ED 417 602		FL 025 124

AUTHOR	Wilde, Judith; Sockey, Suzanne
TITLE	Evaluation Handbook.
INSTITUTION	Evaluation Assistance Center--West, Albuquerque, NM.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	1995-12-00
NOTE	213p.
CONTRACT	T003H10004-94D
PUB TYPE	Guides - Non-Classroom (055)
EDRS PRICE	MF01/PC09 Plus Postage.
DESCRIPTORS	*Bilingual Education Programs; Compliance (Legal); Cultural Awareness; Elementary Secondary Education; *Evaluation Criteria; *Evaluation Methods; Federal Programs; Institutional Evaluation; *Measurement Techniques; Objectives; Program Administration; Program Design; *Program Evaluation; Rating Scales; Scheduling; Technical Writing
IDENTIFIERS	*Improving Americas Schools Act 1994

ABSTRACT

This handbook offers extensive guidance in evaluating programs under the Improving America's Schools Act of 1994, particularly bilingual education programs under Title VII. It is divided into five sections. The first provides an overview of the handbook and some useful definitions. The second section provides background information about and definitions of evaluation, assessment, and analytic techniques. Various types of evaluation are described and guidelines for managing an evaluation are suggested; working with an external evaluator is also addressed. The third section assists in planning the evaluation: writing and modifying objectives that are measurable, creating management timelines; selecting assessments that will measure learner success in a manner that is sensitive to their language, culture, and gender as well as to program needs; and selecting scoring methods. Section four deals with implementation of an evaluation: ensuring timelines are met; training staff to assist in the evaluation; collecting data; analyzing data; and the specific requirements of the legislation. The final section deals with writing the report: interpreting analyses; presenting results; making recommendations; and writing a complete and accurate report. Substantial support materials for each section are appended. Contains 79 references. (MSE)

EVALUATION HANDBOOK

Judith Wilde, PhD
and
Suzanne Sockey, PhD

CENTENNIAL
1893 - 1993

NMHU

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

December 1995

The **Evaluation Assistance Center-West** (EAC-West) is funded by the US Department of Education's Office of Bilingual Education and Minority Languages Affairs (OBEMLA) and is operated under Grant № T003H10004-94D by New Mexico Highlands University, Center for the Education and Study of Diverse Populations. The contents of this publication do not necessarily reflect the views or policies of the Department of Education, nor does the mention of trade names, commercial products, or organizations imply endorsement by the US Government. Readers are free to duplicate and use these materials in keeping with accepted publication standards. EAC-West requests that proper credit be given in the event of reproduction.

TABLE OF CONTENTS

List of Tables and Figures

Tables

Figures

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.

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:

❶ to contribute to decisions about program development and implementation,

❷ to contribute to decisions about program continuation, expansion, or "certification,"

❸ to contribute to decisions about program modification,

❹ to obtain evidence to rally support for a program,

❺ to obtain evidence to rally opposition to a program, and

❻ 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

	Number	%
Students by grade		
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
3	Spanish-speakers	15	19.2
	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 *n*s" (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.

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✖					
Meet with evaluator/staff-discuss evaluation plan	x	xxxx	xxxx	xxxx	xxxx	xxxx
Determine feasibility of evaluation plan	x	xxx✖	xxxx	xxx✖	xxxx	xxx✖
Review objectives & assessment instruments	x	xxxx			xxxx	

x indicates approximately one week; ✖ 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 "✖" 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.

① A drop in NCEs often reflects the expected patterns. These students are behind their grade-peers, and continue to fall further behind.

② 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.

③ 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.

An activity statement is a clear description of the performance or expected behavior.

(1) State activities in specific and measurable terms.
(2) Write activities in a logical sequence.
(3) Relate activities to the program's goals and objectives.

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

① Define the term "novel."

② Identify the different types of novels.

③ Select from one type of novel.

④ Read excerpts from at least five novels.

⑤ Select one novel from those reviewed.

⑥ 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 straightforward 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 Culturally and 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: implications for 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.

D***erived 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.

P***ercentiles*** 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.

G***rade 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 Hills' 1986 monograph *All of Hills' Handy Hints* about the interpretation of widely used test score scales.

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	✖	✖	
Percent Correct	Standard of 100% correct	✖	✖	
Mastery Scores	Mastery/ non-mastery of content	✖	✖	
Grade Equivalents	Norm group		Perhaps	✖
Standard Scores, including NCEs	Norm group	✖	✖	
Stanines	Norm group			✖
Rubrics	Criterion performance	✖	✖	
Percentiles	Norm group		✖	
Gain Scores	N/A		Perhaps	

Changing test scores is possible. That is, if test scores have been recorded in 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, grade equivalents, 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 from a raw score of 11 to a stanine of 3, NCE of 30, and percentile of 17; in addition, the final columns indicate that a score of 11 is at a grade equivalent of 2.7 and an extended scale score of 430. Note that because 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 should be 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 to misinterpreted and misunderstood results for the project. The next section of this *Handbook* deals with the implementation of the planned evaluation.

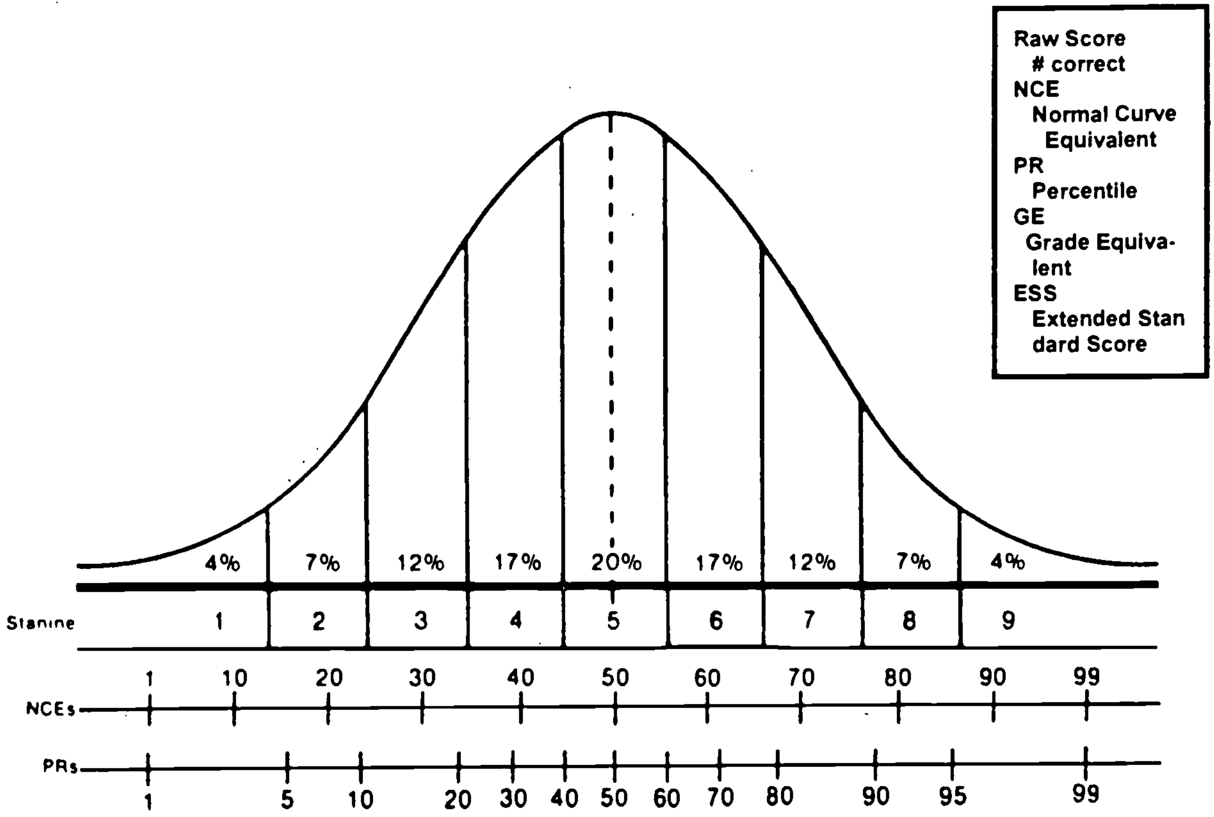
Table 4.
Conversion Table for a Standardized Test

	Stanine	Grade 4							Stanine	Grade 5							All Grades	
		October grade 4.1		February grade 4.5		May grade 4.8				October grade 5.1		February grade 5.5		May grade 5.8				
Raw Score		NCE	PR	NCE	PR	NCE	PR	Raw Score		NCE	PR	NCE	PR	NCE	PR	Raw Score	GE	ESS
1	1	—	—	—	—	—	—	1	1	—	—	And so on ...				1	-.-	342*
2		—	—	—	—	—	—	2		—	—					2	-.-	351*
3		1	1	—	—	—	—	3		—	—					3	-.-	361*
4		3	1	1	1	—	—	4		—	—					4	-.-	370
5		8	2	2	1	1	1	5		1	1					5	-.-	380
6		13	4	7	2	3	1	6		5	2					6	-.-	389
7	2	17	6	11	3	7	2	7		9	3					7	-.-	399
8		20	8	15	5	11	3	8		12	4					8	-.-	406
9	3	24	11	19	7	15	5	9	2	15	5					9	-.-	415
10		27	14	22	9	19	7	10		19	7					10	2.5	424
11		30	17	26	13	23	10	11		21	8					11	2.7	430
12		33	21	29	16	26	13	12	3	24	11					12	2.8	437
13	4	36	25	32	20	29	16	13		27	14					13	3.1	443
14		39	30	35	24	32	20	14		29	16					14	3.3	449
15		41	33	37	27	34	22	15		31	18					15	3.5	454
16		44	39	39	30	36	25	16		33	21					16	3.7	459
17	5	46	42	41	33	38	28	17	4	35	24					17	3.8	464
18		48	46	43	37	40	32	18		37	27					18	4.0	469
19		50	50	45	41	42	35	19		39	30					19	4.1	473
20		52	54	47	44	43	37	20		41	33					20	4.2	478
21		54	58	49	48	45	41	21		43	37					21	4.4	482
22	6	56	61	51	52	47	44	22	5	45	41					22	4.5	487
23		58	65	53	56	49	48	23		47	44					23	4.7	491
24		60	68	55	59	51	52	24		48	46					24	4.9	496
25		62	72	57	63	53	56	25		50	50					25	5.1	500
26		64	75	59	63	55	59	26		52	54					26	5.3	504
27	7	66	78	61	70	57	63	27		53	56					27	5.5	508
28		69	82	63	73	60	68	28		55	59					28	5.6	512
29		71	84	66	78	62	72	29	6	57	63					29	5.8	517
30		73	86	68	80	64	75	30		59	67					30	6.0	522
31	8	75	88	70	83	66	78	31		61	70					31	6.2	527
32		78	91	72	85	68	80	32		63	73					32	6.5	533
33		80	92	75	88	71	84	33	7	66	78					33	6.7	538
34		83	94	77	90	73	86	34		68	80					34	7.0	545
35		85	95	80	92	76	89	35		71	84					35	7.3	551
36	9	87	96	82	94	79	92	36		73	86					36	7.5	557
37		90	97	85	95	82	94	37		76	89					37	7.8	564
38		93	98	88	96	84	95	38	8	80	92					38	8.3	573
39		96	99	91	97	88	96	39		83	94					39	8.7	581
40		99	99	95	98	91	97	40	9	87	96					40	9.2	590
41		—	—	99	99	96	99	41		91	97					41	9.8	600
42		—	—	—	—	99	99	42		95	98					42	10.4	611
43		—	—	—	—	—	—	43		99	99					43	11.1	623
44		—	—	—	—	—	—	44		—	—					44	11.9	637
45		—	—	—	—	—	—	45		—	—					45	12.8	652

Notes: — and -.- represent extreme raw scores that were obtained by very few children and that cannot be assigned truly reliable norms. These scores should be regarded simply as lower or higher than the first or last value given in the table. * represent Extended Scale Scores that are estimates based on extrapolations.

BEST COPY AVAILABLE

Figure 2.
Relationships among Stanines, NCEs, and Percentiles



Raw Score
# correct
NCE
Normal Curve Equivalent
PR
Percentile
GE
Grade Equivalent
ESS
Extended Standard Score

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

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 appropriate; 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.

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

❷ 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.

❸ 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.)

❹ Explain the scoring system. Utilize the first 1 or 2 vignettes in demonstrating the scoring procedures to the group.

❺ Trainees should score the other vignettes on their own during the training session.

❻ Review the scores of the experts. Any vignettes that trainees score differently should be discussed in depth.

❼ 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.)

❽ 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 analyses to be utilized, and
- collection of data across 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 basic design requires that
- students be pretested and posttested with the same assessment instrument,
- appropriate scoring methods be used (preferably NCEs), and
- the same students be utilized for both pretest and posttest.

The design allows
- a comparison against the national norm (50 NCEs) or grade-mates,
- simple analyses based 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

① Subtract the project student's average pretest score from the comparison group's average pretest score -- this is the pretest gap.

② Subtract the project student's average posttest score from the comparison group's average posttest score -- this is the posttest gap.

③ Subtract the posttest gap from the pretest gap -- this is the gap-reduction.

④ 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.

⑤ 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 group, 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:

❶ to test the difference between the two groups' pretest scores (ideally, this should be nonsignificant, indicating that the two groups are similar);

❷ 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);

❸ 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

❹ 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.

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 the 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 <u>why</u> 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:

① 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;

② 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;

③ 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

④ 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,
- □ Avoid 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 justify such conjectures,
- □ Acknowledge the pitfalls encountered,
- □ Write <u>one</u> 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:

① A mater-of-fact statement of major findings as descriptions of results.

② Findings categorized to highlight those that require action.

③ A statement of the findings that require action in terms that specifically indicate the necessary action.

④ A statement of the options that should be considered.

⑤ 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.

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 Nº18. 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, Nº 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.

Appendix I

Summary of the standards for evaluation of educational programs, projects, and materials
About Evaluations
KEYS TO ... Understanding "Title VII-ese"

SUMMARY OF THE STANDARDS FOR EVALUATION OF EDUCATIONAL PROGRAMS, PROJECTS, AND MATERIALS**

A UTILITY STANDARDS

The utility standards are intended to ensure that an evaluation will serve the practical information needs of given audiences. These standards are:

A1 **Audience Identification**
Audiences involved in or affected by the evaluation should be identified, so that their needs can be addressed.

A2 **Evaluator Credibility**
The persons conducting the evaluation should be both trustworthy and competent to perform the evaluation, so that their findings achieve maximum credibility and acceptance.

A3 **Information Scope and Selection**
Information collected should be of such scope and selected in such ways as to address pertinent questions about the object of the evaluation and be responsive to the needs of interests of specified audiences.

A4 **Valuational Interpretation**
The perspectives, procedures, and rationale used to interpret the findings should be carefully described, so that the bases for value judgments are clear.

A5 **Report Clarity**
The evaluation report should describe the object being evaluated and its context, and the purposes, procedures, and findings of the evaluation, so that the audiences will readily understand what was done, why it was done, what information was obtained, what conclusions were drawn, and what recommendations were made.

A6 **Report Dissemination**
Evaluation findings should be disseminated to clients and other right-to-know audiences, so that they can assess and use the findings.

A7 **Report Timeliness**
Release of reports should be timely, so that audiences can best use the reported information.

A8 **Evaluation Impact**
Evaluations should be planned and conducted in ways that encourage follow-through by members of the audiences.

B FEASIBILITY STANDARDS

The feasibility standards are intended to ensure that an evaluation will be realistic, prudent, diplomatic, and frugal; they are:

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

BI **Practical Procedures**
The evaluation procedures should be practical, so that disruption is kept to a minimum and that needed information can be obtained.

B2 **Political Viability**
The evaluation should be planned and conducted with anticipation of the different positions of various interest groups, so that their cooperation may be obtained and so that possible attempts by any of these groups to curtail evaluation operations or to bias or misapply the results can be averted or counteracted.

B3 **Cost Effectiveness**
The evaluation should produce information of sufficient value to justify the resources expended.

C PROPRIETY STANDARDS

The propriety standards are intended are intended to ensure that an evaluation will be conducted legally, ethically, and with due respect for the welfare of those involved in the evaluation, as well as those affected by its results. These standards are:

C1 **Formal Obligation**
Obligations of the formal parties to an evaluation (what is to be done, how, by whom, when) should be agreed to in writing, so that these parties are obligated to adhere to all conditions of the agreement or formally to renegotiate it.

C2 **Conflict of Interest**
Conflict of interest, frequently unavoidable, should be dealt with openly and honestly, so that it does not compromise the evaluation process and results.

