert in data analysis -- in several forms, not just the type of analysis that s/he prefers and routinely uses. In general, most designs can be analyzed using one of three approaches: (1) grade cohort, (2) gap reduction, or (3)

quasi-experimental comparison. Each of these designs allows includes a comparison between students enrolled in the educational program and students who are considered the nonproject comparison group. Currently, IASA Title VII projects, among others, require this comparison when analyzing data for school retention, academic achievement, and gains in language proficiency (EDGAR 34 CFR 75.590[b]). Note, however, that this does not mean that every assessment used must have a nonproject comparison group -- it may be most appropriate to utilize a nonproject comparison group only at the beginning of the program, and then on an annual basis.

In this section we describe these three types of analyses -- with a reminder that not all objectives will require statistical analyses. It may help to review the materials on goals, objectives, and activities at this point.

Evaluating objectives is first a job of reading the objective carefully and determining the type of analysis to be done. In many cases, statistical analyses are not needed. To determine whether statistics are needed, (1) review the requirements, regulations, or statutes of the funding agency (some specify fairly specifically how evaluations should be performed, or at least mention specific comparisons that should be made) and (2) review the program's objectives to determine whether their language necessitates statistics (key words: "significantly higher/lower scores," "statistically higher/lower scores"). See Appendix IV for the document "Matching Objectives to Evaluation Design" for more details on how to assess objectives when statistics are not necessary.

Grade cohort is a technique first developed by Beverly McConnell for evaluations of educational programs serving migratory students (McConnell, 1982). Its key feature is that it allows the evaluation to include students who have been involved with the program for as few as 100 days, instead of requiring that students be in the program for the entire school year. Basic information about the grade cohort design is provided below. In addition, several documents are available in Appendix III: "KEYS TO ... Testing differences between groups: Grade Cohort," "Basic Grade Cohort Design," "Advanced Grade Cohort Design," and "Data Presentation for Grade Cohort Design."

The grade cohort design answers the question *What are the achievement gains of students who have been in the program for 1 "year" (or more) as compared to students who have not yet received the program's services (or have received fewer "years" of services?*

The basic design requires that

- students be pretested and posttested with the same assessment instrument,
- tests be given on a set, periodic basis (e.g., every 100 days), and
- data be collected by language group by grade level.

The design allows

- students to enter the program throughout the school year,
- various data analysis to be utilized, and
- collection of data across the years.

Background. The grade cohort design is a quasi-longitudinal design (which translates to semi-long term) which originally was designed for programs serving migrant populations. In this original form, the design required that students identified as needing a program be pretested with an NRT before they entered the

program. Students can enter the program at any time during the year, as long as they all are pretested with the same NRT. During the school year, students are posttested as soon as they complete 100 days of the program. Because students can enter the program at any time, students may be posttested at different times during the year (e.g., Students 1 through 5 enter the program at the beginning of the school year, they are tested 100 days later; Student 6 enters the program on day 6 of the program, she will be posttested on day 106; Students 7 and 8 enter on day 15 of the program, they will be posttested on day 115). The same NRT is used for all pretesting and all posttesting. When students have completed a second 100-days of the program, they will be posttested a second time with the same NRT. Each 100-day period is considered to be a "year" of education within the program.

In recent years, some modifications of the grade cohort design have been suggested. These modifications include the suggestion that NRTs are not the only assessments appropriate to the design. Instead, any appropriate assessment that can be used to pretest and posttest can be used as long as the instrument's reliability and validity can be documented. Another major suggestion is that the unit of measure need not be 100-day increments. Instead, the unit of measure might be units of an educational program. These units should not be small pieces, but units which demonstrate a major growth on the student's part. This modification is most appropriate for adult students involved in a literacy program with several "tracks," not all of which are required of each student. Finally, a program might prefer to utilize the traditional academic year rather than defining 100-days as a "year."

Nonproject comparison group. The grade cohort design utilizes a "live" comparison group. All students who enter the program are considered the nonproject comparison group, regardless of exactly when they enter the program -- as long as neither the curriculum nor the assessment have changed. The comparison group always will be larger than the project group. This procedure allows the evaluation of a small group of students since the comparison and project groups can be added to across time as more students enroll; this is especially helpful in bilingual education programs that may serve small numbers of students in any given year.

Data for each language group (Spanish-speaking, Farsi-speaking), each language proficiency (LEP, NEP, FEP), within each grade level (grade 2, grade 3) should be maintained separately. Other information can be separated out if that is of interest to the program (gender, length of residency in the school district).

Data analysis. Descriptive statistics should be provided for each group of students (e.g., Farsi-speaking LEP students in 2nd grade). Various analyses are possible based on the expertise of the evaluator and the staff, as well as the number of "years" of data that has been collected. For instance, a simple analysis might determine whether there is a change in 3rd grade scores from the time students entered the program until 100 days later. A more advanced design might determine whether there are changes from the time students enter 1st grade until they complete elementary school after 5th grade. Analyses should be completed for each group of students for whom data has been collected. It may take two to three years before enough data has been collected on each group to do a full analysis of all subjects, but analysis of some student groups may be possible sooner.

Benefits/Problems. The design does meet the requirement for a nonproject comparison group that is similar to the project students. It controls for several problems that can affect the validity of the evaluation; e.g., the history of the students, the maturation process, testing problems, and students who leave the project (mortality). Also, the grade cohort design readily allows a more longitudinal emphasis which may be helpful for IASA and other longer-term projects.

The major problem to this design is the time involved in collecting data on enough students to allow an evaluation. However, as opposed to the designs that do not allow the evaluation of small groups of students at all, this is a fairly minor problem.

Gap reduction is a technique first developed by Tallmadge, Lam, and Gamel (1987) because "we assumed that most [bilingual education] projects would find it difficult or impossible to implement a traditional true or quasi-experimental design. [The design] is easy to implement, satisfies the regulations' requirements, and does not require a nonproject comparison group made up of students similar to those served by the project" (original emphasis, p 3). Basic information about the gap reduction design is provided below. In addition, "KEYS TO ... Testing differences between groups: Gap Reduction," "Gap Reduction Design Elements" (there are no "advanced" techniques within the gap reduction design), and "Data Presentation for the Gap Reduction Design" are presented in Appendix IV.

The gap reduction design answers the question *Has the difference (gap) between the project group's performance (average test scores) and the comparison group's performance been reduced across the school year?*

The basice design requires that

- students be pretested and posttested with the same assessment instrument,
- appropriate scoring methods be used (preferably NCE's), and
- the same students be utilized for both pretest and posttest.

The design allows

- a comparison against the national norm (50 NCE's) or grade-mates,
- simple analysis bases on subtraction (i.e., no statistical tests), and
- comparison(s) of nontest data (e.g., number of absences, library books used).

Background. The gap reduction design was developed to help local evaluators overcome four flaws that the authors saw in bilingual education evaluations: lack of evaluation expertise at the local level, inadequate guidelines for evaluation, insufficient technical assistance for local projects, and limited availability of funds for evaluation purposes (Tallmadge, Lam, & Gamel, 1987). Conceptually, the design is quite easy. The only requirements are for pretest and posttest data for two groups (the project group and the comparison group). The data might be test performance (an NRT or an alternative assessment score) or a behavior (number or percent days absent from school, number of library books checked out during the academic year, number or percent students referred to gifted/talented education or special education). The nonproject group's average score is used as the basis for the comparison with the belief that across the school year the project students should become "more like" the students not in the program. If using NRTs, testing dates should be one year (i.e., 12 months) apart.

In recent years, some modifications of the gap reduction design have been suggested. The most major of these modifications is that criterion-referenced tests might be used with no nonproject comparison group at all. In this case, the comparison is based on the criterion score that has been determined to show success. For instance, if the "passing" score to show mastery of the content area is 80% correct on the project-developed CRT, than 80% correct becomes the score against which the project group's average score is compared. Students are still pre- and posttested, but now their posttest average score should come closer to the 80% criterion than did their pretest average score. Similarly, a predetermined score on an alternative

assessment can be used as the "comparison" score. As an example, scores on a rubric might range from 0 (no response) to 6 (full response), with a score of 4 indicating an adequate response. The score of 4 could be used as the comparison; in a sense this is the criterion for success on this assessment.

Nonproject comparison group. The gap reduction design can utilize either a "live" comparison group or a test-specific comparison group. In the first case, the comparison would be the average score of students in the state, district, similar school, or same school not enrolled in the program being evaluated. In the latter case, the national norm of an NRT (which always will be 50 NCEs), or the criterion score of the CRT or an alternative assessment's rubric. When using nontest data (# of library books or days absent), the comparison should be against other students at the school who are not enrolled in the project.

Ideally, data for each language group, each language proficiency, within each grade level should be maintained separately and analyzed separately. It may not be possible to do this if the numbers become very small. In general, because statistical analyses are not utilized, a minimum of 10 to 15 students in each group would be sufficient. However, with such small numbers, the generalizability of the information is questionable. References and generalizations should be made about the students in the program only, not to the population of students who might be enrolled in such a program.

Data analysis. Descriptive statistics should be provided for each group of students. The analyses for the gap reduction design are quite simple, based simply on subtracting scores from one another. The specific steps involved are

- Subtract the project student's average pretest score from the comparison group's average pretest score -- this is the pretest gap.
- Subtract the project student's average posttest score from the comparison group's average posttest score -- this is the posttest gap.
- Subtract the posttest gap from the pretest gap -- this is the gap-reduction.
- Interpret the gap-reduction:
    - a positive number means the gap has been reduced (a successful program),
    - a negative number means the gap has become greater (not a successful program), and
    - a gap-reduction of zero indicates that the gap has remained the same.
- Create a graph (see the example in Appendix IV) to visually demonstrate the gap-reduction.

Benefits/problems. The design does meet the requirement for a nonproject comparison group that is similar to the project students. It is very easy to use since it requires no statistical tests. It is helpful that the design does allow for the analysis of nontest data that might be appropriate for various projects.

The major problem is that there is no statement of what amount of gap-reduction is "good." It would seem logical that 5 points of gap-reduction is better than 1 point, but there is no minimum that should be considered to indicate a successful program. A related problem is that gap-reduction removes evaluation from the actual scores of the students. Rather than reporting the actual scores, some reports have provided only the gap-reduction. This provides little information for the reader about how well, in an "absolute" sense, the students performed.

Non-equivalent comparison group designs (the t-test design) are frequently used to evaluate educational programs. This is a traditional method for analyzing the results of two groups of students (those enrolled in the project and the comparison group) at two different times (the pretest before the program began and the posttest at the end of the program year). Basic information about this design is provided below with more

details in the documents "KEYS TO ... Testing differences between groups: t-tests," "Basic Nonproject Comparison Group Design," "Advanced Nonproject Comparison Group Design," and "Data Presentation for t-tests" in Appendix IV.

The t-test, or non-equivalent comparison group, design answers the question *How does the performance of students who participate in the program (treatment group) compare statistically to the performance of similar students who are not in the program?*

The basic design requires

- similar treatment and comparison groups,
- the same test for each grooup, with appropriate scoring methods,
- similar scores for both groups on the pretest, and
- a statistical test of significance be performed on the test scores

The design allows

- various data analysis techniques to be used

Background . The non-equivalent comparison group is a quasi-experimental design. It recognizes that having identical project and comparison groups is not realistic, particularly when dealing with programs to serve culturally and linguistically diverse students. For example, in the case of bilingual education, a student identified as needing services cannot legally be denied services.