C3 **Full and Frank Disclosure**
Oral and written evaluation reports should be open, direct, and honest in their disclosure of pertinent findings, including the limitations of the evaluation.

C4 **Public's Right to Know**
The formal parties to an evaluation should respect and assure the public's right to know, within the limits of other related principles and statutes, such as those dealing with public safety and the right to privacy.

C5 **Rights of Human Subjects**
Evaluations should be designed and conducted so that the rights and welfare of the human subjects are respected and protected.

C6 **Human Interactions**
Evaluators should respect human dignity and worth in their interactions with other persons associated with an evaluation.

C7 **Balanced Reporting**
The evaluation should be complete and fair in its presentation of strengths and weaknesses of the object under investigation, so that strengths can be built upon and problem areas addressed.

C8 **Fiscal Responsibility**
The evaluator's allocation and expenditure of resources should reflect sound accountability procedures and otherwise be prudent and ethically responsible.

D ACCURACY STANDARDS

The accuracy standards are intended to ensure that an evaluation will reveal and convey technically adequate information about the features of the object being studied that determine its worth or merit. These standards are:

D1 **Object Identification**
The object of the evaluation (program, project, material) should be sufficiently examined, so that the form(s) of the object being considered in the evaluation can be clearly identified.

D2 **Context Analysis**
The context in which the program, project, or material exists should be examined in enough detail so that its likely influences on the object can be identified.

D3 **Described Purposes and Procedures**
The purposes and procedures of the evaluation should be monitored and described in enough detail so that they can be identified and assessed.

D4 **Defensible Information Sources**
The sources of information should be described in enough detail so that the adequacy of the information can be assessed.

D5 **Valid Measurement**
The information-gathering instruments and procedures should be chosen or developed and then implemented in ways that will assure that the interpretation arrived at is valid for the given use.

D6 **Reliable Measurement**
The information-gathering instruments and procedures should be chosen or developed and then implemented in ways that will assure that the information obtained is sufficiently reliable for the intended use.

D7 **Systematic Data Control**
The data collected, processed, and reported in an evaluation should be reviewed and corrected, so that the results of the evaluation will not be flawed.

D8 **Analysis of Quantitative Information**
Quantitative information in an evaluation should be appropriately and systematically analyzed to ensure supportable interpretations.

D9 **Analysis of Qualitative Information**
Qualitative information in an evaluation should be appropriately and systematically analyzed to ensure supportable interpretations.

D10 **Justified Conclusions**
The conclusions reached in an evaluation should be explicitly justified, so that the audiences can assess them.

D11 **Objective Reporting**
The evaluation procedures should provide safeguards to protect the evaluation findings and reports against distortion by the personal feelings and biases of any party to the evaluation.

About Evaluations

A sound evaluation can

1. provide a rich source of information for teaching and guiding students' learning,
2. assist in monitoring programs,
3. assist in evaluating program effectiveness,
4. act as a source of student motivation,
5. contribute to student improvement, and
6. meet federal and state requirements if implemented appropriately.

A poor evaluation can

1. misrepresent what a student can do,
2. measure something other than what a student learns in the classroom,
3. measure only rote recall of facts and figures, not in-depth thinking,
4. lead to misinterpreting student performance, and
5. not be useful as a "sound evaluation" as defined above.

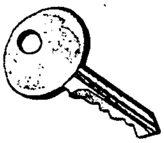
KEYS TO ... Understanding "Title VII-ese"

We have had several requests recently to explain what we mean by some of the jargon typically used in Title VII and by the US Department of Education. We hope we have included all the terms that you find troublesome. If not, please let us know so we can add to the list.

In order to understand Title VII-ese, you first must understand the organizational terms, and the organizations, with which we work. These include:

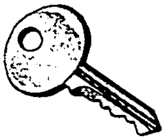
- SEA State Education Agency; not a particular state, but any department of education at a statewide level
- IHE Institution of Higher Education; not a specific site, but any college or university
- LEA Local Education Agency; not a specific site, but any local school district, part of a local school district, or a consortium of school districts
- CBO Community Based Organization
- NPO Non Profit Organization

In addition, there are several organizations related to the US Department of Education. These organizations are national or regional in scope; all are funded by DoE or are part of the DoE.



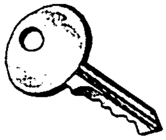
- TAC Technical Assistance Center; at the time that we write this, there are many TACs across the country. The Evaluation Assistance Centers (EACs) and Multi-functional Resource Centers (MRCs) provide assistance to Title VII projects. Funding will end March 1996
- CC Comprehensive Regional Assistance Centers; also referred to as C-TACs. Provide technical assistance to all federally-funded education programs; there are 15 regional CCs. Full funding should begin April 1996
- NCBE National Clearinghouse for Bilingual Education; provides information (monographs, articles, etc.) on the educational needs of limited-English proficient students.
- CAL Center for Applied Linguistics
- NCES National Center for Educational Statistics, part of OERI
- OERI Office of Educational and Research Improvement
- OBEMLA Office of Bilingual Education and Minority Languages Affairs
- OMB Office of Management and Budget
- OCR Office of Civil Rights

There are several programs funded (by grants, funding formulae, or other means) through IASA. These are defined below. In addition EDGAR is the *Education Department General Administrative Regulations*. EDGAR contains the "rules" for grant applications and evaluations.



- Title I Programs for students living in poverty.
- Migrant Title I funding for the education of migrant students
- Even Start Title I funding for preschool students and their families living in poverty
- N&D Title I funding for Neglected and Delinquent students
- Title VII Programs for limited-English proficient students
- Title IX Programs for American Indian students (including native Alaskans and native Hawaiians)
- Title IV Safe and Drug Free Schools programs

Finally, some terms that are commonly used which may need some definition to understand their special meaning within the bilingual education context.



- NEP Non-English Proficient; students whose English skills are minimal or who do not speak English at all
- LEP Limited English Proficient; refers to students whose English skills do not allow them to communicate or learn effectively in English-only classrooms
- FEP Fluent English Proficient; refers to students who can communicate effectively in English; previously LEP students
- EO English Only; monolingual English speakers

Modified from **KEYS TO ... Understanding "Title VIII-ese"** *in* *EAC-West news, December 1990 by the Evaluation Assistance Center-West/NMHU, Albuquerque, NM* *Revised 11/95*

Appendix II

Conceptualizing Educational Evaluation

Recent decades have produced many article on the conceptualization of educational evaluation. Most evaluation designs respond to ten questions, or issues. D. Nevo (1983) in "The conceptualization of educational evaluation: An analytical review of the literature" (*Review of Educational Research, 53*, pages 117-128) summarized the evaluation literature. His research can be modified to provide the most common answers to ten evaluation questions.

✧ How is evaluation defined?

Educational evaluation is a systematic description of educational objects and/or an assessment of their merit or worth.

✧ What are the functions of evaluation?

Educational evaluation can serve four different functions: (1) formative (for improvement); (2) summative (for selection of "best" and accountability); (3) sociopolitical (to motivate and gain pubic support); and (4) administrative (to exercise authority).

✧ What are the objects of evaluation?

Any entity can be an evaluation object. Typical evaluation objects in education are students, educational and administrative personnel, curricula, instructional materials, programs, projects, and institutions. Within IASA, educational programs are evaluated to determine their worth in improving the educational status of students at risk of educational failure. Given this definition, the term "program" will be substituted for "object" in the rest of this document.

✧ What kinds of information should be collected regarding each program?

Four groups of variables should be considered regarding each educational program. They focus on (1) the goals of the program; (2) its strategies and plans; (3) its process of implementation; and (4) its outcomes and impacts regarding students, staff, families, and others involved in the educational program.

✧ What criteria should be used to judge the merit of a program?

The following criteria should be considered in judging the merit or worth of an educational program: (1) responding to identified needs of actual and potential clients (students, staff, and/or families); (2) achieving national goals (e.g., *Goals 2000*), ideals, or social values; (3) meeting agreed-upon standards and norms (e.g., high content and performance standards); (4) outdoing alternative educational programs; and (5) achieving (important) stated goals of the object. Multiple criteria (assessments, viewpoints) should be used for any educational program.

✧ Who should be served by an evaluation?

Evaluation should serve the information needs of all actual and potential parties interested in the program being evaluated ("stakeholders"). It is the responsibility of the evaluator(s) to delineate the stakeholders of an evaluation and to identify or project their information needs.

✧ What is the process of doing an evaluation?

Regardless of its method of inquiry (or evaluation design), an evaluation process should include the following three activities: (1) focusing the evaluation problem; (2) collecting and analyzing empirical data; and (3) communicating findings to evaluation audiences. There is more than one appropriate sequence for implementing these activities, and any such sequence can (and sometimes should) be repeated several times during the life span of an evaluation study.

✧ What method of inquiry should be used in evaluation?

Being a complex task, evaluation needs to mobilize many alternative methods of inquiry from the behavioral sciences and related fields of study and to utilize them according to the nature of a specific evaluation problem (e.g., combine naturalistic and object-oriented designs). At the present state of the art, a preference for any specific method of inquiry, if stated before the evaluation is begun, is not warranted.

✧ Who should do evaluation?

Evaluation should be conducted by individuals or teams possessing (1) extensive competencies in research methodology and other data analysis techniques; (2) understanding of the social context and the unique substance of the educational program being evaluated (e.g., bilingual education, education for linguistically and culturally diverse students, ESL methodologies, educational programs for those at risk of education failure -- those living in poverty, who are delinquent or neglected); (3) the ability to maintain correct human relations and to develop rapport with individuals and groups involved in the evaluation; and (4) a conceptual framework to integrate the above-mentioned capabilities.

✧ By what standards should evaluation be judged?

Evaluation should strike for an optimal balance in meeting standards of (1) utility (to be useful and practical); (2) accuracy (to be technically adequate); (3) feasibility (to be realistic and prudent); and (4) propriety (to be conducted legally and ethically).

CURRENT FRAMEWORKS FOR PROGRAM EVALUATION

The purpose of this document is to define various program evaluation models. The models can (1) stimulate thinking, (2) provide sources of new ideas and techniques, and (3) serve as a mental checklist for conducting an evaluation. These models are not prescriptive guidelines but should be considered as approaches that might be used to focus a program evaluation. Also, these frameworks describe processes for thinking about evaluation, they do not describe or limit the statistical approaches that can be used to analyze the data collected for evaluative purposes.

Objectives-Oriented

This approach is designed as a process for determining the extent to which the educational objectives of a program actually are accomplished. The objectives-oriented approach was originally formulated by Ralph Tyler, who proposed that goals or objectives should be established or identified and defined in behavioral, measurable terms relevant to participant (e.g., students) behaviors, using standardized norm-referenced tests or other types of instruments. The outcome data collected from the measures then are compared with the behavioral objectives to determine the extent to which participant performance matched the stated expectations. Discrepancies between the actual, observed performance and the stated objectives are defined as objectives; modifications in the objectives would be necessary to correct the deficiencies. Examples of objectives-oriented approaches include Provus' discrepancy model, Popham's instructional objectives approach, and Hammond's evaluation approach.

Strengths

- Easily understood; easy to follow and implement
- Produces information generally relevant to educators' mission
- Extensive literature available on applications to classrooms and other educational settings
- Assists in clarifying ambiguous generalities about educational outcomes
- Has been used extensively to improve test development

Limitations

- Does not measure the merit or worth of a program
- Lacks standards to judge the importance of discrepancies between objectives and performance levels
- Ignores outcomes not anticipated through objectives
- Neglects transactions that occur within the program or activity being evaluated
- Training needed to ensure that objectives are written in behavioral terms and are appropriate for curriculum

Decision-Management

The most important contributions to this approach come from Stufflebeam and Adkin who drew their work from management theory. Both contend that pivotal evaluation decisions are made by the program manager and that program objectives are not the primary concern of the program or the evaluation. An evaluator, working closely with the program manager, identifies the decisions the latter makes and collects sufficient information about the advantages and disadvantages of each decision. The information is used by the program manager to make a fair judgement. The evaluation becomes an explicitly shared function based on dependent teamwork between the evaluator and the program manager.

Examples of the management-oriented approach are the Context-Input-Process-Product (CIPP) evaluation model and the UCLA evaluation model.

Strengths
- Focuses on information needs and pending decisions to be made by a program manager
- Stresses the utility of information
- Useful for shaping an evaluation in reference to actual decision-making considerations
- Preferred choice of many administrators and boards
- Generates potentially important questions to be addressed in the evaluation

Limitations
- Evaluators occasionally are unable to respond to questions or issues that may be significant
- Can be costly and lead to complex evaluation
- Assumes that important decisions can be made in advance
- Frequent adjustment may need to be made in the original plan if this approach is to work well

Judgement-Oriented

This is the oldest and most widely used approach to evaluation. The judgement-oriented approach is dependent on the professional expertise of a program manager in making judgement about a program, product, or activity. Some examples of this approach include formal and informal professional review systems as well as special panel and individual reviews. Samples of the approach include Goal-free evaluation, Stakes's countenance evaluation, Ad hoc reviews, and Eisner's educational connoisseurship.

Strengths
- Emphasizes expert judgement and human wisdom in the evaluative process
- Focuses attention on standards that should be used in rendering judgements about educational programs
- Can be cost efficient
- Translates educated observations into statements about education quality

Limitations
- Can permit judgements that reflect personal bias
- Presumes decision-maker's expertise
- Highly dependent on inter-judge and inter-panel reliability
- Demands for objectivity are more rigorous in the evaluation of public programs

Adversarial

This approach is made up from a collection of evaluation practices which contain an adversarial component. The term "adversarial" refers to all evaluations in which a planned effort is made to incorporate opposing (pro and con) views within a single evaluation. Deficiencies as well a strengths are represented to assure fairness and balance in the evaluation. Samples of this approach include judicial and congressional hearings and debate models.

Strengths
- Opposing viewpoints illuminate positive and negative aspects of evaluation
- Broad range of information collected
- Satisfies informational needs of audiences in an interesting and informative manner

- Can be combined readily with other approaches
- Diffuses opposition because both sides are represented
- Clarifies issues

Limitations

- Requires rigorous planning
- Useful only for solving problems, not for making program improvement or measuring progress
- Strong pro and con positioning may not allow for the full range of information needed
- Does not eliminate bias but balances and publicizes bias
- Time consuming and expensive
- Not sufficiently developed to serve as a standard or model for educational evaluations

Pluralistic-Intuition

This approach is used by managers who incorporate the values and perspectives of all individuals and groups into the decision-making process. Judgements are weighed and balanced largely in an intuitive manner. There is little, if any, preoccupation with stating and classifying objectives, designing elaborate evaluation designs, instrumentation, and/or preparing long technical reports. Types of this approach include the countenance model, illuminative model, and democratic evaluation.

Strengths

- Can be used by any sensitive individual
- Flexible and rich with information
- Emphasizes the human element in evaluation
- Incorporates multiple viewpoints and multiple data techniques

Limitations

- Loose and unsubstantiated evaluations
- Time consuming, labor intensive, and costly--especially for large school institutions
- Lacks clear approaches on how to weigh or combine individual standards into overall judgements
- Reliance on open-ended techniques makes the evaluator a potential problem

Note

Many funding agencies require specific types of evaluations or types of data. For instance, federally-funded IASA programs require that behaviorally-based objectives be defined and evaluated. Be sure that the chosen framework meets the needs of the program, and of the funding agency.

USES FOR EVALUATION DATA

Evaluation can be used both for (1) accountability purposes, to show year-end achievement gains, and (2) utility, to inform about the instructional program's strengths and weaknesses. The data collected throughout the year can be used for these purposes. Below are some definitions which may assist in the development of an evaluation system.

Assessment *vs* Evaluation

While there is some overlap between assessment and evaluation, and the terms often are used interchangeably, it is easiest to talk about them separately. ***Assessment*** is the use of various written and oral measures and tests to determine the progress of students toward reaching the program goals. To be informative, assessment must be done in a systematic manner; consistency within measures (from one assessment period to the next with the same instrument) and across measures (similar results achieved with different instruments) must be ensured. ***Evaluation*** is the summarization and presentation of the assessment results for the purpose of determining the overall effectiveness of the program. Evaluation may help inform about modifications that need to be made in the program or may present an end-of-project picture of the program.

Formative Evaluation

Formative evaluation is conducted at various points during the planning and early stages of the program. A good formative evaluation can identify problems and suggest ways to modify the program for improvement. The formative evaluation can be in written form, or might take the form of oral reports to specific stakeholder groups (parents, teachers, administrators).

Dynamic Evaluation

Although similar to formative evaluation, dynamic evaluation is more on-going in nature. Dynamic evaluation suggests that teachers and administrators should look at various measures on a continuous basis. By "keeping tabs" on the progress of students, the instructional program can be modified immediately to meet the current needs of participants.

Summative Evaluation

Summative evaluation generally requires a more formal year-end or end-of-project written report. The assessment results from across the life of the program are presented, analyzed, summarized, and interpreted. The report addresses the strengths and weaknesses of the program as well as suggesting ways in which the program might be improved further. The summative evaluation frequently is written for accountability purposes and must meet the requirements of a funding agency or an oversight agency.

Responsibility for Evaluation

The services of a professional evaluator may be needed to ensure a well-designed evaluation utilizing valid and reliable assessment instruments. While many portions of the evaluation can be performed by the program staff, it usually is best to have a more professional, objective view of the program for the summative evaluation. Frequently, a team approach involving teachers, administrators, and professional evaluator(s) is used to provide in-depth information and to encourage staff "buy-in" to the evaluation process. Regardless of whether or not a professional evaluator is used, the program director is responsible for ensuring a reliable and valid evaluation report, completed in a timely manner.

Guidelines for Managing the Evaluation Plan

Maintaining quality control is essential to any evaluation plan. Provided below are activities to ensure that the evaluation is the best possible. Remember, too, that even if a professional evaluator has been hired, the responsibility of managing the evaluation plan is that of the program director.

1. Assess the adequacy of the evaluation design

Make sure that the design of the evaluation is valid, reliable, credible, and realistic before the start of the program. One approach for ensuring adequacy is to design a set of standards by which to review the evaluation design -- these standards would be based on the features of the evaluation that are most important to this site, for this program. Another approach is to follow the criteria developed by the Joint Committee for Standards for Educational Evaluation. This Joint Committee offers a comprehensive framework for developing standards in defining, designing, administering, collecting, analyzing, budgeting, contracting, reporting, and staffing an evaluation.

2. Monitor the practice of the evaluation design

Every good evaluation plan specifies evaluation activities that should be monitored to ascertain that the original design is implemented faithfully. Strategies to follow in monitoring evaluation practices include:

- Develop time frames to mark the milestones or dates on which products must be delivered and/or major activities must be concluded.

- Interview and observe key personnel to determine whether project activities conform to the approved evaluation plan.

- Ensure that the data collection efforts are carried out as planned by creating information checks. Train staff on proper test/assessment administration and data collection procedures. Create filing systems in which to store information as it is collected -- train staff to utilize these as well. Systematically check all data gathering activities.

3. Revise the evaluation design as needed

Unanticipated circumstances in a project's activities, or in the general school context, may require changes in an evaluation plan. Arrangements should be made for periodic examination of the original evaluation plan and for modifications as necessary. When making changes,

- Contact your project officer to ensure that the changes are approved by OBEMLA,

- Update key personnel, including the evaluator, regarding the approved changes in activities and timeline.

- Document the changes and include them in the annual performance report.

FINDING AN EVALUATOR

One of the most important activities of the program director is the selection of an evaluator. EAC-West suggests here guidelines for selecting an evaluator and for contract negotiations.

Types of Evaluators

- *Internal evaluators* -- selected from district or program staff; familiar with personnel, district, and program policies and procedures; easier access to district data bases. Considered less credible because of their possible connection with the project.
- *External evaluators* -- selected from district consultants or others outside the educational institution; hired as the result of a bidding process. Considered more objective because they have less program affiliation.
- *Team approach* -- utilizes both internal and external evaluators and/or both qualitative and quantitative evaluators, combining the strengths of each and providing a good way to evaluate the project within a limited budget.

Selecting an Evaluator

The best method for finding an evaluator is to ask other program directors whom they might recommend; to contact the district's evaluation, planning, or research offices to determine who their consultants are; or to advertise in local and regional newspapers or journals. Overall, the skills you need can be described in three general areas. You may want to develop a checklist that defines the skills needed for this project.

1. *A broad technical background* in program evaluation, research, statistics, and computers. This should include experience in designing and implementing various data collection techniques and data analysis procedures. Credibility in the field can be shown through other evaluations, written reports, and presentations at conferences.
2. *Experience in managing program evaluations* including the ability to deal with, and communicate clearly in writing and orally to, different types of people. A demonstrated ability to work with district personnel and to follow timelines and budgets is essential.
3. *Experience specific to this program*. In the case of a Title VII program, experience with Title VII, ESL programs, or other background with LEP students. The evaluator must have knowledge of federal and state evaluation requirements for Title VII programs and knowledge of tests appropriate for program evaluation purposes, including commercially produced and locally developed tests (including alternative assessments) for English language proficiency, native languages, and other subject areas. Also, the evaluator should understand the process involved in language acquisition and related implications for assessing student growth.

When considering a particular evaluator, ask for references and check them. Also, ask the evaluator to provide copies of evaluation reports s/he has written in the past, particularly those dealing with similar projects.

Contract Information

The cost of an evaluation depends on the scope of the program and of the expected evaluation. With limited monies, expect less from the evaluator and more from the project director and staff; for instance, staff can collect data and the evaluator can analyze data. Depending on the size of the project and the budget, consider the following:

- amount of data to be collected, who will collect it, and when it will be collected;
- data entry costs and who is responsible for data entry;
- number of meetings requiring the evaluator's presence;
- program documentation necessary for the evaluator to review; and
- the evaluation report(s) -- oral and written, formative and summative.

In addition, include who has final editing privileges for written reports and who will disseminate reports (usually this should be the program director). A payment schedule should be included with the contract. The last payment should be made after, or concommitant to, the delivery of the final, revised evaluation report.

ROLE OF PROJECT DIRECTOR IN EVALUATION

Ultimately, the project director is responsible for all phases of the educational program, including the evaluation. Some of the activities in which the project director should be involved include those listed below.

- Hiring the evaluator early in the project year. This can be done as early as during the application process. (Some evaluators will assist in proposal writing with the understanding that they will work on the evaluation.)

- Identifying program staff who will be involved in the evaluation process. Staff should be involved as much as possible to ensure their buy-in to the procedures.

- Negotiating the evaluator's contract--specifying tasks and responsibilities of evaluator and program staff during the evaluation.

- Creating a timeline for the whole project and assuring that the evaluation schedule is met.

- Providing the evaluator with the funding agency's regulations on evaluation, the funded proposal, and any approved amendments to the proposal.

- Ensuring that the evaluator has appropriate student outcome and project implementation data, including documentation about all components of the project (e.g., the instructional services, materials development, staff development, and parent development).

- Ensuring that project staff are trained in keeping accurate project records and that the evaluator has reasonable access to project staff, teachers, students, parents, and others whose insights will be useful to the evaluation.

- Ensuring the usefulness of the evaluation for project improvement and for planning. This includes requesting formative (on-going, periodic [oral] mini-reports) as well as summative (final, year-end report) evaluation.

- Approving the annual reports and determining to whom and how the report(s) are disseminated.

- Presenting a summary of the annual report to local administrators and, if appropriate, to the local media.

ROLE OF STAFF IN EVALUATION

The role of project staff will vary from site to site depending upon the goals and objectives of that particular educational program. Provided below is a list of suggested activities on which staff can collaborate with project directors and evaluators in order to provide a sound and useful evaluation. Review each item below and decide to what extent staff members can participate in your program evaluation. This list is not exhaustive, each program may determine other activities with which staff should be involved.

Student Outcomes

- Maintain information on project students in terms of tests or assessments (e.g., academic achievement, language proficiency, attitude) and the scores they receive on each (i.e., number correct, NCEs, standard scores).
- Record students' academic achievement by subject matter where appropriate.
- Document students' participation in programs such as gifted and talented programs, special education, and so on; document students' absenteeism, grade retention, and drop-out from the program or the school.
- Collect information on the academic progress of students who formerly participated in the program and are now in English-language classrooms.

Assessment

- Participate in modifying objectives to match student performance more closely.
- Coordinate with program directors in developing clear and observable criteria for alternative assessment instruments.
- Review test items to ensure a close link with program objectives.
- Assist in the development of alternative assessment to be used in the program.
- Maintain agreement between raters or judges when more than one person is involved in scoring student performance.
- Keep records of the tests, assessments, and observations as well as the dates they were completed.

Program Implementation

- Record educational background, needs, and competencies of project students.
- Report the amount of time in years or school months that students participate In the program.
- Provide a brief description of the specific educational activities of the program.
- List the instructional activities and accompanying educational materials and instructional methods and techniques.

- Report the staff's educational and professional qualifications, including language competencies acquired each year.

Evaluation Report

- Meet with evaluator regarding the program--complete interview(s), survey instrument(s), etc.
- Prepare draft materials describing the portions of the program with which staff is familiar.
- Allow evaluator access to classroom for observation purposes.
- Review and discuss draft evaluation document(s).

Remember:
Staff participation = staff buy-in = a more successful program!

ROLE OF THE EVALUATOR IN A TITLE VII PROJECT

Obviously, the evaluator's primary role is to evaluate the Title VII project in a manner that is helpful to the project and meets the funding agency's regulations for evaluation (e.g., the IASA and EDGAR regulations). In addition to writing the evaluation report, the evaluator **must** become thoroughly familiar with the project's proposal, including characteristics of the persons served, project design, staffing, materials, methods, and process and product objectives. In addition, s/he generally

- assists in the selection or development of assessment instruments, develops the interview protocols and other questionnaires, and may collect data;
- refines the evaluation design by observing project operations, reviewing relevant documents, and determining the nonproject comparison group (if one is needed);
- analyzes data and interprets findings, creates tables and figures, and provides information to the project staff; and
- creates various reports--these may be informal, oral reports on a periodic basis as well as the annual reports and evaluations based on data analysis and work with project staff.

Other activities may be included in the evaluator's responsibilities, depending upon the specific needs of the project. For instance, an evaluator may

- conduct the needs assessment for the project's grant application,
- be involved in proposal writing, and/or
- progress reports to ensure the continued funding of the program.

WORKING WITH YOUR EVALUATOR ON THE FINAL REPORT

Your evaluator will have expertise in a variety of issues related to educational evaluation in general and your program in particular. The project director also has expertise in a variety of issues related to the program, the school, children in the school, and the program and school staff. Finally, the staff also has knowledge and expertise in issues that are important to the evaluation; this also will make them stakeholders in the results of the evaluation.

When it comes to the evaluation report, the **evaluator** should have expertise in

- the creation of alternative assessments,
- data collection and analysis,
- interpreting and explaining the data, and
- creating graphs and tables.

The **project director** has expertise in

- the history of the project,
- pedagogical materials and techniques utilized,
- how/why standardized tests were selected, and
- what has happened in the project in the past.

Staff can be involved by writing, or assisting with, sections on

- classroom implementation,
- in-service training,
- parent involvement, and
- anecdotal evidence to support the findings.

Consider all these factors when determining the content of the evaluation report. Who should write what portions of the report? The factors about which project director or staff are experts should be written by them; the factors about which the evaluator is an expert should be written by him/her. This will result in a more complete and accurate report. Also, if the evaluator is not responsible for all of the report, s/he may be able to spend more time working with you on another facet of the evaluation.

For questions about doing the evaluation, and what should go into the report, contact the funding agency and/or local technical assistance agencies (e.g., the 15 Regional Comprehensive Centers for IASA-funded education programs).

Appendix III

MANAGEMENT TIME SCHEDULE FOR EVALUATION

Year: 19____ Total years of project: _____; this is year ① ② ③ ④ ⑤

Management Tasks ▾ Months ▸	1	2	3	4	5	6	7	8	9	10	11	12
Planning												
Determine need for evaluator; hire evaluator												
Review evaluation plan with evaluator												
Meet with evaluator & staff re: evaluation plan												
Assign staff tasks re: evaluation												
Determine current feasibility of evaluation plan												
Review objectives of project; modify if needed												
Select available assessment instruments												
Create needed assessment instruments												
Create staff observation/interview forms												
Ensure funding agency requirements are met												
Approve evaluation plan												
Other:												
Data Collection												
Schedule evaluator meetings with staff												
Train staff in data collection methods												
Monitor administration of high-stakes assessments (NRTs or locally-developed)												
Schedule evaluator to observe classrooms												
Monitor collection of non-test data												

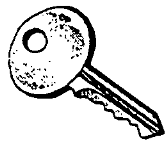
Created by the Evaluation Assistance Center-West/NMHU, Albuquerque, NM (Revised 9/95)

Management Tasks ▾ Months ▸	1	2	3	4	5	6	7	8	9	10	11	12
Monitor collection of other outcome assessments (attitudes, achievement)												
Supervise completion of data collection forms												
Approve completed data collection forms												
Contact technical assistance agencies if needed (local, state, Comprehensive Centers)												
Other:												
Analysis												
Review analyses with evaluator												
Monitor adjustments to analyses, as needed												
Review tables, figures with evaluator												
Approve analyses												
Other:												
Reporting												
Determine number, type of reports needed												
Review formative reports (oral or written)												
Supervise writing of draft summative report; ensure requirements are addressed												
Schedule meeting with evaluator & staff to review summative report recommendations												
Approve summative report												
Disseminate report												
Other:												

Created by the Evaluation Assistance Center-West/NMHU, Albuquerque, NM (Revised 9/95)

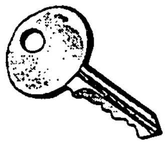
KEYS TO ... Planning Data Management

The data collection process is considered by many evaluators to be the "heart" of evaluation. During this phase of the project, information critical to the evaluation of the program is collected. To have "healthy" data at the end of the collection effort, a sound management plan is required. Some logistical questions to consider in managing data collection are listed below. .



- Who will be responsible for collecting the data? Is training needed for the person(s) to carry out the interviews, observations, ratings, testing, or other data collection procedures? Is a special incentive required to ensure full partake-pat of the data collector(s) or of those from whom data will be collected?
- When are data to be collected? When and how are evaluation instruments to be delivered and returned? What timelines or schedules are to be followed?
- Where will observations be made, interviews conducted, and tests administered?

An aspect of the data management plan that is troublesome in many evaluations is the actual handling of data. The following four steps will help ensure that the data re collected and maintain din an orderly fashion.



- Set up a filing and organization system for the information at the beginning of the data collection process (not half-way through);
- Safeguard the information from loss, premature release, or inappropriate use;
- Make sure quantitative and qualitative information is recorded accurately; and
- Store the raw data in a safe place for at least three years (the American Psychological Association recommends five years). Follow-up studies, reanalyses, or questions about the evaluation may require access to the raw data at a later date.

To maintain quality control of the data collection activities, be sure to

- Follow appropriate customs, good manners, procedures, and protocols;
- Attain needed clearances, permissions, and releases;
- Monitor for consistency and good practices in the distribution, administration, and return of the instruments, as well as in the recording of data;
- Modify the data collection plan when needed; and
- Use appropriate aggregation and reduction techniques when summarizing the data.

For information on the actual data to be collected, see the materials sent from the appropriate funding agency. While there are some general rules to be followed, each agency will have slightly different data collection requirements.

MANAGEMENT PLAN

This Management Plan is modified from work by Educational System Planning, Woodland, CA. The Project Director and the Evaluator work cooperatively to determine the completion date and person responsible for each task. Staff is involved in decision-making as appropriate.

Activities & Procedures	Completion Date	Person Responsible
PLANNING		
1. Arrange for external evaluator; establish an evaluation team: Director, Resource Teachers, Evaluator, & Asst. Superintendent.		
2. Develop an evaluation management plan including questions, activities, instruments, team, responsibilities, & schedule.		
3. Establish an evaluation team consisting of the project director, evaluator, resource teacher, & assistant superintendent of curriculum.		
4. Determine evaluation questions to be answered through the evaluation. The questions should be relevant to the program design addressing tasks, approaches, and special features of the program.		
5. Review & revise as necessary the evaluation design of the program. Consideration should be given to testing, instruments, sampling procedures, & the collection and reporting of data.		
6. Determine instruments to be used in the evaluation for: assessments, testing, data collection, & program monitoring.		
7. Develop a uniform set of criteria & procedures for identifying LEP students, assessing language proficiency, selection of program participants, & enrolling students in the appropriate bilingual program. Similarly, uniform procedures should be developed for moving the students through different levels of the program & for exiting students when they achieve Fluent English Proficiency.		
8. Establish an evaluation monitoring process with associated instruments, roles, & responsibilities. Identify members of the monitoring team.		
9. Determine the schedule for evaluator visits as specified below: ✦ Program monitoring, ✦ Interviews with staff & parents, and ✦ Review/collection of data.		
10. Conduct a training session/orientation on the evaluation design, instruments, & procedures for project staff & key personnel.		
11. Conduct an orientation for district administrators and key personnel on the evaluation design & data collection requirements & responsibilities.		