Nonproject comparison group. Both "live" and paper comparison groups are possible with this design. The nonproject comparison group can be the norm group of the NRT being used, as long as the norm group contains students who are similar to those in the program being evaluated. The average score for the state, district, or school also are possible. The comparison group should match the group of students enrolled in the project as closely as possible. Thus the best comparison group usually will be students in a similar school that does not have a program similar to that being evaluated, or students within the same school who are not enrolled in the program (e.g., English-speaking students of the same cultural heritage and same socio-economic status as the students in the program).

If possible, data for each language group, each language proficiency, within each grade level should be maintained separately and analyzed separately. However, there frequently are not enough students to allow this type of analysis. In this case, report the information in descriptive fashion, then explain how the students were aggregated to allow the analyses to be performed.

Data analysis. Descriptive statistics should be provided for each group of students. Various analyses are possible based on the expertise of the evaluator and the staff, as well as the number and type of tests being used and so on. In its simplest form, the analyses can be performed through a series of t-tests (hence another name for the design). t-tests are analyses utilized to determine whether there is a statistical difference between two average scores. In this case, four t-tests would be needed:

- to test the difference between the two groups' pretest scores (ideally, this should be nonsignificant, indicating that the two groups are similar);
- to test the difference between the project group from pretest to posttest (this should be significant, with the posttest average score significantly higher than the pretest score -- the students' achievement

level has increased);
- to test the difference between the pretest and the posttest for the comparison group (again, a significant difference is anticipated, with posttest scores higher than pretest scores); and
- to test the difference between the average posttest scores for the two groups (ideally, the project group's average score should be higher than the comparison group's score, indicating that their achievement level has outdistanced their cohorts).

In addition, providing a pictorial representation of the data is helpful.

As a general rule of thumb, a minimum of 10 to 15 students in each group is necessary for a t-test analysis; if another type of analysis is performed, more students will be needed. Again, more students would be needed (at least 30 in each group) in order to generalize to a larger population of students.

Benefits/problems. The design meets the requirement for a nonproject comparison group. Because of the way in which the nonproject comparison group is selected, many problems such as mortality, history of the students, maturation of the students and so on are controlled for. In addition, some researchers prefer to see actual data analyses with statistical results before they will support the success of a program.

The major problem with the t-test design is the difficulty in locating a nonproject comparison group that truly is similar to the group in the educational program. The more dissimilar the two groups are, the less valid the design, and the less believable the overall results. A related issue is the costs involved because more students must be tested to allow appropriate analyses. The manner in which students are selected for the two groups also can cause problems, especially if the students in the nonproject comparison group are aware that the project students are receiving special treatment.

Summarizing, actually implementing an evaluation requires expertise and attention to detail. It will be helpful to have a detailed plan, including who is responsible for each aspect of the evaluation and a timeline for completing the various tasks, an experienced evaluator, and a staff training program. While analyses are not required for formative evaluations, they usually are needed, even if not required, for summative evaluations.

Maintaining quality control is essential when implementing the evaluation. The operative word here is "quality" -- the quality of the assessment instruments, training of staff, the evaluator, the evaluation plan (both for formative and summative evaluations), and, as a result, of the overall educational program.

Now that the evaluation plan has been implemented, the last stage of the evaluation is the report itself. The next section of the Handbook deals with the evaluation report.

(table of contents)

V: Writing an Evaluation

> Most reports of educational evaluations are not understood by laypeople and are widely misinterpreted. In fact, they are not generally understood by teachers and administrators, and, as a result, the information that could provide a basis for improving the educational

> program or institution is not communicated.
>
> Tyler (1990, p 733)

Evaluation reports usually are written either to show progress toward reaching the stated goals and objectives and find areas that can be improved (a formative report) or to summarize the overall effects of the program (a summative report). The overall purpose of all reports is to communicate the effects of the program to the program staff, "clients" of the education programs (e.g., students, parents), funding agency personnel, and the community at large. Unfortunately, many reports are sent to the funding agency for accountability purposes and are ignored by the other potential audiences. To some extent this may be the fault of professional evaluators who write in technical jargon without acknowledging the needs of various lay audiences or the program staff itself. As Nevo points out, "It is the responsibility of the evaluator(s) to delineate the stakeholders of an evaluation and to identify or project their information needs" (1983, p 125). Because there are so many potential audiences for the evaluation, this section of the Handbook will focus on the needs and requirements for Title VII evaluation although it should be remembered that these needs and requirements can be generalized to other funding agencies and evaluation as well.

Tyler (1990) suggests that there are several problems with most evaluations that minimize their usefulness to practitioners in the field. It is worth mentioning his primary concern: the "prevailing practice" of reporting test results in abstract numbers such as grade equivalents and percentiles that "have the appearance of clarity, but, in fact, are interpretations of hypothetical referents that are often different from the actual situation" (p 733). To remedy this situation, Tyler suggests that "the results of school learning can be much more directly defined, identified, and described in meaningful terms that are relatively concrete" (p 734). The evaluation of products of learning is recommended along with dropping the practice of reporting average group scores and moving toward reporting numbers and percentages of students who reach certain criterion levels of work. In addition, there are some suggestions that can be made about practices in writing evaluation reports that are applicable to all types of evaluations for all audiences. For instance, ensure that the report is visually appealing with minimal use of technical terms. Be objective, providing both positive and negative findings with plausible explanations that are based on the data and other specific information available.

Program improvement is the purpose of most formative evaluations. This implies that the program will be continued and can be bettered. This is the origin of the term "formative" evaluation -- it is a report written to show what is happening to the program while it is in its formative stages, written to explain what is happening, how, and why. The formative report will identify any problems that may be occurring (or potentially may occur) and will suggest that they can be ameliorated.

> The purpose of the formative report (in IASA Title VII-ese the "annual progress report") is to
>
> - provide background information about program,
> - demonstrate progress toward meeting the goals and objectives,
> - explain why activities or objectives have not been implemented as planned,
> - furnish information about current budget expenditures, and
> - give any other information requested by the Department of Education.

When reviewing test data and comparing the results to the goals and objectives of the project, it is easy to say "we made our goals" or "we need to improve our project." More important, however, is to go a step further and determine why this happened and what can be done about it. Below are a series of questions that

should be considered when preparing a formative evaluation. While reviewing these questions, remember that although it is easy to focus on test results alone, the answers to the questions also should include such information as attendance records, enrollment in postsecondary education, gifted/talented education programs and special education, grade retention, and so on.

How well was the program implemented? When reviewing the program that was planned, and comparing it to the program that actually exists, what are the differences? When determining the degree of program implementation, it may be important to consider cultural and ethnic sensitivity in the school curricula, support services, and extra-curricular activities; flexibility in the curriculum; teaching strategies in both native language and English; staff knowledge and experience with linguistically and culturally diverse students, autonomy in the decision-making process, and collaboration with the community; and administrators' understanding and knowledge of all facets of the school program, collaboration among school, agencies, and organizations, and support of the program. Finally, the impact of the parents and family on the program should be considered.

What is the context within which the program is working? The school program does not operate within a vacuum. It will be important to determine the overall climate of the school in the way it values students' languages and cultures, maintains high expectations for all students, and demonstrates high morale; the way management integrates the needs of all students into aspects of the school; and how the various resources, including capacity and time as well as financial, are allocated to various programs within the school. Title VII specifically requests information about how the Title VII bilingual education program is coordinated with any other federal, state, or locally-funded programs also operating on the school campus.

Are outcomes due to program effects? A review of program implementation as related to the assessment data will indicate those areas in which the outcomes are due to the program implemented. For instance, students' assessment data should improve primarily in areas in which the program truly was implemented. In areas in which program implementation was weak, student improvement should be minimal. If, however, there are major areas in which the students did improve in spite of weak program implementation, outcomes are not due to program effects.

Was the program differentially effective? Results of the program should be compared across ethnic or language groups to ensure that the effects were similar for all groups. Similarly, the results should be compared across grade levels to ensure that the program is equally effective for all age groups. If differences are found by ethnic/language group or by grade level, (a) ensure that assessments are appropriate and without bias, (b) determine whether the program was implemented fully for all groups, and/or (c) modify the approach used with the lower-scoring group(s).

Are program effects lower than expected? If the assessment data show that program effects are lower than had been expected (and was stated in program goals and objectives), four areas must be considered:

- Was the program fully implemented? If not, this can explain the discrepancy; steps should be taken to ensure that the program is implemented as planned as soon as possible. If it has been fully implemented, consider modifying the curriculum to meet the needs of the students more closely;
- Were expectations reasonable? Perhaps the expectations for student gains were unreasonably high. Project staff should investigate whether the expectations should be lowered to a more moderate level;
- Was the curriculum appropriate for the students? A curriculum that is too difficult for the students also can explain lower assessment scores than anticipated; and
- Did the student population change? If the program had been planned for one groups of students, but

the population changed during the planning phases of the program so that a different population actually is being served, the program's effects might be quite different from those anticipated.

In addition, various extraneous variables, such as the opening or close of a factory in the community, an outbreak of a contagious disease, or a local disaster might affect the students enrolled in the program.

Are program effects higher than expected? If program effects are higher than expected, investigate whether the expectations were reasonable. Program effects may be high because the program is new and exciting to staff and students, leading to higher-than-normal participation. If this is the case, modifying the expectations may be premature; a second year of the program may be necessary to verify that the expectations truly were too low. On the other hand, it may be obvious upon further examination that the expectations were too low and should be raised immediately. The test(s) being used also should be reviewed to ensure that the students are not "topping out" due to an easy test. In addition, changes in student population and the effects of extraneous variables should be considered.

Are all instructional components successful? The outcome data should be examined in detail to determine whether some instructional components are more successful than others. If this is the case, implementation of those less successful components should be investigated. In addition, revised instructional methods may be needed for these components. Don't forget that just because something is successful doesn't mean that it cannot be improved further.

Should the project be institutionalized? The goal of the specially funded project should be to mainstream its methods into the school/school district for the benefit of all students. Demonstrating a high success rate through increasing test scores is one way to argue for the institutionalization of project methods and practices.

Title VII reporting methods require an annual progress report, which is a brief formative evaluation report. This is necessary to receive continued funding. The purpose of the annual progress report is to report progress toward accomplishing the objectives of the project, explain why a planned activity or objective was not attained and how this will be remedied, furnish financial information, and to provide any other information the Department of Education may require. Many funding agencies have such requirements; be sure to follow the guidelines provided. For more information on the Department of Education-required progress report for program improvement, see Appendix V for the document "Instructions for the Annual Progress Report."

Summative reports generally are required at the end of the funding period. For Department of Education-funded programs, a biennial (every 2 years) evaluation report is required, plus a final report for the entire funding period. The summative report should include information from the formative reports (annual progress reports) as well as provide an overview of the overall success of the program. How much did students progress during the life of the program? As an additional purpose, the evaluation report should disseminate information about the program to others who might be interested in implementing such a program, who are researching the topic, or who might be interested in policy issues related to the topic.

The purpose of the summative evaluation (in IASA Title VII-ese the " biennial report") is to

- provide information for program improvement,
- define further goals and objectives,
- determine program effectiveness, and

- fulfill the requirements of the Department of Education

A specific format generally is not required for the evaluation report. In general, an all-purpose format would include information about the background of the project, program context indicators, program implementation indicators, data pertaining to students' academic achievement and language proficiency, data regarding changes in self-esteem or attitudes, and any other information requested by the funding agency. A potential "Evaluation Outline" is included in Appendix V. Its various components are briefly defined and explained below.