Activities & Procedures	Completion Date	Person Responsible
TESTING AND ASSESSMENTS: IDENTIFICATION & ENROLLMENT		
1. Establish a standardized test database on all LEP students to be served or currently being served by the program. Test data for NRTs should be in NCEs, scaled scores, or % correct. Test data for alternative assessments should be raw scores or % correct. All test data should be organized by grade level and/or target site. Scores should be posted in each student file allowing for longitudinal study of the student during the course of the program. Test scores should be collected on all students receiving services for 100 days or more. Record pretest and posttest dates and scores on all students enrolling late or leaving early in the program.		
2. Collect NRT and/or alternative assessment data in similar scores for all nonLep students served by the program and FEP students who have been reclassified.		
3. Collect NRT and/or alternative assessment data in similar scores for all mainstream students (nonLEP peers at the target sites) to serve as a comparison group.		
4. Collect district proficiency data on all LEP students, nonLEP students, & FEP students as appropriate. Data should be summarized by grade level. Summarize according to the number of proficiencies administered & the percentage pass/fail.		
5. Develop a summary of language assessment results using the appropriate instrument (SOLOM, IPT, LAS, BSM, other ...). Identify the number of students by district & level of language proficiency.		
6. Assess LEP students' language proficiency in the primary language & in second language.		
7. Maintain records on LEP student progress in the primary language & English in predetermined content areas (reading, language, math, other ...) using the appropriate instrument (list:).		
8. Develop & administer other assessments as appropriate for the specific program.		
9. Conduct a curriculum assessment on project organized, compiled, or developed materials, manuals, & guides. The assessment will be performed according to a curriculum rating scale available through the evaluator.		
10. For bilingual programs (not ESL or SDAIE), determine proficiency of staff in language(s) of instruction.		
11. Collect data to show that all assessments are valid, reliable, fair, and appropriate for these students, for this curriculum.		

Activities & Procedures	Completion Date	Person Responsible
MANAGING EVALUATION: STUDENTS 1. Establish a management information system that provides for project records, students records, & the organized collection & compilation of appropriate data & information. 2. Establish a student folder on all students served by the program. The folder will include: assessment, identification, enrollment, and student background information. Refer to the "student data" format for minimum background information. The "student data" form represents a computerized management information system. 3. Each staff member turns in a monthly summary report of services using the "staff report - time on task" or a similar instrument developed for that purpose. 4. Collect required data and information on dropout, retention, and other factors as appropriate. 5. Collect information specific to any traditionally underrepresented groups such as the physically challenged, LEP students in GATE programs, and so on.		
MANAGING EVALUATION: PARENTS 1. Assure that parents have been involved actively in enrollment of LEP students with adequate opportunity to decline enrollment. This may be accomplished through a letter with the appropriate parent sign-off or conferences and home visits with appropriate documentation. 2. Maintain records on parent group meetings and activities -- agendas, minutes, participation lists, and evaluations of associated training. 3. Monitor a parent activity -- a parent training session, parent group meeting, or parent-student activity. 4. Administer an annual parent group evaluation provided by the evaluator.		
MANAGING EVALUATION: IMPLEMENTATION 1. Develop schedules and instruments/protocols to assess implementation of the program. This should include staff development activities, administrative activities, and curriculum and materials development. 2. Conduct focus groups of students, parents, and staff to assess the implementation of the program.		
MANAGING EVALUATION: CONTEXT 1. For IASA programs, determine relationship of this program to other federally, state, and locally funded educational programs in the same school.		

Activities & Procedures	Completion Date	Person Responsible
MANAGING EVALUATION: STAFF 1. Complete a training evaluation form on all training activities sponsored by the project or participated in by project personnel. Put a summary of the training evaluations on the top of backup information along with an agenda and sign-in sheet identifying the participants by position (aide, teacher, administrator). 2. Complete a visitation report on all schools and programs visited by project staff and key personnel. 3. Conduct an annual assessment of staff training needs. The instrument provided should be administered to all project personnel and key classroom teachers (for a Comprehensive School program administer to all staff) in May or September of each year. Revise instrument. 4. Monitor at least one staff development activity according to a predetermined schedule. (Define:) 5. Develop a staff development higher education plan for each project staff and key personnel participating in higher education courses supported by the Title VII project. This includes a career ladder for paraprofessionals. 6. Evaluate the experience, training, and qualifications of key personnel on an annual basis using the instrument provided.		
MANAGING EVALUATION: PROGRAM 1. Collect information on how the activities funded by Title VII are coordinated with, and integrated into, the overall school program and to other Federal, State, or local programs serving LEP children and youth. 2. Utilize the form developed by the evaluator to assess the overall management of the program.		
REPORTING 1. Provide feedback to project staff and key personnel on evaluation findings. Make formal presentations after interim report and annual progress report. 2. Develop an interim "annual progress report" addressing preservice activities and/or monitoring results. This report will provide information on the progress of the project toward meeting its goals, and problems and corrective actions. 3. Complete biennial "evaluation report" per IASA statutes and regulations from EDGAR. This evaluation should address program improvement, further define the program's goals and objectives, and determine program effectiveness.		

EXAMPLE TITLE VII MANAGEMENT PLAN

This example was developed by the Title VII, ESEA, Bilingual Education Technical Assistance Unit based on work in Espanola, NM.

Activities		Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Person Responsible
Conduct training, Needs assessment; Set up Preservice & inservice plans				✗	✗							✗	✗	✗	✗	✗	Project Director (PD)
Deliver Preservice Plan-Year 1		✗	-	-	-	-	-	-	-	-	-	-	-	-	-	✗	PD, CC, IHE, Consultants
Deliver Inservice & Staff Development Plan		✗	-	-	-	-	-	-	-	-	-	-	-	-	-	✗	PD, CC, IHE, Consultants
Select Instructional Materials		✗	-	-	-	-	-	-	-	-	-	-	-	-	-	✗	Classroom & Resource Teachers
Implement Instructional Objectives		✗	-	-	-	-	-	-	-	-	-	-	-	-	-	✗	Teachers, Aides, Community Liaison
Implement Community Involvement Objectives		✗	-	-	-	-	-	-	-	-	-	-	-	-	-	✗	PD, Community Liaison
Monitor delivery of objectives				✗	-	-	-	-	-	-	-	-	✗				PD & Evaluator
Obtain student enrollment & test data					✗								✗				PD & Resource Teacher
Conduct Pretests:	CTBS										✗						Classroom & Resource Teachers
	LAS											✗					Classroom & Resource Teachers
	CRT					✗			✗		✗		✗				Classroom & Resource Teachers
Conduct Parent Meetings						✗		✗		✗			✗				PD & Community Liaison
Schedule & Monitor Program Evaluation Visits					✗	✗			✗		✗		✗				PD & Evaluator
Order Testing Materials for Next Year		✗	✗	✗													PD
Gather required data & information for evaluation		✗	-	-	-	-	-	-	-	-	-	-	-	-	-	✗	PD
Conduct Posttests:	CTBS										✗						Classroom & Resource Teachers
	LAS											✗					Classroom & Resource Teachers
Analyze & Disseminate Results		✗	✗	✗	✗												PD & Evaluator
Submit Data to Evaluator			✗														PD
Prepare end-of-year Progress Reports		✗	✗														PD
Prepare Biennial Evaluator's Report																✗	PD, Evaluator

Produced by the Evaluation Assistance Center-West/NMHU, Albuquerque, NM (revised 10/95)

Implementation Checklist for Title I Schoolwide Programs

This checklist is from work by Diane August, Kenji Hakuta, Fernando Olguin, and Delia Pompa (September 1995), Guidance on LEP Students and Title I, supported by Carnegie Corporation.

§ 1114(a) ... A Local Education Agency may use title I funds in combination with other Federal, State, and local funds, in order to upgrade the entire educational program in a school that for the school year 1995-6 serves an eligible school attendance area in which not less than 60% of the children are from low-income families; or for the school year 1996-97 an eligible school attendance area in which not less than 50% of the children enrolled in the school are from such families.

Activity	Position Responsible	Date to Start	Date to Complete
Ensure that the plan:			
1. Is developed during a 1-year period, unless the local education agency, after considering the recommendation of the technical assistance providers, deter-mines that less time is needed to develop and implement the schoolwide program or the school is operating a schoolwide program on the day preceding the date of enactment of the Improving America's Schools Act of 1994.			
2. Is developed with the involvement of the community to be served & individuals who will carry out such plan, including teachers principals, other staff, and, where appropriate pupil services personnel, parents, and if the plan related to a secondary school, students from such school.			
3. Will be in effect for the duration of the school's participation under Title I and reviewed and revised, as necessary, by the school.			
4. Is made available to the local educational agency, parents, and the public. ■ Translate, to the extent feasible, into any language that a significant percentage of the parents of participating children in the school speak as their primary language.			
5. Will be, where appropriate, developed in coordination with programs under the School-to-Work Opportunities Act of 1994, the Carl D. Perkins vocational and Applied Technology Education Act, and the National and Community Service Act of 1990.			
Ensure that the plan accomplishes the following:			
1. Incorporates the eight required components.			
2. Describes how the school will use resources under Title I and from other sources to implement those components.			
3. Includes a list of state and local educational agency programs and other federal programs under §1114(a)(4) that will be included in the schoolwide program.			

Activity	Position Responsible	Date to Start	Date to Complete
4. Describes how the school will provide individual student assessment results, including an interpretation of those results, to the parents of children who participate in the assessment.			
5. Provides for the collection of data on the achievement and assessment results of students disaggregated by gender, major ethnic or racial groups, limited English proficiency status, migrant students, and by children with disabilities as compared to other students, and by economically disadvantaged students as compared to students who are not economically disadvantaged.			
6. Seeks to produce statistically sound results for each category for which assessment results are disaggregated through the use of over-sampling or other means.			
7. Provides for the public reporting of dsaggregated data only when such reporting is statistically sound.			
Ensure that the following fiscal responsibilities are met:			
1. Use funds available to carry out the schoolwide program to supplement the amount of funds that would, in the absence of funds under Title I, be made available from nonFederal sources for the school, including funds needed to provide services that are required by law for children with disabilities and children with limited English proficiency.			
2. A school that chooses to use funds from such other programs shall not be relieved of the requirements relating to health, safety, civil rights, gender equity, student and parental participation and involvement, services to private school children, maintenance of effort, comparability of services, or the distribution of funds to state or local educational agencies that apply to the receipt of funds from such programs.			
Ensure that the plan includes professional development to be carried out:			
1. Where appropriate, as determined by the local educational agency, include strategies for developing curricula and teaching methods that integrate academic and vocational instruction (including applied learning and team teaching strategies).			
Ensure that the following schoolwide program accountability requirements are met:			
1. Use the state assessments described in the state plan.			
2. Use any additional measures or indicators described in the local educational agency's plan to review annually the progress of each school served with Title I to determine whether the school is making adequate progress.			

Activity	Position Responsible	Date to Start	Date to Complete
3. Publicize and disseminate to teachers and other staff, parents, students, and the community, the results of the annual review of all school performance profiles that include statistically sound disaggregated results.			
4. Provide the results of the local annual review to schools so that the schools can continually refine the program of instruction to help all children served under Title I in those schools meet the state's performance standards.			
Ensure that the schoolwide program includes the following components:			
1. A comprehensive needs assessment of the entire schools t hat is based on information on the performance of children in relation to the state content standards (essential elements) and the state student performance standards (TAAS).			
2. Schoolwide reform strategies that: ■ provide opportunities for all children to meet the state's proficient and advanced levels of student performance standards; ■ are based on effective means of improving the achievement of children; ■ use effective instructional strategies, which may include the integration of vocational and academic learning (including applied learning and team teaching strategies); ■ increase the amount and quality of learning time, such as providing an extended school year and before-and-after-school and summer programs and opportunities, and help provide an enriched and accelerated curriculum; ■ include strategies for meeting the educational needs of historically underserved populations, including girls and women; and ■ address the needs of all children in the school, but particularly the needs of children who are members of the target population of any program, which may include: • counseling, pupil services, and mentoring services; • college and career awareness and preparation, such as college and career guidance, comprehensive career development, occupational information, enhancement of employability skills and occupational skills, personal finance education job placement services, and innovative teaching methods which may include applied learning and team teaching strategies; • services to prepare students for the transition from school to work, including the formation of partnerships between elementary, middle, secondary schools, and local businesses, and the integration of school-based and work-based learning;			

Activity	Position Responsible	Date to Start	Date to Complete
2., cont'd • incorporation of gender-equitable methods and practices; • how the school will determine if such needs have been met; and • are consistent with, and are designed to implement, the state and local improvement plans, if any, approved under Title III of the Goals 2000 Educate America Act.			
3. Instruction by highly qualified professional staff.			
4. Professional development for teachers and aides, and; where appropriate, pupil services personnel, parents, principals, and other staff to enable all children in the school to meet the state's student performance standards.			
5. Strategies to increase parental involvement, such as family literacy services..			
6. Plans for assisting preschool children in the transition from early childhood programs, such as Head Start, Even start or a state-run preschool program, to local elementary school programs.			
7. Measures to include teachers in the decisions regarding the use of assessments in order to provide information on, and to improve, the performance of individual students and the overall instructional program.			
8. Activities to ensure that students who experience difficulty mastering any of the standards required by the state plan during the course of the school year shall be provided with effective, and timely additional assistance, which may include ■ measures to ensure that students' difficulties are identified on a timely basis and to provide sufficient information on which to base effective assistance; ■ to the extent the school determines feasible using funds under this part, periodic training for teachers in how to identify sch difficulties and to provide assistance to individual students; and ■ for any student who has not met such standards, teacher-parent conferences, at which time the teacher and parent shall discuss: • what the school will do to help the student meet such standards; • what the parents can do to help the student improve the students' performance; and • additional assistance which may be available to the student at the school or elsewhere in the community.			

SPECIFYING GOALS

Defining a goal

The first step in specifying goals is to have a clear understanding of what a goal is. A goal generally is defined as a statement of the program's intent, purpose, or expected outcome(s). Another way to think about a goal is by referring to it as a statement about where the project is headed. For example

To increase the proficiency of LEP students.

This statement is a goal because it describes what the program ultimately intends to achieve. Compare this with the following statement:

To provide English language instruction to LEP students.

This statement is not a goal because it refers to what the program does, now what it intends to accomplish. When goals are stated as "end results," programs are able to identify effectively and to prioritize the direction for addressing the needs of LEP students.

Three characteristics to consider when writing goals

1. **State goals in broad and general terms.**
   Goals are abstract, idealized statements of what we want ourselves or others to achieve. Goals focus on the purpose or desired achievement in general terms without specifying the performance, criteria, or conditions under which they will be achieved.

2. **Identify the target group to be involved in each goal.**
   When writing a goal statement, identify the target group, classroom, or school designated to achieve the goal. Identifying the target population for whom the goal is intended clearly communicates the goal to planners, implementers, and evaluators.

3. **Describe the goal as an intended outcome rather than as a process.**
   Goals should be used to help focus on what you are trying to attain rather than on how you plan to accomplish it. Some example goals statements are provided below.

   To increase English proficiency of LEP students.

   To have the teaching staff integrate students' native language as a medium of instruction in all content areas.

   To improve the teaching staffs skills in presenting cultural awareness activities.

   To have the bilingual instructional staff develop instructional science materials in the students' native language.

   To enhance parent participation in the proposed program.

DETERMINING APPROPRIATE GOALS

In determining the appropriateness of a goal, it is important to ask yourself, "How can I be sure that a goal accurately describes the project's intent or purpose?" To determine appropriateness, ensure that:

1. **Goals relate to the educational aims.**

   A major aim of Title VII is "to establish educational programs using bilingual or special alternative instructional practices (where appropriate), techniques, and methods for limited English proficient students" (20 USC 3282). In an attempt to address the educational aims of Title VII, efforts should be made to relate goals to effective programming for culturally and linguistically diverse students.

2. **Goals are linked directly to needs.**

   Goal statements should be generated from a comprehensive needs assessment of the project students. For example, if linguistically diverse students have low scores in vocabulary, grammar, spelling, and reading comprehension, one likely goal would be "to improve students' skill in reading."

3. **Goals correspond to student outcomes and project components.**

   When goals relate to expected student outcomes and components of the program, they provide a dialogue for identifying and operationalizing (1) the skills students will be expected to learn in order to participate successfully in the mainstream classroom; (2) the type of training required by the teaching staff; (3) the expected classroom instruction and learning materials; and (4) the extent of parent involvement.

4. **Goals are prioritized according to the realities of the school or school district.**

   After identifying goals for the program, you may find yourself with too many goals to achieve. In such a case, you may want to use one of several methods for prioritizing goals: random selection, importance and feasibility, importance of the component, or hierarchical order.