Although the ***Executive Summary*** is the first portion of the report that anyone will read, it should be the last one written. This section should be from 3 to 5 pages long, with bullets providing as much information as possible. For many readers, this will be the only section read, so make the important points, make them quickly and succinctly, and leave the reader with an overall feeling that s/he understands the basics of the project.

Some of the key questions to consider when writing the Executive Summary include:

- What was evaluated? What does the program "look like?"
- How was the evaluation conducted? Were there any major constraints?
- What are the major findings and recommendations of the evaluation?
- Is there any other information someone should have to understand the project?

The ***Introduction*** should include all information necessary to understand the context and implementation of the program from its inception through the current reporting period. It describes how the program was initiated and what it was supposed to do. The amount of detail presented will depend upon the audience(s) for whom the report is prepared. If the audience has no knowledge of the program, it must be fully explained; if the report is primarily intended for internal use and the audience is familiar with the program, this section can be fairly brief, setting down information as a reminder of what occurred. Regardless of the audience, if the evaluation report will be the sole lasting record of the program (e.g., the final report for a Title VII project), then this section should contain considerable detail. Much of this can be written during the course of the program by the program director or staff -- the evaluator does not need to be involved in this aspect of the report. Besides the evaluator and program staff, information might be gained from sources such as the program proposal, minutes of faculty and parent meetings, curriculum outlines, budget forms, district information, and so on.

Some of the questions that should be answered in the introductory section of the report include those listed below.

- Where was the program implemented? What sort of communities? How many people were involved (students, families, staff)? What special groups were involved?
- How did the program get started?
- What kind of needs assessment or screening procedures were utilized? What were the results?
- What was the program designed to accomplish? What were the goals and objectives? Were there local, state, or federal constraints on the project?
- What has happened in previous year(s) of the program? What improvements have been made? Why?

The program staff and director probably know much of this information, but references to documents and cross-checking with other individuals will help ensure the consistency of the information.

The ***Methodology*** section describes and delimits the evaluation study undertaken by the evaluator and the program staff. It explains how the evaluation was conducted. It is important to provide enough detail that readers will have faith in the outcomes and conclusions of the report. Descriptions of the selection and development of instruments should be included as well as their technical qualities (i.e., reliability and validity) in relation to this group of students. Samples of instruments should be included although well-known tests only need be referenced. It is essential that program staff as well as the various audiences for the evaluation report agree that this is a fair measure of the program. Some of the questions that should be considered in this section are listed below.

- What is the evaluation design? Why was this one chosen? What are the limits of this design?
- How were instruments selected? Are they the most appropriate available for these students, this curriculum? How can validity and reliability be demonstrated?
- Were instruments developed by the staff? What was the development process? What was done to ensure the validity and reliability of the instruments?
- Were instructors trained to administer, score, and interpret the results of the testing instruments? How was this done?
- What was the schedule for data collection? When were instruments administered? Were all students measured, or were sapling procedures used?
- What other types of information were collected, by whom, and when? What was the purpose of the various data collected (e.g., context indicators, implementation indicators)?

The ***Findings*** are the heart of the evaluation report. This section presents the results of the various instruments described in the Methodology section. If the instruments and other data were relevant, reliable, and valid, these results constitute hard data about the program. In addition, this section also might include some soft data that will enliven the report and provide results that cannot be expressed in numbers -- anecdotes, testimonials.

Before writing any of this section, all data analysis should be completed. Scores from tests should be presented in tables, charts, or graphs; results of questionnaires frequently are summarized on a copy of the questionnaire itself, which may appear in the text or in an appendix. Three general areas of data collection will be presented here: program context, program implementation, and student outcomes (all of which were described in the "What to assess?" section of Planning the Evaluation). In addition, the success of the program in meeting the goals and objectives and a discussion of any unanticipated results should be included.

Some of the questions that should be considered are listed below.

- What is the climate of the school regarding culturally and linguistically diverse students?
- How has management reacted to the program? What support has been received?
- How are various resources allocated to school programs, including the one being evaluated?
- How does this program interact with other programs on campus?
- Was the program implemented as planned? If not, what happened? Were some components dropped or modified?
- How were staff, administrators, families, community members involved?
- Were all curricula available, with appropriate materials? Does the curriculum match the state content standards?
- What numbers and kinds of professional activities were offered? How successful were they? How

were they selected? Who was involved?
- What were the language(s) of instruction? Were these appropriate for all students?
- Did changes occur in the program? Why? What?
- What did the program finally look like?
- How many students were involved? What did they "look like?"
- What are the students learning about themselves, language(s), and content areas? Are students achieving the state performance standards?
- Were the goals and objectives of the program met?
- Did anything unanticipated happen? Why? What?

The last section of the report is the ***Conclusions, Discussions, and Recommendations***. This will provide a final interpretation of what happened during the program, why it happened, to whom it happened, and how the program can be improved. In presenting this information, some of the key questions are:

- How certain is it that the program caused the results?
- How good were the results of the program?
- Why did the anticipated results not match the actual results?
- What happened within the program (context and implementation) that impacted the results (student outcomes) most greatly?
- What can be done to improve the program?

The two most frequently read sections of evaluation reports are the Executive Summary and the Conclusions, Discussion, and Recommendations. These should be written as strongly and as clearly as possible.

Presenting information in a logical, clear fashion can be difficult if the program is complex and the evaluation design is difficult. When writing the report, there are some practices that can help provide the information in the best possible manner. Some of these are presented in the box below.

When writing an evaluation report (formative or summative):

- Address all points specified in the funding agency's guidelines,
- Aviod using technical terms or jargon,
- Write in the active voice,
- Use a visually appealing format, including tables and figures,
- Organize the findings around objectives or evaluation questions,
- Be objective, reporting both positive and negative findings
  - Provide plausible explanations wherever possible,
- Speculate about findings only when the data or reasoned arguments justifiy such conjectures,
- Acknowledge the pitfalls encountered,
- Write one report that will meet the needs of various audiences,
- Solicit comments on the draft report from various audiences, and
- Present oral evaluation report(s) before finalizing the written document.

Data presentation is another key issue. It is common to have many tables and figures within an evaluation report. These must be logical, legible, and understandable. Tips for creating tables and figures, and on presenting numerical data in the text are presented in Appendix V as "Guidelines for presenting data." The time necessary for creating good, clear tables, graphs, and figures is well worth the effort; it is not

uncommon for people to "look at the pictures" rather than carefully reading the results.

Some general rules include using figures to express numbers 10 and above while using words to express numbers below 10. Also, words can be used to express commonly used numbers that do not have a precise meaning (e.g., one-half). When creating tables and figures, ensure that they are clear, including a brief but self-explanatory title; it should not be necessary to read any text in order to understand tables and figures.

Audiences are a key to the evaluation report. As indicated earlier, the evaluator must determine who the audiences are and be prepared to provide information to several of them. This does not necessarily mean that separate reports will need to be written. A good Executive Summary of the evaluation report can be used for several audiences. In addition, the needs of various audiences can be met in one report. A table at the beginning of the report might "point" these audiences towards the sections that will be of greatest importance to them.

Tyler (1990) suggests that four audiences typically need four somewhat different kinds of information. Teachers and parents need student-specific information. They need to know what students learned and what is still "missing." Parent needs to know what is expected of their children while teachers need to know which students need further specialized assistance.

School principals need classroom-oriented information so they can provide assistance to teachers to ensure that the year's goals can be met. For their purposes, the evaluation report might include the percentages of students in each classroom who are meeting the goals and objectives.

School district personnel want information about schools in order to identify problems serious enough,; or opportunities great enough, to justify a considerable commitment of their time (and potentially money). For their purposes, the evaluation report should include information about the proportion of students at each site who are reading the learning objectives.

In many cases, oral reports for reach group of stakeholders can be presented. In others, brief appendices of the one major report will provide the information needed. Providing tables of information can be enough for some groups to receive the information the need. For tips on ensuring that the audiences' needs are anticipated, see that document "KEYS TO ... Reporting evaluation results to different audiences" in Appendix V.

Recommendations for improvement are one of the key sections of the evaluation report. However, evaluator(s) and program staff must be careful that the recommendations made are feasible for the program. When one or more of the objectives of an educational program have not been met, it is especially important to determine why and to make recommendations about how to proceed. Lignon and Jackson (1989) suggest that there are five different levels of recommendations that can be made:

- A mater-of-fact statement of major findings as descriptions of results.
- Findings categorized to highlight those that require action.
- A statement of the findings that require action in terms that specifically indicate the necessary action.
- A statement of the options that should be considered.
- A recommendation that a specific action be taken.

Each of these levels of recommendation are more prescriptive than the previous level(s). Again, it will be important that the evaluator and the program staff work together to ensure that the recommendations are

feasible -- programmatically, fiscally, and personnel-wise. For a more detailed version of this information, see Appendix V "Types of Evaluation Conclusions and Recommendations."

Title VII evaluation is prescribed within EDGAR and the Improving America's Schools Act of 1994. The specific "does and don'ts" have not yet been published, but in general the IASA/EDGAR statements about evaluation are more flexible than was the case under ESEA regulations.

The Title VII evaluation will be used by the program

- for program improvement,
- to further define the program's goals and objectives, and
- to determine program effectiveness

In general, programs funded under IASA will not be terminated for failure to meet the objectives, as long as good faith efforts are being made to ameliorate the situation. There are two caveats to this statement: (1) dual language programs (developmental bilingual programs, two-way programs) may be terminated if students are not learning both of the target languages (English and another language) and (2) both schoolwide and system-wide programs may be terminated if students are not being taught to and making adequate progress toward achieving the state content and performance standards. While it is unclear what constitutes "adequate progress," it probably can be assumed that program that can show that their curricula match the state standards or frameworks, and whose students are gaining in achievement levels, will not be terminated.

For further information on Title VII evaluation standards, see the document "Evaluation for IASA Title VII" in Appendix V.

Combining evaluation results can be helpful for a variety of purposes. Such information can help design a new program, provide information to support bilingual education, and lead toward the development of new theories. Some methods for integrating evaluation data are described in "Methods for integrating findings." For those contemplating such an activity, see "Suggested form for summarizing report results." Both documents are in Appendix V.

Summarizing, the evaluation report is an essential part of the evaluation process. For those who are involved with Title VII evaluations, two final documents are included in the Appendix, "IASA Title VII reporting procedures" and "Evaluation design checklist." The former describes the annual progress (formative) report and the biennial evaluation report required by that agency. In general, the guidelines provided can be used by most agencies evaluating an educational program. The latter document provides the evaluation items required by IASA and/or EDGAR; again, most of these would be appropriate for any type of evaluation. The checklist is set up to provide information about the adequacy of the reporting of each evaluation item and allows other comments to be made as well. We suggest that anyone preparing an evaluation should create such a checklist, specific to their own situation, before the report is completed. This will allow an objective determination of whether the final report meets the needs of the educational program.