Methods for Prioritizing Goals

When a program is developed in an area of high need, it is not uncommon for a program staff to find that they have developed more goals than they can realistically expect to meet within the program period. Below are described four different methods for determining which of the goals should be selected for implementation. Note that this procedure should be used before the program is submitted to a funding agency. Once the agency has funded the program it is difficult to delete goals (unless a needs assessment determines that they are no longer required), although it will be possible to modify them to ensure evaluability.

Methods	Advantages	Disadvantages
Random selection of goals		
Randomly select the goals to be used from the larger total set of goals.	✧ Can be done by one person ✧ Quickest, simplest method ✧ All goals are equally important	✧ Risks missing goals that might be important ✧ Risks credibility of the evaluation (especially if "wrong" goals are selected)
Rate goals according to level of importance and feasibility		
Identify a team to participate in goal selection. Rate the goals from most important to least important and from most feasible to least feasible. Use only those goals that are rated the highest.	✧ Fast method ✧ Gives other people input into selection process ✧ Makes it likely that important goals will be selected	✧ Depends on cooperation among team members ✧ Raters may not represent the opinions of other individuals
Rate goals according to the importance of their components.		
Identify a team to participate in goal selection. Determine the components of the project (e.g., student achievement,. staff development). Assign each goal to a component. Rate each set of goals according to the importance and feasibility of its component.	✧ Fast method ✧ Gives others input to selection process ✧ Makes it likely that important goals will be selected	✧ Depends on the cooperation of the team ✧ Raters may not represent the opinions of others
Select goals by hierarchical order		
Assign each goal to a component. Chart the goals within each component from most simple to most complex. The most complex, or "terminal," goals are given highest priority.	✧ Can be done or without a team ✧ Assigns priority to the most logically complex goals	✧ Relatively time consuming, depending on the number of goals ✧ May not be desirable to test only complex goals

REVIEWING OBJECTIVES

What are the qualities of a "good" objective? There generally are four qualities to consider when writing objectives.

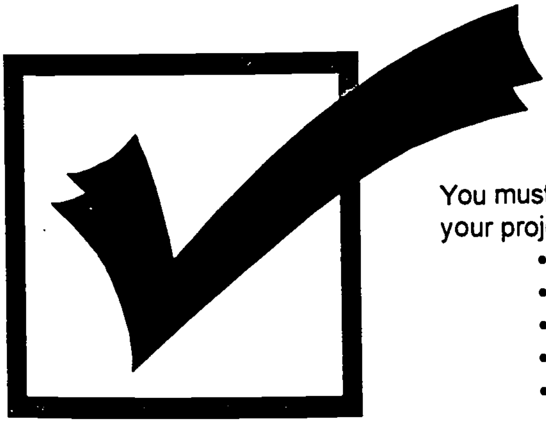
1. A good objective specifies an outcome rather than a process.
2. A good objective is stated as an overt behavior.
3. A good objective uses strong action verbs.
4. A good objective describes only a single outcome.

What should be included in an objective? Keep in mind your <u>ABCDs</u>:

- Audience, the learner: WHO will be involved?
- Behavior, the target performance: WHAT must the learner do to provide evidence that the objective has been attained?
- Condition(s), the circumstances under which the behavior will be demonstrated: GIVEN what?
- Degree, the criterion of success: HOW WELL will the behavior be performed?

Example objectives

- Given that students regularly attend class, the third grade project students will show an increase in language proficiency of at least one level on the SOLOM by the end of the first year.
- By the end of the project, each instructional staff person will have successfully completed (as measured by a grade of "B" or better) a course in English language development at the local university.
- During small group work, project teachers will use cooperative learning instructional techniques with 85% competency as measured by the Cooperative Observation Checklist.
- Curriculum development staff will complete one training guide for science education by the end of the second year of the project. The guide will be utilized, revised, and evaluated during the third year of the project. (Note that this is two objectives.)
- Given that they have transportation, parents of project students will attend at least four PAC meetings during the first year of the project.



You must have at least 1 goal and objective for each component of your project. This usually involves

- student achievement and proficiency,
- staff development,
- instructional design/curriculum,
- materials development, and
- family involvement.

CREATING ACTIVITIES

DEFINING AN ACTIVITY

Another element in communicating the intent of the program is through its activities. The purpose of the activities is to describe in detail any prerequisites or actions necessary to ensure achievement of the objective. By prerequisites, we are referring to the conditions and/or criterion in which the objective is to be achieved.

CHARACTERISTIC OF AN ACTIVITY STATEMENT

An activity statement consists of one major characteristic -- a clear description of the performance or expected behavior. Some examples include:

- ... will attend four workshops
- ... will have developed instructional materials
- ... will implement instructional approach one hour each day

In most cases, activity statements will be very similar to the objective performance statements. The purpose of the activity is to describe more precisely the intent of the objective.

EXAMPLE OBJECTIVE AND ITS ACTIVITIES

Activities might include a description of the prerequisites necessary to achieve the objective. For example, the objective *Instructional staff will use the cooperative learning approach when teaching mathematics* might have activity statements such as:

1. Attend one full week of initial training on cooperative teaching approaches.

2. Train students on cooperative learning strategies during the first 2 months of school.

3. By the end of the third month of school, use cooperative approaches in mathematics at least four days a week.

4. Attend a minimum of four out of eight in-services provided by the school district on cooperative learning.

MODIFYING OBJECTIVES

Problem

You (and/or your evaluator) realize that one of your objectives is not appropriate. This might be because it no longer relates to the needs of the students, it is not quantifiable, it is not written in an appropriate manner, the "objective" really is an activity, or for some other reason.

Solution

Modify the objective. This can be done as long as the overall scope of the project does not change. The procedure suggested by many funding agencies is as follows:

1. Determine why this objective is not appropriate as currently written.
2. Modify the objective so that it is appropriate.
3. Justify the modification.
4. Contact appropriate personnel within the funding agency by telephone (within the Office of Bilingual Education and Minority Languages Affairs, this is the project officer). Explain 1-3 above; be sure to stress that this modification will not change the scope of the project. Ask for permission to make the modification. This usually will be approved. Ask how to proceed to ensure written permission; for instance, some have suggested that you should write a letter listing a date and indicate that you will assume that the objective can be modified unless you hear differently by that date -- others want you to ask for (and receive) a letter from the agency.
5. Follow-up: contact the funding agency by letter. Repeat your reasoning as to why the modification is necessary and how the objective should be written. Carefully remind the person that permission was given in the phone conversation. Follow his/her suggestion for ensuring written approval. OBEMLA guidelines suggest that you should hear from the program officer within 30 days.
6. In the next annual report, and in the evaluation report, include the original objective, why it was modified, and that permission to modify was granted by the project officer. Then provide the "new" objective.

Planning Goals and Objectives

Goal: *Example: To increase English language proficiency in all four modalities.*				
Objective(s): *Example: Students who attend class regularly will read one novel of their choice by the end of the school year.*	Activity(ies): *Example: Students will (1) identify genres of novels, (2) read introductory chapter from 1 novel in each of 4 genres, (3) read 1 novel in its entirety, (4) complete the class reading log questionnaire.*	Assessment(s): *Example: Class reading log questionnaire (4-point rubric plus written responses)*	Person(s) Responsible: *Example: Classroom teacher and evaluator*	Timeline: *Example: Beginning 2nd quarter of school year; complete by end of year.*
Goal:				
Objective:	Activity(ies):	Assessment:	Responsible:	Timeline:
Goal:				
Objective:	Activity(ies):	Assessment	Responsible:	Timeline:

Goal: A broad, general statement about where the program is going.
Objective: A more specific statement about the expectations for the program. Don't forget the "ABCD"s of each objective.
Activity: A statement that describes in detail any prerequisites or actions necessary to ensure the achievement of the objective.
Assessment: How do you know that the objective was met? (This might be a checklist of behavior[s], a holistic rubric used to assess the progress made on a particular project, or any other type of assessment.)
Responsible: Who is "in charge" of ensuring that the activities take place and the evaluation is completed?
Timeline: When will the activity(ies) occur? When will the assessment take place?

Created by the Evaluation Assistance Center-West/NMHU; Albuquerque, NM (Revised 12/95)

PURPOSES FOR ASSESSMENT

Revised from E.D. Roeber (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.

Monitoring

- Assessment provides periodic measurements of student progress in order to determine the educational "growth" of a student from one time to another.
- Assessment provides a periodic measure of the performance of groups of students to track performance over time.

Information/Accountability

- Assessment informs parents and students about student performance in order to encourage students and/or teachers to improve performance.
- Assessment provides the public with information about the performance of groups of students in order to encourage schools to improve the system.

Improving Student Performance

- Assessment provides data to teachers and students which encourages instruction geared to the needs of individual students.
- Assessment provides information to educators on groups of students which can be used to review current instructional strategies and materials and to make improvements where needed.

Allocation of Resources

- Assessment provides information to determine where instructional staff are needed.
- Assessment provides information to determine where financial resources are most needed.

Selection/Placement of Students

- Assessment helps determine the eligibility of students for various educational programs or services.
- Assessment determines the program or service most appropriate for the instructional level of the student.

Certification

- Assessment provides a means of determining the competence level of individual students.
- Assessment provides data to certify the adequacy of an educational program.
- Assessment provides data to certify the acceptability of an educational system (accreditation).

Program Evaluation

- Assessment provides the information needed to determine the effectiveness of an educational program or intervention.

Issues in Designing a Student Assessment System

Basic Issues

Purpose
□ Academic Achievement
□ Language Proficiency
□ Attitude
□ Other

Type
□ NRT
□ CRT
□ Alternative

Reason
□ Progress
□ Grading
□ Placement
□ Accountability

Language(s) of Questions
□ English
□ Student's L1
□ Both

Language(s) of Answers
□ English
□ Student's L1
□ Both

The instrument/procedure
□ Exists, OK as is
□ Exists, modify
□ To be developed

Frequency of Measurement

□ Yearly, spring-to-spring
□ Yearly, other time frame--justify: ____________________
□ Academic year, fall-to-spring
□ Once each semester
□ Quarterly □ Monthly □ Weekly
□ Other: ____________________

Rating Scales

□ Holistic — An overall, impressionistic score. Usually ranges from 1-3 up to 1-10; can be numbers, symbols (+, ✔, -), or letters (A-F).

□ Primary Trait — A method that scores particular parts of a performance or product, usually on a 1-4 or 1-6 basis. Each of these parts, or traits, is scored separately. A final score may be the individual traits, or may be the sum of the trait scores.

□ Analytic — A variation of primary trait in which particular trait(s) are considered more important than others and are weighted accordingly. Weightings might be based on recent topics within the curriculum, weaker areas for a student, or others.

Note that for each type of scoring, the rubrics must be defined and described. What is it that makes a performance a + rather than a -, or that gives a trait 6 points instead of 5 points?

□ Frequency — A count of the number of times something occurs; e.g., the number of books read in a month, the number of homework assignments turned in.

Scores to be Used

Used only for commercially-available test:

□ Stanine
Divides the students into 9 groups, gives gross description of student's performance. Cannot be used for evaluation purposes; cannot be used to average scores. Might be used for descriptive purposes.

□ Grade Equivalent
Achievement expressed in terms of grade level, provides information about performance at grade level but scores off grade level are only estimates. Cannot be used for evaluation purposes; cannot be used to average scores. Might be used for descriptive purposes.

Can be used for commercial or classroom assessment (with enough students, usually over 75-100):

□ Percentile
Represents the percent of students who scored at or below that score. Of limited use for evaluation purposes; cannot be averaged across different assessments or to indicate gains.

□ Scale or Standard*
Provide information about achievement that can be used longitudinally, and across tests or schools. There are several types of scale and standard scores, so also must have knowledge about this particular form of score.

□ NCE*
The Normal Curve Equivalent is a particular type of standard score, ranging from 1-99. Designed for use with "Gap Reduction" evaluation design; useful for all evaluations. Can be used longitudinally, and across schools and tests. Must indicate raw scores or percentage correct as well as NCEs.

Can be used with any type of assessment instrument:

□ Raw; Percentage
Number (or percentage) of items answered correctly. Of limited use for evaluation purposes; but can be useful to determine mastery. Cannot be averaged across different assessments All other scores are converted raw scores.

*With scale scores, standard scores, and NCEs, obtaining the same score from one year to the next indicates that the student is learning the material expected of him/her.

Test Administration: Who will administer the instrument? Is training needed?

□ Program Director
□ Teachers
□ Students
□ Other: ______________________________

□ School Administrator
□ Paraprofessionals
□ Evaluator

STANDARDS FOR TESTING BILINGUAL PERSONS

1. For a bilingual person, any test that relies on English becomes confounded since in unknown degrees it becomes an English test.

2. Bilingualism is a complex phenomenon involving all aspects of literacy, communication, and social functions.

3. Mental processing in the weaker language may be slower, less efficient, and less effective.

4. Language *background*, not just language proficiency, must be taken into account in every facet of assessment such as test development, selection, administration, and interpretation.

5. Tests developed without accounting for language differences are limited in their validity and in how they can be interpreted.

6. Psychometric properties (e.g., reliability, validity) do not translate from one language to another and hence translations do not work.

7. Measuring proficiency in L1 and L2 "may be necessary" to design instructional programs.

8. Proficiency in English should be determined along *several* dimensions.

9. The ability to speak English in naturalistic situations may not predict the ability to learn academic material in English.

10. Assessment of nonnative speakers of English will take extra time (more tests and observations).

11. Particularities of cultural background can lower test performance.

12. Special training for bilingual communication in testing may be profitable and beneficial.

13. Tests must be proven to be equivalent if they are formulated in L1 and L2.

Produced by the Evaluation Assistance Center-West/NMHU, Albuquerque, NM (Revised 10/95). Revised from the Joint Committee of the American Educational Research Association, American Psychological Association, & American Council on Measurement, *Standards for Educational and Psychological Testing*, Washington, DC: APA, 1985.

SELECTING APPROPRIATE ACHIEVEMENT AND PROFICIENCY TESTS

1. Identify assessments that might be appropriate. Ask other personnel from other programs, utilize published lists of tests; consider the use of district/state recommended or mandated tests to help reduce the testing load on students.

2. Operationalize the instructional objectives of the program. Match the test items or subtest descriptions to the instructional objectives in terms of content and level of difficulty. If this test is to measure language proficiency, be sure the test's purpose matches the program's definition of proficiency. This may include oral production oral comprehension, writing production, and reading comprehension. It may be necessary to utilize more than one assessment.

3. Utilize a published critique of tests. Find what experts think of the technical aspects, item construction, and test construction of each assessment under consideration.

4. Review the Test Manual to determine the characteristics of the test's norm group (they should be similar to the program's or school's students), reliability and validity, type of norms available for the time of the year you will be administering the test (empirical, based on actual administration, or interpolated, based on a "statistical guess" about what students would score if it were administered at this time).

5. Examine the scoring services and reporting formats available. Be sure the scores needed for evaluation purposes (NCEs, percent correct, raw scores, scaled scores, or another type of standard score) are available.

6. Determine whether the test is designed for individual or group administration (some can be used in either manner).

7. Decide whether the time needed for testing is realistic for the program (e.g., an 80-minute test for class periods of 50 minutes will not work without modification of the school's day). Also, ensure that the cost of the test is appropriate for the program.

8. If the test is still considered appropriate, obtain a specimen set or sample of the test before the final decision is made. Create a panel (administrator, teacher, program director, parent, and, if for a secondary program, a student) to review the sample for cultural, linguistic, and gender bias; appropriateness of items for this geographic area.

9. Determine whether staff have the necessary expertise to administer the test.

10. If the test still is considered appropriate, try it for one year.

CHOOSING AN ASSESSMENT APPROPRIATE FOR YOUR PROGRAM

Many assessment instruments, both standardized and alternative, are available. What is the best for your program? How can you choose the best for your students? Although EAC-West cannot recommend a particular test for any purpose, there are some general guidelines that should be followed.

1. When selecting an instrument that already exists, follow the guidelines in the EAC-West handout *Selecting Appropriate Achievement and Proficiency Tests.*

2. When developing an alternative instrument, follow commonly used guidelines and plan on at least 1 year to produce a really good instrument. This may seem like a long time, but you should be able to use the instrument for several years.

3. When assessing language proficiency, be sure to assess all four areas of reading, writing, speaking, and listening. This may require more than one instrument.

4. To meet Title VII evaluation regulations, be sure that
   - reliability and validity can be documented (i.e., how good are the results? does the test give you good, valuable information about the students?),
   - the scoring is objective and the administration of the assessment is standardized, and
   - scores are collected on enough students to ensure that the results will apply to all students in the program.