(table of contents)

References

> The reference list ... provides the information necessary to identify and retrieve each source. ... Note that a reference list cites works that specifically supports a particular [program]. In contrast, a bibliography cites works for background or for further reading, and may include descriptive notes.
>
> APA (1994, 174)

Alderson, J.C., Krahnke, K.J., & Stansfield, C.W. (Eds.) (1987). Reviews of English language proficiency tests. Washington, DC: Teachers of English to Speakers of Other Languages.

American Psychological Association (1994). Publication manual (4th ed.). Washington, DC: Author.

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

Benson, J. & Michael, W.B. (1990). A twenty-year perspective on evaluation study design. In H.J. Walberg & G.D. Haertel (Eds.), The international encyclopedia of educational evaluation (pp 545-553). Oxford, England: Pergamon.

Bernharadt, V.L. (1994). The school portfolio: A comprehensive framework for school improvement. Princeton Junction, NJ: Eye on Education.

Beyer, B.K. (1995). How to conduct a formative evaluation. Alexandria, VA: Association for Supervision and Curriculum Development.

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

Charles, R., Lester, F., & O'Daffer, P. (1987). How to evaluate progress in problem solving. Reston, VA: National Council of Teachers of Mathematics.

Chronbach, L.J. (1982). Designing evaluation of educational and social programs. San Francisco: Jossey-Bass.

Chronbach, L.J., Ambrong, S.R., Dornbusch, S.M., Hess, R.D., Hornick, R.C., Phillips, D.C., Walker, D.E., & Weiner, S.S. (1980). Toward reform of program evaluation. San Francisco: Jossey-Bass.

Conoley, J.C. & Kramer, J.J. (Eds.) (1989). The tenth mental measurements yearbook.

Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska.

Curriculum Office (1994a). Language arts portfolio handbook for intermediate grades 3-5. Juneau, AK: Juneau School District.

Curriculum Office (1994b). Language arts portfolio handbook for the primary grades (3rd ed.). Juneau, AK: Juneau School District.

Del Vecchio, A., & Guerrero, M. (1995). Handbook of English language proficiency tests. Albuquerque, NM: Evaluation Assistance Center-West, New Mexico Highlands University.

Del Vecchio, A., Guerrero, M., Gustke, C., Martínez, P., Navarrete, C.J., & Wilde, J.B. (1994). Whole-school bilingual education program: Approaches for sound assessment. Program Information Guide 18. Washington, DC: NCBE.

Durán, R.P. (1990) Validity and language skills assessment: Non-English background students. In H. Wainer & H.I. Braun (Eds.), Test validity (pp. 105-128). Hillsdale, NJ: Lawrence Earlbaum.

Elementary Grades Task Force (1992). It's elementary! Sacramento, CA: CA Department of Education.

Eraut, M.R. (1990). Educational objectives. In H.J. Walberg & G.D. Haertel (Eds.), The international encyclopedia of educational evaluation (pp 171-179). Oxford, England: Pergamon

FairTest (1995). Bilingual assessment fact sheet. Cambridge, MA: National Center for Fair and Open Testing.

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

Fitz-Gibbon, C.T. & Morris, L.L. (1978b). How to design a program evaluation. Beverly Hills: Sage.

Fitz-Gibbon, C.T. & Morris, L.L. (1978c). How to measure program implementation. Beverly Hills: Sage.

Goodman, K.S., Bird, L.B., & Goodman, Y.M. (1992). The whole language catalog. Santa Rosa, CA: American School.

Guba, E.G. & Lincoln, Y.S. (1981). Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approaches. San Francisco: Jossey-Bass.

Hays, W.L. (1988). Statistics (4th ed.). New York: Holt, Rinehart and Winston.

Henerson, M.E., Morris, L.L., & Fitz-Gibbon, C.T. (1978). How to measure attitudes. Beverly Hills: Sage.

Herman, J.L. (1990). Item writing techniques. In H.J. Walberg & G.D. Haertel (Eds.), The international encyclopedia of educational evaluation (pp 355-359). Oxford, England: Pergamon

Hills, J.R. (1986). All of Hills' handy hints. Washington, DC: National Council on Measurement in Education.

Holt, D. D. (Ed.) (1994). Assessing success in family literacy projects: Alternative approaches to assessment and evaluation. McHenry, IL: Delta Systems.

Huitema, B.E. (1980). The analysis of covariance and alternatives. New York: John Wiley and Sons.

Joint Committee of the American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). Standards for educational and psychological testing, Washington, DC: American Psychological Association.

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

Kerlinger, F.N. (1986). Foundations of behavioral research (3rd ed.). New York: Holt, Rinehart, and Winston.

Keyser, D.J. & Sweetland, R.C. (Compilers.) (1994). Test critiques, Vols. I-X. Austin: Pro-Ed.

Kirk, R.E. (1982). Experimental design: Procedures for the behavioral sciences (2nd ed.). Belmont, CA: Brooks/Cole.

Levy, P. & Goldstein, H. (1984). Tests in education: A book of critical reviews.

Orlando, FL: Academic Press.

Lignon, G. & Jackson, E.E. (1989, March). Who writes this junk? Who reads evaluation reports anyway? Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Linn, R.L. (Ed.) (1989). Educational measurement (3rd Ed.). New York: American Council on Education.

Madaus, G.F., Scriven, M., & Stufflebeam, D.L. (1993). Evaluation models: Viewpoints on educational and human services evaluation. Boston: Kluwer-Nijhoff.

Mager, R.F. (1962). Preparing objectives for programmed instruction. Palo Alto, CA: Fearon.

Marzano, R.J., Pickering, D., & McTighe, J. (1993). Assessing student outcomes: Performance assessment using the dimensions of learning model. Alexandria, VA: Association for Supervision and Curriculum Development.

McConnell, B. (1982). Evaluating bilingual education using a time series design. In G.A. Forehand (Ed.), New directions for program evaluation: Applications of time series analysis to evaluation (pp 19-32). San Francisco: Jossey-Bass.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H.I. Braun (Eds.), Test validity (pp. 33-46). Hillsdale, NJ: Lawrence Earlbaum.