   In addition, Title VII required that you assess each content area (including English language arts achievement and language proficiency). At least one assessment in each area should have a nonproject comparison group (the norm group for the instrument or another group of students in your school/district/state).

5. Use appropriate scores, such as NCEs, standard scores, raw scores, or percentage correct scores. These can be manipulated (that is, used in mathematical computations) and compared across the years (using raw scores or percentage correct scores requires that the same instrument be used each year).

6. Do not use a set cut-off score to determine eligibility for a program (entering or exiting). Instead, create a lower limit cut-off and a higher limit cut-off; scores in the bandwidth between these two cut-offs indicate that further information is needed to make a decision. As an example,
   - lower than 40% correct indicates that the student needs further assistance before moving on to the next curriculum unit;
   - higher than 50% correct indicates that the student is ready to move on to the next unit; but
   - scores between 41% correct and 49% correct indicate that further assessment is necessary to determine whether the student has the skills to allow success in the next unit.

7. Use multiple measures to assess students' proficiency and achievement. This might include a standardized instrument, a behavioral checklist, a written work, and/or other alternative assessments. More instruments will give a better overall view of the child's skills; don't be surprised if different results appear on different instruments (especially standardized test vs alternative instruments).

8. Do not test unless there is a purpose to the testing. Remember that much of alternative assessment can be curriculum-embedded. That is, it is not "test time."

9. Always be sensitive not only to the language of the children (both their English and home language proficiencies), but also to the culture of the children.

GUIDELINES FOR DEVELOPING RELIABLE AND VALID ALTERNATIVE ASSESSMENTS

In order to be effective, alternative assessment techniques must be carefully planned. Alternative assessments can serve diagnostic purposes as well as formative and summative evaluation purposes. General guidelines for planning an alternative assessment are presented below.

1. Begin with a clear statement or operationalization of the instructional objectives. If the content of a test or assessment does not match the instructional objectives of the program (the student's expected performance), the test's validity may be negated.

2. Develop clear and observable scoring criteria that span the potential performance of students (e.g., highest quality to lowest quality level). Careful thought should be given to selecting the adjectives or verbs in order to clarify interpretation of the students' performance.

3. Select an appropriate alternative assessment technique to measure student performance. Some popular techniques used to assess student performance include checklists, rating scales, holistic scales, and questionnaires.

4. Conduct a judgmental review of the criteria described in the selected alternative assessment technique. Select judges to rate the statements used in each category or item with respect to how they relate to program goals and objectives.

5. Allow time to test the assessment instrument and its ability to draw the information desired.

6. Draw directly from students' work to ensure evidence of whether students are mastering the intended objectives.

7. When more than one rater is involved in assessing student performance with the selected assessment instrument, validity of the judgements can be ensured by training raters to meet a set criterion. For example, when judging 10 student papers, raters should give the same score to at least 8 of the 10 papers and be no more than 1 or 2 points apart (depending on the scale used) on any paper.

8. Maintain objectivity in assessing student work by periodically checking the consistency between assessments given to students' work.

9. Keep consistent and continuous records of the students to measure their development and learning outcomes. Multiple assessments (at least six observations or judgements across the course of the year) allow a more comprehensive and objective assessment of the students' performance.

10. Consider using multiple measures such as other tests or performance-based assessments to assist in validating the alternative assessment.

11. Score assessments in such a way that they allow for aggregation or summary of individual data into group data. Maintain uniform administration and scoring procedures to ensure the reliability and validity of the data.

HOW TO DEVELOP A HOLISTIC ASSESSMENT

1. Based on the content area and skills to be assessed, develop rubrics or criteria for scoring students' work. As an example, assume an assessment of writing. Rubrics should reflect the full range of work students may potentially produce -- from the lowest quality level to the highest quality level. Possible ranges include scales from 1-10 to 1-3. Remember that the broader the possible range (e.g., 1-6 as opposed to 1-3), the more detailed the information about the student's progress.

2. Select a topic for students to write about that is of interest and appropriate to their age or grade level. Use evocative materials that will stimulate student's writing (e.g., photographs, objects, experiences).

3. Develop instructions for students that provide enough detail so they know what is expected of them. For instance, directions to "write about Valley Forge" are much less interesting and inviting than "you are a soldier at Valley Forge. You are cold and hungry. You are writing a letter to your parents to tell them all about what you see and how you feel as you wait for the British. What would you write in the letter?"

4. Have all students from each grade level write on the topic selected. Make sure papers are anonymous until scoring is completed. Students can use their social security number or an assigned number of some type.

5. After students have completed their writing assignment, randomly select at least ten papers from each grade level. Duplicate these sample papers for scoring. These papers will become the **benchmarks** for scoring the rest of the students' papers.

6. Identify raters who will be involved in the rating process and have them score the sample papers, one grade level at a time.

7. After individually scoring the sample writings, raters should compare scores to determine the extent of agreement and reliability of the scoring procedure. Raters should seek to reach agreement on at least 9 out of 10 writing samples and no more than one score level difference on any writing sample. Be sure that those receiving the highest score deserve that score -- i.e., their paper matches the rubric description. This is especially important at the top end of the rubric to ensure that papers scoring at the highest level are not just the best available papers.

8. Once a high level of agreement has been established between raters, have them independently score the rest of the students' writing papers. Conduct intermittent sessions to ensure that the standards of reliability remain consistent between raters.

9. After completing all the scoring, tabulate results for each student. To calculate the average score by classroom, add al the students' scores and divide the total by the number of students who participated and turned in papers.

TWO MAJOR ASSESSMENT ISSUES: VALIDITY AND RELIABILITY

Revised from J.L. Galvan. Validity and Reliability Issues in Portfolio Assessment. *Paper presented at the California Association for Bilingual Education conference, Anaheim, CA., February 1993.*

You have decided to develop a new method of assessment for your program. This may be one individual new alternative assessment, or it may be a program of portfolio assessment. In either of these cases, two issues, in particular, must be considered. The first is whether the assessment will be a valid measure(s) of your students' abilities. The second is whether the assessment you develop will be reliable.

First, let's consider whether the assessment you develop is **valid**. In other words, do you and others think it measures what it is supposed to measure? Can you be sure that it does?

- **Can it be trusted?** Does the procedure appear to measure what we intend to measure? Does the teacher believe in the process? Does the student believe it is fair? Do the parents feel comfortable with it? Do the administrators trust it?

- **Is it an accurate measure of what the students know?** Do we know that a student's progress as measured by the assessment procedure is a correct measure of what the student knows?

- **Does the procedure compare well with other kinds of measures of the same thing?** If we're measuring reading, would the student be able to perform as well or as poorly on some other kind of reading test? If we're measuring second language proficiency, does the assessment give us an accurate picture of how well/poorly that student knows the target language?

Next, let's consider whether the assessment you develop is **reliable**. In other words, are we measuring the same thing every time we use it? Does everyone who uses it measure the same thing as everyone else?

- **Are the instructions written clearly?** Can you be sure that the assessment is put together in the same way for all students? Does every student receive the same level of assistance on the test? or in the development of their portfolios? Are the procedures clear enough that everyone understands the same thing about what goes in and how that is decided?

- **Are all teachers trained in the same way?** In other words, does your training plan clearly spell out the kinds of training experiences all teachers will receive?

- **How will you know that ALL teachers and students follow the procedures in the same way?** Have you thought about how you will ensure that the contents of the portfolios, or the procedures for administering and grading the assessment, will be comparable across classrooms? If you want to provide for some teacher flexibility, have you thought about how you will allow for some freedom of choice about contents while ensuring that the evaluation results are comparable?

- **Are the evaluation criteria spelled out clearly enough that they will be used in the same way by all teachers?** In other words, once the portfolios are assembled, or the assessment completed, are all of your teachers evaluating it the same way for all of their students?

- **Have you planned to conduct inter-rater checks?** Do you have a plan for ensuring that your teachers' judgements about their students' work are based on the same scale? In other words, if a score of "5" means outstanding, have you checked to make sure that a "5" rating by Teacher A is comparable to a "5" rating by Teacher B and with one by Teacher C?

Given these potential problems, what are some things we can do to ensure validity and reliability? Again, these issues are related to portfolio assessment, an individual alternative assessment, and even for the use of standardized tests.

- **Be very clear about what you want to measure.** Make sure that you describe what you want to measure in as much detail as you can. It's not enough to say that you want to measure language proficiency you need to decide whether that means oral language skills (such as delivering a speech, describing a picture, or participating in a dialogue) or whether it means producing an error-free composition.

- **Compare with other measures.** Once you know what you want to measure, figure out how you will know whether the assessment plan will result in measurements of these things. Can you think of other kinds of measures of these same things which you can compare with the results of the assessment? Can you compare them with the CTBS scores, or can you use the SOLOM, etc.?

- **Work as a team** at the District (or, at the very least, at the School) level to develop the plan. Make sure that the plan makes sense to all participants and that it is detailed enough to ensure that everyone will interpret it in the same way.

- **Make sure that you develop a training plan** along with the assessment plan, so that everyone who will use the plan will receive the same kind of training in its use.

- **Try out the plan** on a limited basis and evaluate the procedures and the results, and, if necessary, revise it and try it out again, before distributing it to everyone to use.

- **Follow all procedures carefully.** In other words, as part of your training plan, make sure that you include a way to ensure that everyone will use the plan in exactly the same way, what we call "inter-rater reliability."

- **Conduct inter-rater checks often.** Once the assessment is in place, make sure that you re-check for inter-rater reliability at least once each year, more often if possible.

ENSURING THE RELIABILITY AND VALIDITY OF ASSESSMENTS

"Assessment" is a broader concept than just "testing" -- it includes paper-and-pencil tests, rating items on scales, observation of student performance, critiquing student products, conducting interviews, and reviewing students' background or previous performances. Any assessment of students must be thought out and planned very carefully. EAC-West has developed *Guidelines for Developing Reliable and Valid Alternative Assessments* which describes some of the planning procedures that should be followed. Below is a framework for evaluating the technical validity of both standardized and alternative assessments.

Consider the **consequences**, intended and unintended effects of assessments on the ways teachers and students spend their time. For example, if students always are allowed 20 minutes in which to write an essay, what will happen if they are allowed only 10 minutes, or are allowed 30 minutes instead -- will they be able to produce a well-written document?

The **fairness** of the assessment is especially important. Included in this category are fairness in item development, scoring procedures (including the training and calibrating of raters), access to the teaching of the topic, and so on.

The results must **transfer** to other situations and must **generalize** to other groups of students and to other tasks. Sometimes it will be necessary to include measures of different types in order to assure transferability and generalizability.

Consider the **cognitive complexity** of the task. This typically has been a criticism of paper-and-pencil tests, but we cannot assume that all alternative assessments measure cognitively complex tasks.

The **quality of the content** must be consistent with the best current understanding of the field and reflective of what are judged to be aspects of quality that will stand up to the test of time. Perhaps more importantly, the tasks to measure a given domain should be worthy of the students' and teachers' time and effort.

The scope of teaching and the measurement tool must match in their **content coverage**. In addition, the assessment must be **meaningful** to students in order to elicit their best efforts and to increase their motivation level. Finally, **cost efficiency** is important in any assessment endeavor.

roduced by the Evaluation Assistance Center-West; Albuquerque, NM (Revised 10/95). Modified from R.L. Linn, E.L. Baker, & S.G. unbar (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20*(8), pgs. 5-21.*

Creating Your Own Rubric

"Rubric" comes from the Latin word for "rule." A rubric is a set of scores that describes how well a student, or group of students, is performing. The rubric not only reflects the scores (e.g., 0-3, 0-6) but each score includes a description of its meaning (e.g., the score of 0 means "no response on paper"). There generally are five steps to developing a rubric for use in the classroom. These are defined below, with examples included.

1. Determine the type of rubric to be used.

Three types of rubric are used:
① Holistic: provides an overall score for the effectiveness of the work.
② Primary trait: provides a score for each of several "parts" of the work--can be summed for a total score, or left as separate scores.
③ Analytic: weights the scores for some parts of the work more heavily than other parts of the work. Such a score might reflect aspects that have been the focus of the classroom recently, or parts of the work that are more important than others.
Note: the latter two types of rubric are sometimes combined and referred to jointly as "analytic." They are separated here for ease of discussion.

Any of these types can be used for most types of work. Which is needed is based on the purpose of the assessment (is it a quick view of where students are? is it the basis for a final grade? do you want to emphasize current classroom topics?) and the teacher's preferences.

2. Determine the spread of scores to be assigned.

Rubrics generally begin with a zero point that indicates no response on the student's part. Then, rubrics usually range from 1-3 (low, average, high) to 1-8 (providing more discrimination between scores). As an example, see the generalized rubric below.

Score 6	Exemplary Achievement
Score 5	Commendable Achievement
Score 4	Adequate Achievement
(Demonstrates	general understanding)
Score 3	Some evidence of achievement
Score 2	Limited evidence of achievement
Score 1	Minimal evidence of achievement
Score 0	No response

3. Create the descriptors for each score.

The descriptors need to be clear enough that others could use them, and to cover the various ranges of student responses. While not all scores need descriptors, at least every other and the end-points (anchors) need descriptors. As an example, see the rubric below.

5 points Excellent Achievement
Demonstrates internalized understanding of major concepts
- Solves problem and gets correct answer
- Shows computation, draws and label diagram
- Includes alternative attempts to think about problem
- Reflects on and generalizes about methods and solutions

3 points Adequate Achievement
Demonstrates a general understanding of major concepts
- Solves problem and gets correct answer
- Shows computation and gets correct answer

1 point Limited Evidence of Achievement
Demonstrates a lack of skills necessary to reach solution
- Draws incorrect diagram or makes incorrect computation

4. Use the scoring rubric on a set of papers. Revise descriptors as necessary.

After scoring the papers, consider the types of scores given. Do the scores vary? do they vary in a way you would anticipate (good students have higher scores, weaker students have lower scores)? Consider the following:

No high scoring papers. Did the activity require something not taught in class? Is something required in the rubric that was not asked for in the task? Is the rubric too difficult for students?

All high scoring papers. Is the rubric too easy for students? Are expectations too low? Is this result acceptable?

No passing papers. Were the directions misinterpreted? Were the directions wrong? Was this task very different from the usual assignment? Was something "going on" in class, or in the community, that might have impacted students?

Results do not seem to match expectations--do not match other measures of achievement. Is there a problem with the task? Is there a problem with the rubric? Do the task and rubric form a cohesive package? Should this be used in another class as further testing of the package?

5. Standardize the process with a set of anchor papers.

When the rubric development process seems fairly comfortable, use a set of papers to find those that best exemplify the specific scores/descriptors. Be sure that a paper, or a few papers, are not given the highest or lowest score simply because they are the best (or worst) papers in the group. The purpose of the descriptors is to ensure that papers meeting a set criterion receive the score that characterizes that paper; scoring rubric do not compare students to one another. The anchor papers can be used to further define the score or to train others to use the rubric.

Consider primary trait rubrics

The above boxed examples measure progress in one area. It is possible to use primary trait-types of rubrics to measure more than one feature of the work within one assessment. For instance, the SOLOM measures oral proficiency in each of five areas (comprehension, fluency, vocabulary, pronunciation, and grammar) on a 5-point scale. This method can provide a great deal of information to the teacher and to the student, all on one assessment form.

GOALS → OBJECTIVES → ACTIVITIES→ ASSESSMENT → EVALUATION

The essence of program design is the relationship of its various components. In fact, this relationship is circular: the completed evaluation suggests modifications to the goals, which starts the cycle again.

Goals

- are stated in broad and general terms,
- identify the target group to be involved, and
- describe an intended outcome rather than a process.

Objectives

- specify outcomes rather than a process,
- are stated as overt behaviors,
- use strong action verbs, and
- describe a single outcome.

Activities

- describe in detail any prerequisites or actions necessary to ensure achievement of the corresponding objective.

Assessment

- identifies students to participate in the project,
- measures progress of students towards achieving the objectives of the program, and
- provides year-end summaries of students' skills and proficiencies.

Evaluation

- determines whether the activities have been completed,
- ascertains whether the objectives have been met,
- concludes whether the goals were achieved,
- determines whether the program has been successful, and
- suggests ways in which the program can be improved -- modified goals, objectives, and activities.

Appendix IV

Guidelines for Managing the Evaluation Plan
Student Oral Language Observation Matrix (SOLOM)
Student Data Sheet for Use in Multi-Year Evaluations
Matching Objectives to Evaluation Design
KEYS TO ... Testing differences between groups: Grade Cohort
Basic Grade Cohort Design
Data Presentation for Grade Cohort Design
Advanced Grade Cohort Design
KEYS TO ... Testing differences between groups: Gap Reduction
Gap Reduction Design
Data Presentation for Gap Reduction Design
KEYS TO ... Testing differences between groups: t-tests
Basic Nonproject Comparison Group Design
Data Presentation for Nonproject Comparison Group Design
Advanced Nonproject Comparison Group Design

GUIDELINES FOR MANAGING THE EVALUATION PLAN

Maintaining quality control is essential to any evaluation plan. Provided below are activities to ensure that the evaluation is the best possible. Remember, too, that even if a professional evaluator has been hired, the responsibility of managing the evaluation plan is that of the program director.

1. Assess the adequacy of the evaluation design

Make sure that the design of the evaluation is valid, reliable, credible, and realistic before the start of the program. One approach for ensuring adequacy is to design a set of standards by which to review the evaluation design -- these standards would be based on the features of the evaluation that are most important to this site, for this program. Another approach is to follow the criteria developed by the Joint Committee for Standards for Educational Evaluation. This Joint Committee offers a comprehensive framework for developing standards in defining, designing, administering, collecting, analyzing, budgeting, contracting, reporting, and staffing an evaluation.

2. Monitor the practice of the evaluation design

Every good evaluation plan specifies evaluation activities that should be monitored to ascertain that the original design is implemented faithfully. Strategies to follow in monitoring evaluation practices include:

- Develop time frames to mark the milestones or dates on which products must be delivered and/or major activities must be concluded.
- Interview and observe key personnel to determine whether project activities conform to the approved evaluation plan.
- Ensure that the data collection efforts are carried out as planned by creating information checks. Train staff on proper test/assessment administration and data collection procedures. Create filing systems in which to store information as it is collected -- train staff to utilize these as well. Systematically check all data gathering activities.

3. Revise the evaluation design as needed

Unanticipated circumstances in a project's activities, or in the general school context, may require changes in an evaluation plan. Arrangements should be made for periodic examination of the original evaluation plan and for modifications as necessary. When making changes,

- Contact your project officer to ensure that the changes are approved by OBEMLA,
- Update key personnel, including the evaluator, regarding the approved changes in activities and timeline.
- Document the changes and include them in the annual performance report.

STUDENT ORAL LANGUAGE OBSERVATION MATRIX (SOLOM)

Purpose
The SOLOM is an informal rating tool that has proven a useful guide for teacher judgement of oral language proficiency as observed in a school setting. It can b used to determine English acquisition phase, diagnose student needs, and record the progress of individuals and groups. Some success has been reported in using the SOLOM to rate languages other than English.

Description
The SOLOM provides five scales for rating key dimensions of language proficiency. Each of these five scales may be rated from 1 to 5, yielding a total score range of from 5 to 25. The scales are

- ✳ Comprehension
- ✳ Fluency
- ✳ Vocabulary
- ✳ Pronunciation
- ✳ Grammar.

The SOLOM is not a standardized test, but has been used widely throughout California since about 1978 to supplement assessments garnered through standardized tests of language. Preliminary work is being conducted to standardize training for raters, and to ascertain the validity and reliability of the SOLOM. A one-hour training session is recommended for those who will use this instrument.

Administration
The SOLOM should be used by persons who are native speakers of the language (or who, at a minimum score at a level "4" in all categories of the language being assessed), and who are familiar with the student to be rated. Ideally, the classroom teacher will rate the English language proficiency of a student after several weeks of instruction. There is no test to be administered; rather, the teacher needs a few quiet moments to reflect on the language skill of a given student, and to select the description which most closely matches the current proficiency of that student. Based on the teacher's observation of the student, an "X" is indicated through the square in each category which best describes the students abilities.

A rating is immediately available, and can be used to group or regroup students for ESL lessons, to report student progress, or to guide refinements of instruction. Students scoring at a level ll"1" can be said to have no proficiency in the language.

Produced by the Evaluation Assistance Center-West/NMHU; Albuquerque, NM (10/95). Based on materials from the California Association for Bilingual Education dated 5/92

SOLOM Teacher Observation

Student's name ____________________ Observer's name ____________________

Language observed________________ Grade ______ Date Completed____________

	1	2	3	4	5
Comprehension	Cannot be said to understand even simple conversation.	Has great difficulty following what is said. Comprehends only "social conversation" spoken slowly & with frequent repetitions.	Understands most of what is said at slower-than-normal speed with repetition.	Understands nearly everything at normal speech although occasional repetition may necessary.	Understands everyday conversation l& normal classroom discussion without difficulty.
Fluency	Speech is so halting & fragmentary as to make conversation virtually impossible.	Usually hesitant; often forced into silence by language limitations.	Speech in everyday conversation & classroom discussion frequently disrupted by student's search for the correct manner of expression.	Speech in everyday conversation & classroom discussions generally fluent, with occasional lapses while student searches for the correct manner of expression.	Speech in everyday conversation & classroom discussions fluent & effortless, approximating that of a native speaker.
Vocabulary	Vocabulary limitations so extreme as to make conversation virtually impossible.	Misuse of words & very limited vocabulary; Comprehension quite difficult.	Student frequently uses the wrong words; conversation somewhat limited because of inadequate vocabulary.	Student occasionally uses inappropriate terms &/or must rephrase ideas because of lexical inadequacies.	Use of vocabulary & idioms approximate that of a native speaker.
Pronunciation	Pronunciation problems so severe as to make speech virtually unintelligible.	Very hard to understand because of pronunciation problems. Must frequently repeat in order to make self understood.	Pronunciation problems necessitate concentration on the part of the listener & occasionally lead to misunderstanding.	Always intelligible, though one is conscious of a definite accent & occasional inappropriate intonation patterns.	Pronunciation & intonation approximate that of a native speaker.
Grammar	Errors in grammar & word order so severe as to make speech virtually unintelligible.	Grammar & word-order errors make Comprehension difficult. Must often rephrase &/or restrict self to basic patterns.	Makes frequent errors of grammar & word-order which occasionally obscure meaning.	Occasionally makes grammatical &/or word-order errors which do not obscure meaning.	Grammatical usage & word order approximate that of a native speaker.

Produced by the Evaluation Assistance Center-West/NMHU; Albuquerque, NM (10/95). Based on materials from the California Association for Bilingual Education dated 5/92

STUDENT DATA SHEET FOR USE IN MULTI-YEAR EVALUATIONS

The fist academic year of the project was 19______; This is a □ 2-year, □ 3-year, or □ 5-year project.
For the purposes of evaluation,"Year" is defined as: □ 100 days □ 180 days □ units of instruction □ Other: ____________

Student Background

Name ____________ ID # ____________ Date of Birth ________ Age ________ Gender ________

Ethnic group ____________ Native language (L1) ____________ Home language ____________

US arrival date (if appropriate)________ District entry date________ SES (free lunch, partial, no subsidy)________

Last school year attended ________ Where attended ____________

Ever retained? □ Yes □ No Reason ____________

Special education? □ Yes □ No Process for referral ____________

Gifted/talented? □ Yes □ No Process for referral ____________

Student Assessment Data

	Pre-Entry			Year 1			Year 2			Year 3			Year 4			Year 5		
Tests (NRTs) / (Alternative) Assessments	Raw Score	NCE	%tile	Raw Score	NCE	%tile	Raw Score	NCE	%tile	Raw Score	NCE	%tile	Raw Score	NCE	%tile	Raw Score	NCE	%tile
Language Proficiency Tests																		
Language Proficiency Assessments																		
Academic Achievement Tests																		
Academic Achievement Assessments																		

Developed by the Evaluation Assistance Center-West/NMHU; Albuquerque, NM (Revised 12/95)

School Context

For IASA-funded programs, the relationship between this and other specially-funded projects must be described. As part of that, indicate below (1) whether the student ever participates in the program listed (yes/no) and (2) during which years of this project s/he is enrolled in the other program (e.g., the student might be enrolled in Even Start during Year 1 of the program and then move to Title I-Basic during Year 2 and Year 3).

Title I-Basic	□ Yes	□ No	If yes, □ Year 1	□ Year 2	□ Year 3	□ Year 4	□ Year 5
Title I-Even Start	□ Yes	□ No	If yes, □ Year 1	□ Year 2	□ Year 3	□ Year 4	□ Year 5
Title I-Migrant	□ Yes	□ No	If yes, □ Year 1	□ Year 2	□ Year 3	□ Year 4	□ Year 5
Title I-Delinq/Negl	□ Yes	□ No	If yes, □ Year 1	□ Year 2	□ Year 3	□ Year 4	□ Year 5
Title IV	□ Yes	□ No	If yes, □ Year 1	□ Year 2	□ Year 3	□ Year 4	□ Year 5
Title VII	□ Yes	□ No	If yes, □ Year 1	□ Year 2	□ Year 3	□ Year 4	□ Year 5
Other: ____________			□ Year 1	□ Year 2	□ Year 3	□ Year 4	□ Year 5

While participating in the project, was the student retained in a grade?

□ Yes □ No If yes, □ Year 1 □ Year 2 □ Year 3 □ Year 4 □ Year 5

While participating in the project, did the student drop-out of school?

□ Yes □ No If yes, □ Year 1 □ Year 2 □ Year 3 □ Year 4 □ Year 5

(For these two questions, consider the years that the project exists, not the years in which the student participated in the project.)

For Title VII only:

Reason for leaving program (check as many as appropriate):

□ Met exit criteria □ Transferred within district □ Transferred out of district

□ Parent pull-out □ Dropped out of school □ Other: ____________

For IASA programs:

Does student have a disability? □ Yes □ No

If yes, describe briefly ____________

MATCHING OBJECTIVES TO EVALUATION DESIGN

Within current Title VII evaluation guidelines, statistics are implied in only three areas: when comparing limited English proficient children and youth with non-limited English proficient children and youth with regard to (1) school retention, (2) academic achievement, and (3) gains in language proficiency. In addition, some statistics might be required based on how the project's objectives are written. Below are ways to analyze/report data based on how the objectives are written. Specifically, some example objectives for the typical components of a Title VII project are provided, with a suggestion of how they might be analyzed to determine whether they have been successfully met.

By the end of the project year, 80% of project students who attend classes will increase their comprehension, when reading aloud grade appropriate books, by 60% as measured by the classroom miscue assessment.

1. Define "attend class." Possible options include attending a minimum of 100 days in class, or a percentage of classes attended (e.g., 75% or "over half"). Do not analyze data for students who have not attended enough classes. Report the attendance criterion, and the number of students who did meet it.
2. Look at pretest and posttest data (for instance, beginning of school year and end of school year or spring 1995 and spring 1996 -- as appropriate). Determine how much each student increased his/her score. Make this a percentage (e.g., increased the score by 75%).
3. Report the number (and percentage) of students who increased at least 60% and those who did not increase at least 60%. You may want to be more definitive and report the number of students who gain various percentages in their comprehension scores.
4. Determine whether at least 80% of these students did gain at least 60% on their comprehension scores. If this is the case, the objective has been met; if fewer than 80% of the students gained 60% on their comprehension scores -- why?

Students who attend 80% of the project's English language development (ELD) classes will increase their English language achievement by an average of at least 2 NCEs each year on the CTBS reading subtest.

1. Look at the attendance records for project students. Include in this analysis only those who attended at least 80% of the ELD classes.
2. Collect these students' CTBS reading subtest NCE scores for last year and for this year (e.g., spring 1995 and spring 1996). Calculate the number of NCEs gained for each student. This information might be reported in a table.
3. Average the number of NCEs gained. Remember, this is based only on those who attended at least 80% of the ELD classes.
4. If the average gain is at least 2 NCEs, this objective has been met. If not, why?

Project student who attend 75% of math classes will have a higher average CTBS math subtest score than project students at the end of the year.

1. Look at attendance records. Collect CTBS math data only for students who attend at least 75% of their math classes.
2. Determine the nonproject comparison group: national norm, state, or local average on the CTBS math subtest; students at the same school not participating in the project; or students at a similar school not receiving Title VII funds. Collect data for the appropriate nonproject group.

3. Report the average score for project and nonproject groups. Which is higher? (No statistics are needed because the objective does not say "will be significantly higher [use the quasi-experimental *t*-test design]" or "the gap between the two groups will be lessened [gap reduction analysis]" or "scores of those in the program will be higher than a comparison group of individuals who are join the project [grade cohort].")

During each year of the project, staff will experience increased appreciation for the culture(s) and traditions of their students by attending more than half of the school cultural celebrations and events.

1. In the planning phases, be sure to develop sign-in sheets or another method to determine participation in school cultural celebrations and events.
2. Determine the number of cultural celebrations and events held at the school this year.
3. After each event, record information from sign-in sheets (or other method of determining participation) to determine which staff attended.
4. At the end of the year, report the number of events each staff person attended (e.g., 4 staff attended 50% of events, 2 staff attended 80% and 1 staff attended all events.) Is the number more than half? If so, the objective has been met; if not, why?

Parents of children in the project will attend at least 3 (of a possible 7) PAC meetings during the first project year.

1. During the planning phase, be sure to create a method of knowing who attends each PAC. This might be a sign-in list, a check-in process with a teacher, or some other method.
2. Determine the number of parents who attend each PAC meeting.
3. At the end of the year, determine how many PAC meetings each parent attended. Report the number of parents attending none, at least 1, at least 3, or more than 5 of the meetings.
4. Did all parents attend at least 3 meetings? If so, the objective was met; if not, why?

Teachers will develop curriculum materials for social studies regarding <u>the</u> <u>Long</u> <u>Walk</u> for the Navajo-speaking children during the first semester of the project.

1. Determine whether the social studies teachers did, indeed, create curriculum materials based on <u>the</u> <u>Long</u> <u>Walk</u>. (If not, why? Were some purchased instead? Was that unit eliminated from the curriculum?)
2. If there is materials available on <u>the</u> <u>Long</u> <u>Walk</u>, determine whether it is available in Navajo.
3. If the materials were available in Navajo at some point during the first semester of the project, the objective has been met.
4. You may choose to evaluate the effectiveness of this material, but it is not part of the objective. To evaluate effectiveness,
   - develop minimal criteria for the "goodness" of the materials,
   - ask other social studies teachers and/or parents to read the materials -- develop a rating sheet for their comments,
   - ask Navajo speakers to rate the Navajo used in the materials -- develop a rating sheet for their comments, and
   - ask students [perhaps even some outside the project] for their comments on the materials -- develop a rating sheet or an interview form.

   If the content of the materials and the Navajo language use is acceptable, based on the previously determined minimal criteria, then the materials are effective.

KEYS TO ... Testing differences between groups: GRADE COHORT

Among the regulations in the Education Department General Administrative Regulations (EDGAR) is the following:

§75.590(b) *How students are achieving State performance standards, including comparison with nonlimited English proficient students with regard to school retention, academic achievement, English proficiency, and native language proficiency, as appropriate.*

The usual method for obtaining a nonproject comparison group is to use students not currently in a Title VII project, state/district average scores, or standardized test norms. However, there are times none of these options is appropriate; for instance, when you have a highly mobile population or if you have a project with adult students who can choose to participate in particular portions of the program. In such cases the **grade cohort** design may be the best available. This design uses entry data from current and post students in the program as the comparison group; assessment data is collected on the same type of students at regular intervals before, during, and after the program. You may use the number correct, % correct, or a standard score (such as the NCE) to indicate the students' knowledge base, and what they have learned during the program. Do **NOT** use stanines, grade equivalents, or percentiles; these scores cannot be used for comparison purposes.

To use this design with a **mobile population**, analyses must be performed within categories of students; e.g., same grade level, same language group. As the program progresses, collect data as students enter the program and every 100 days. The aggregated scores of current and past students become the comparison group; the newer (current) students are the project group – compare new students' posttest scores against the group of pretest scores. For instance, as the 4th grade LEP students begin their reading lessons, give them a pretest on a standardized reading instrument. At the end of a 100-day learning period, posttest the students with the same reading test. As the students complete 100, 200, and then 300 days of instruction, these can be referred to as project "years." More specifically, you will be able to answer the following questions:

1. *Do students in the title VII program have higher scores than similar students just entering the Title VII program?* Comparing the scores of students who are just completing the program (based on 100 days) with the aggregated pretest scores of those students who have entered the program at some time in the recent past will allow you to answer this question.
2. *Do students who stay in the Title program longer (i.e., more "years" of attendance) have higher scores than those students who are entering and/or have been in the program less time?* This design can become a longitudinal design quite easily. A comparison can be made among new students, those who have had 100 days of instruction, 200 days of instruction, and so on.