Millman, J. & Greene, J. (1989). The specification and development of tests of achievement and ability. In R.L. Linn (Ed.), Educational measurement (3rd Ed.) (pp 335-366). New York: American Council on Education.

Morris, L.L., Fitz-Gibbon, C., & Lindheim, E. (1987). How to measure performance and use tests. Newbury Park, CA: Sage.

Morris, L.L. & Fitz-Gibbon, C.T. (1978). How to present an evaluation report. Beverly Hills: Sage.

National Council of Teachers of Mathematics (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: Author.

National Council of Teachers of Mathematics (1992). Mathematics assessment: Alternative approaches. Reston, VA: Author.

National Evaluation Systems (1991). Bias issues in test development. Amherst, MA: author.

Navarrete, C. & Gustke, C. (1995). A guide to performance assessment for culturally and linguistically diverse students. Albuquerque, NM: Evaluation Assistance Center-West, New Mexico Highlands University.

Nevo, D. (1983). The conceptualization of educational evaluation: An analytical review of the literature. Review of Educational Research, 53, 117-128.

Nevo, D. (1990). Role of the evaluator. In H.J. Walberg & G.D. Haertel (Eds.), The international encyclopedia of educational evaluation (pp 89-91). Oxford, England: Pergamon.

Newmark, C.S. (Ed.) (1985). Major psychological assessment instruments. Boston: Allyn and Bacon.

Newmark, C.S. (Ed.) (1989). Major psychological assessment instruments, Vol. II. Boston: Allyn and Bacon.

Nitko, A.J. (1989). Designing tests that are integrated with instruction. In R.L. Linn (Ed.), Educational measurement (3rd Ed.) (pp 447-474). New York: American Council on Education.

Office of Bilingual Education and Minority Languages Affairs (1994). Improving America's schools--Challenges, opportunities, expectations. Washington, DC: author.

Parlett, M. & Hamilton, D. (1972). Evaluation as illumination: A new approach to the study of innovatory programs. (Occasional Paper 9). Edinburgh, Scotland: Center for Research in the Educational Sciences, University of Edinburgh.

Patton, M.Q. (1975). Alternative evaluation research paradigm. Grand Forks, ND: Study Group on Evaluation.

Payne, D.A. (1994). Designing educational project and program evaluations: A practical overview based on research and experience. Boston: Kluwer.

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

Popham, W.J. (1990). A twenty-year perspective on educational objectives. In H.J. Walberg & G.D. Haertel (Eds.), The international encyclopedia of educational evaluation (pp 189-195). Oxford, England: Pergamon.

Popham, W.J. & Sirotnik, K.A. (1992). Understanding statistics in education. Itasca, IL: F.E. Peacock.

Puckett, M.B. & Black, J.K. (1994). Authentic assessment of the young child: Celebrating development and learning. New York: Macmillan.

Reeves, C.A. (1987). Problem-solving techniques helpful in mathematics and science. Reston, VA: National

Council of Teachers of Mathematics.

Roeber, E.D. (1995). How should the comprehensive assessment system be designed? A. Top down? B. Bottom up? C. Both? D. Neither? Washington, DC: Council of Chief State School Officers.

Rossi, P.H. & Freeman, H.E. (1982). Evaluation: A systematic approach (2nd ed.). Beverly Hills: Sage.

Scriven, M. (1967). The methodology of evaluation. In R.E. Stake (Ed.), Curriculum evaluation. American Educational Research Association Monograph Series on Evaluation, 1. Chicago: Rand McNally.

Sharp, Q.Q. (Compiler.) (1989). Evaluation: Whole language checklists for evaluating your children. New York: Scholastic.

Stake, R.E. (1967). The countenance of educational evaluation. Teachers College Record, 68, 523-540.

Stiggins, R.J. (Ed.) (1981). A guide to published tests of writing proficiency. Portland, OR: Clearinghouse for Applied Performance Testing.

Stufflebeam, D.L., Folely, W.J., Gephart, W.J., Guba, E.G., Hammond R.L., Merriman, H.O., & Provus, M.M. (1971). Educational evaluation and decision-making. Itasca, IL: F.E. Peacock.

Tallmadge, G.K., Lam, T.C.M., and Gamel, N.N. Bilingual education evaluation system users' guide. Volume I: Recommended procedures. Mountain View, CA: RMC Research Corporation.

Test Collection, Educational Testing Service (1991). The ETS test collection catalog, vols 1-6. Phoenix, AZ: Oryx.

Thorndike, R.L. (1990). Reliability. In H.J. Walberg & G.D. Haertel (Eds.), The international encyclopedia of educational evaluation (pp 260-272). Oxford, England: Pergamon.

Tierney, R.J., Carter, M.A., & Desai, L.E. (1991). Portfolio assessment in the reading-writing classroom. Norwood, MA: Christopher Gordon.

Tyler, R. (1950). Basic principles of curriculum and instruction. Chicago: University of Chicago.

Tyler, R. (1990). Reporting evaluations of learning outcomes. In H.J. Walberg & G.D. Haertel (Eds.). The international encyclopedia of educational evaluation (pp 733-738). Oxford, England: Pergamon.

Walberg, H.J. & Haertel, G.D. (1990) (Eds.). The international encyclopedia of educational evaluation. Oxford, England: Pergamon.

Zeller, R.A. (1990). Validity. In H.J. Walberg & G.D. Haertel (Eds.), The international encyclopedia of educational evaluation (pp 189-195). Oxford, England: Pergamon.

(table of contents)

NOTE: ***Appendices are not included in the electronic version of this publication.***

---

go to HOME PAGE
Online Library
Subject Area:Assessment & Accountability