To use this design with an **adult literacy program**, analyses must be performed within language group and level of ability. Collect data as adults enter the program and as they **complete instructional units.** The scores collected across time (i.e., all pretests for a specific instructional unit) become the comparison group; the newer adults entering that instructional unit are the project group – compare the adults' posttest scores against the larger group of pretest scores. For instance, as the LEP adults begin their writing lessons, give them a pretest. At the end of the first instructional unit (e.g., writing sentences), posttest the students on the same writing test. As the students complete various instructional units (e.g., writing sentences, simple stories, autobiographies), these then can be referred to as project "years." More specifically, you will be able to answer the following questions:

1. *Do adults in the literacy writing program have higher scores than similar adults just entering the literacy writing program?* Comparing the scores of students who are just completing the program (based on a specific instructional unit) with the aggregated pretest scores of those adults who have entered the same writing unit in the recent past will allow you to answer this question.
2. *Do adults who stay in the adult literacy program longer (i.e., more "years" of instructional units) have higher scores than those adults who are entering and/or have been in the program less time?* A comparison can be made among new students, those who have completed 10 writing units, 20 writing units, and so on. This visual representation of the results makes it clear whether the gap has become smaller or larger.

Using this method is one way to meet the EDGAR regulations. **Cautions**: (1) because you may have small numbers of students entering the program at a given time, the pretest and posttest groups of students may need to be built up over time. It may take 2 to 3 years before you can fully implement this design; (2) a traditional pre/post design does not meet the EDGAR regulations on evaluation. Because the grade cohort comparison group is made up of previous students as well as current students, this design does meet the Regulations; and (3) finding that the students have better scores on the posttest doe snot necessarily mean that the Title VII project **caused** this change – however, it is one factor to consider.

Modified from KEYS TO ... testing differences between groups: grade cohort *in* EAC-West news, *December 1991 by the Evaluation Assistance Center-West/NMHU, Albuquerque, NM* *Revised 10/95*

BASIC GRADE COHORT DESIGN

BACKGROUND OF THE DESIGN

The grade cohort design is a form of quasi-longitudinal design (that is, it can allow you to look at students over several years). It was developed in 1983 specifically for use with programs serving migrant populations. Since then, it has been modified somewhat to allow its use in any program with small numbers of students. It allows a "year" to be defined in various terms (e.g., 100 days of instruction, a particular unit of instruction), thus maintaining the largest number of students possible within the program for evaluation purposes.

QUESTIONS ANSWERED

- What kind of achievement gains are shown by students who have been in the Title VII program for one year as compared to students with similar characteristics who have not received Title VII services?
- How has children's achievement changed with on-going Title VII services (i.e., after 2 or 3 years of Title VII services)?
- How effective was this Title VII program (as opposed to no program) for students?

BASIC DESIGN REQUIREMENTS

1. All students identified for the program are given a pretest. Students may enter at different times during the year, so long as each is given the same pretest.

2. Baseline data for each grade level consists of the pretest scores of students entering the program at that grade level, regardless of what time of year they entered or what year they entered.

3. Students are given a posttest at periodic intervals -- every 12 months, every 100 days they complete within Title VII, or after a given instructional unit (e.g., math [addition, subtraction, multiplication, and division] with single digit numbers). These posttests are the pretest for the next interval of education.

4. Data from pre- and posttests should be collected by grade level by language group by number of years within the project. A one-year evaluation would consist of comparing students who have completed one grade level (e.g., 4th) with the pretest scores of students who entered that grade with no prior Title VII experience.

5. Data can be compared cross-sectionally (e.g., beginning 3rd grade to beginning 4th grade) or longitudinally (e.g., beginning 3rd grade to beginning 6th grade). Data also can be compared to determine the effects of different amounts of Title VII experience. For example, comparing 3rd grade students with no previous Title VII experience to 3rd grade students who have had 2 years of Title VII experience.

6. It may be necessary to collect data across 2 or more years before there is enough to allow statistical comparisons. The more sophisticated the questions asked, the longer it may take to collect the data.

7. Data from the past may be collected from the files of current students. For instance, if the program is serving children in grades 3-5, data may be collected about those children's achievement in K-2 as well.

MAKING SENSE OF THE DATA

- Collect scores for students by grade level by language group by experience within the Title VII program. The scores should be NCEs (normal curve equivalents), percent correct, raw score correct, or a form of standard score (Scale Scores, Standard Scores, etc.). *Do not use percentiles, stanines, or grade equivalents*; these scores cannot be used in calculations.

- Calculate average scores for each category of students (3rd grade Spanish-speaking in Title VII for first year, or 5th grade Mandarin-speaking in Title VII for second year). Present the data in a table, such as Example Table 1, that includes the number of students in each group.

- Calculate a comparison between the Title VII students and the nonproject comparison group (the pretest scores for all students who have ever entered that group). A simple method is to create a figure such as Example Figure 1 which shows data for both program and nonproject comparison group children for 4 years (1 before the program began, then 3 years' of project data from Example Table 1).

- Look at the change in the students from entering the program to the completion of one year, two years, and so on. These numbers can be seen in Example Table 1; Example Figure 1 provides a visual illustration of this comparison.

ADVANTAGES OF THE GRADE COHORT DESIGN

1. The design meets the Title VII evaluation requirements. It requires annual testing (although "annual" can be defined in different ways) and includes a nonproject comparison group.

2. The design eliminates some of the reasons not related to the project that children might increase their achievement scores. For instance, the same assessments must be used across time -- this eliminates the problems that might occur by changing instruments (no instrument is exactly equivalent to another instrument). Also, because of the various definitions that can be used for "annual" testing, fewer students are lost to the program, thus increasing the number of students included in the evaluation.

3. Because of the way in which the nonproject comparison group is created, it is not necessary to use national or statewide norm groups as the nonproject comparison group. This improves the validity of the evaluation.

4. Small groups of students are less problem with this design. Because of the way students' scores can be aggregated across years, it is possible to have a complete and valid evaluation with sufficient students to make the results generalizable to other situations.

DISADVANTAGES OF THE GRADE COHORT DESIGN

1. It may take some time to identify and accumulate scores for enough students at each grade level. Thus for at least one year, and possibly two, the evaluation will be incomplete. Graphs and tables can be presented, but few statistics are appropriate when the groups of students are very small.

2. The design does not eliminate all possible "other" explanations of the increased achievement scores. Students get older (and know more) and they come to you from many different backgrounds. They are selected for the program based on certain criteria, some of which may explain the changes in achievement level of the students.

ANALYZING GRADE COHORT DATA: Tables and Graphs, Summarizing

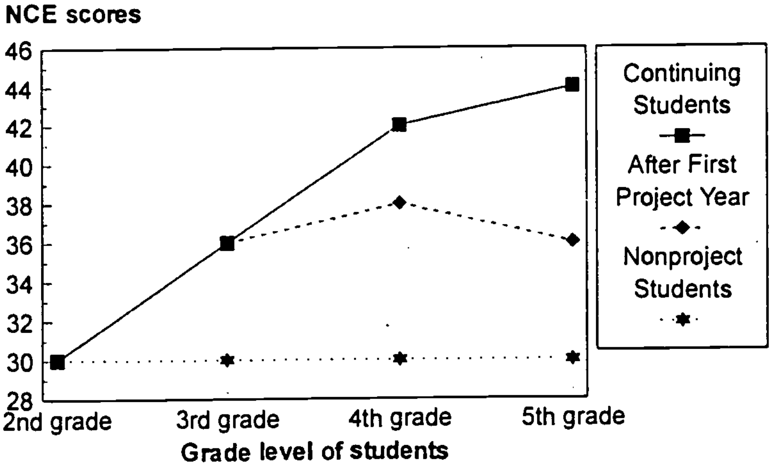
The easiest way to analyze the data for a grade cohort design is to compare the average scores of students before and after their Title VII experience. For instance, looking at Example Table 1 and Example Figures 1 and 2, it is easy to see that the longer the students stayed in the program, the greater their increase in achievement. If possible, it would be appropriate to collect the same set of scores for students the year after they finish the Title VII program. In this example, collect the same scores in 6th grade of students who began the program in the 3rd grade and stayed with the program through theth5 grade. This would allow the comparison of their "growth curve" with and without the program. It might be expected that their curve would flatten out (i.e., their scores would be higher than 5th grade, but would not increase at the same rate as their 3rd-5th grade increases).

DATA PRESENTATION FOR GRADE COHORT DESIGN

Example Table 1.
Average NCEs (and number of students) on the CTBS Math Achievement Subtest by Language Group by Grade Level, and by Year in Program

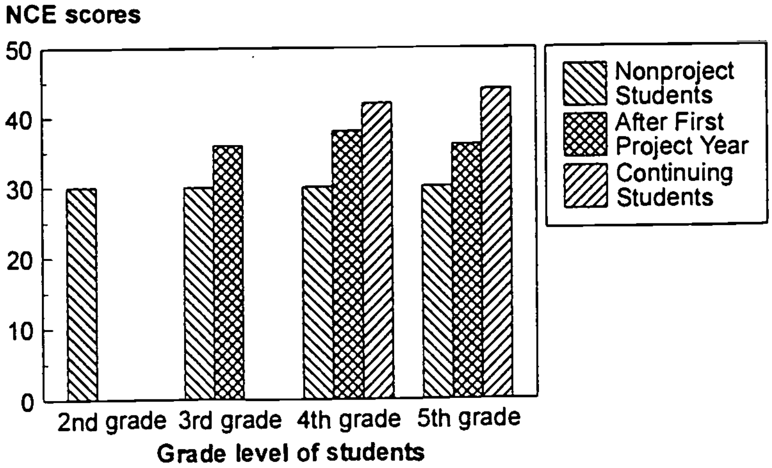
Home language → Grade level & program year ↓	Spanish		Ukraine		Lao		All students	
	NCE	(#)	NCE	(#)	NCE	(#)	NCE	(#)
3rd grade								
1st year	36	(25)	38	(10)	35	(19)	36	(54)
4th grade								
1st year	38	(15)	35	(9)	37	(22)	37	(46)
2nd year	42	(12)	39	(8)	40	(12)	40½	(32)
5th grade								
1st year	36	(18)	32	(15)	30	(20)	33	(53)
2nd year	39	(14)	37	(9)	38	(20)	38	(43)
3rd year	44	(12)	43	(8)	40	(11)	42	(31)

Example Figure 1.
Comparison of Spanish-speaking students' (Program & Nonproject) average Math Subtest NCEs Grade 2 through Grade 5:
Line graph.



Nonproject students are those who are just entering the program.

Example Figure 2.
Comparison of Spanish-speaking students' (Program & Nonproject) average Math Subtest NCEs Grade 2 through Grade 5:
Bar graph.



Nonproject students are those who are just entering the program.

ADVANCED GRADE COHORT DESIGN

Note: Analyses are suggested but not required for this design. You may wish to read "Interpreting Graphs & Tables" and only the introductory paragraphs under "Statistical Analyses." For those who want to do statistical analyses, or who want their evaluators to do statistical analyses, the analyses are presented in the ▸ paragraphs below.

INTERPRETING GRAPHS AND TABLES

Data presented in tables and graphs should be self-evident. That is, little verbal or written explanation should be needed. However, summarizing tables and graphs within the text of the annual report is helpful. From Example Table 1 and Example Figures 1 and 2, the following text might be written.

> *Example Table 1 presents the standardized NRT math subtest data (in Normal Curve Equivalents) collected on all students involved in the project from 1991-1994. The students were from three language groups: Spanish, Ukraine, and Lao. Data are presented by language group within each grade, and for all students within each grade. Additionally, the data are presented based on the number of years the students had been in the Title VII program. For instance, there were two groups of fourth grade students: (1) those who had completed both 3rd and 4th grade within the project and (2) those who began the program in 4th grade. The data show that the average scores increased during each year that the student was in the project. Example Figure 1 shows the same data, with pretest information added, for the 5th grade Spanish-speaking students who had been in the program for all three years, and the data for the Spanish-speaking students who completed one year of the program at each grade level. Again, continued time in the program resulted in increased achievement on the part of the students.*

ANALYZING GRADE COHORT DATA: Statistical Analyses

In addition, statistics can be done on the average scores. This is when it will be especially important to have enough students within each group. While tables and graphs can be prepared with as few as 10 students in each group, it is not appropriate to do statistics on any fewer than 15-20 students per group; the greater the number of students, the better chance of finding a statistically significant increase in the students' scores. Grade cohort is helpful here because of its use of (1) smaller time periods defined as a "year" and (2) adding students to groups across various years as long as the same pre- and posttest assessments are used.

Four types of procedures could be used with the data:

1. Compare the test scores of one group of students as they proceed through the program. This measures the progress of the students who began the program in the 3rd grade and continued through the 5th grade (looking at Example Table 1: the third grade row, the fourth grade second year row, and the fifth grade third year row). This is a longitudinal analysis, looking at one group of students across a period of time.

    ▸ The best type of analysis would be a regression or analysis of variance that looks at whether there is a statistically important [significant] increase from one year to the next and whether this increase is a straight line or curves in some way (trend analysis). One multivariate analysis could be done that included all language groups separately, or several analyses could be done that involved one language group at a time.

    ▸ More simply, a dependent t-test can be done comparing each grade level's score to the next grade level's score. This technique will determine whether each year's average score was statistically greater than the previous year's average score.

2. Compare the 3rd grade test scores for students entering the program to the test scores for 3rd grade students who have been in the program for a year (looking at Example Figure 1: compare the bottom line with the second line). A comparison like this could be made at each grade level. This is a test of how well students do when they have had a year of program instruction as compared to students in the same grade and language group who have not had any program instruction. This is a cross-sectional type of analysis.

- The best statistical analysis for this would be an analysis of covariance that would consider the pretest and posttest data for each language group in one analysis. This would allow a clean determination of how much each group gained from pretest to posttest.

- More simply, a series of t-tests could be completed; one t-test for each language group, comparing each pretest with each posttest. If the same students were in the pretest group and the posttest group, the dependent t-test would be used; if the groups involve different students [e.g., this year's incoming 3rd graders compared to those who have completed the 3 rdgrade year of the program], then the independent t-test would be appropriate.

3. Compare those who have completed one year of the program with those who have completed two years, with those who have completed 3 years of the program (e.g., use 5th graders only, compare the scores in the last row of Example Table 1). This will allow a comparison of the effectiveness of the full program with a "partial" program. In effect, this would allow the program to assure parents that although their children's scores increased with one year of the program, that the students should stay in the program in order to receive the full benefits.

   - The best analysis for this comparison would be an analysis of variance that would allow the comparison of all years' data simultaneously. This is appropriate because different students would be in each year of the program being tested.

   - Alternatively, a series of independent t-tests could be done comparing each pair of grade levels (e.g., comparing third grade first year with third grade second year, then third grade second year with third grade third year).

4. Collect the same data for one year after the program is completed. Then compare students' average scores from the year before they entered the program, all through the program, and then for the year after the program. This may provide more evidence for the effectiveness of the program. Ideally, the analysis (and graph that could be done) will show that during the program, achievement progressed more rapidly than either before or after the program. These data also could be compared to the pretest scores of students who initially entered the program at each grade level. Then one line of the graph would show the scores of students before they entered the program and could be considered as the nonproject comparison group (see the lowest line, Example Figure 1). The second line will present the achievement of students before, during, and after the program (see the highest line, Example Figure 1 -- although "after" is not presented). The graph should show that students' in the program progressed more rapidly than the nonproject comparison group and that the students' achievement progressed more rapidly during the program than either before or after the program.

   - The best method for analyzing the data would be through a multivariate multiple regression -- this would allow the analysis of all language groups, project and nonproject data, in one analysis. Also, it would allow the determination of whether students learned in a predictable manner (a straight line), or in a "bumpier" manner that curves more strikingly at certain times (trend analysis).

   - The data also could be analyzed for each language group separately, with a multiple regression or an analysis of variance.

   - More simply, a series of independent and dependent t-tests could be utilized. Independent ts would be used to compare project and nonproject students at each grade level; the dependent ts would be used to compare within the project group and within the nonproject group.

KEYS TO ... Testing differences between groups: GAP REDUCTION

Among the regulations in the Education Department General Administrative Regulations (EDGAR) is the following:

§75.590(b) *How students are achieving State performance standards, including comparison with nonlimited English proficient students with regard to school retention, academic achievement, English proficiency, and native language proficiency, as appropriate.*

Frequently it is difficult to find an appropriate nonproject comparison group. One suggestion is to use a norm group. This might be the group of students used by the test developer and described in the testing manual of the instrument you plan to use. If this group of students is similar to your title VII project students; e.g., both groups include Navajo and Korean students), then the national norm is an appropriate comparison group. It also is appropriate to use the national norm group if you want to compare your students against mainstream, majority-culture students. You might also consider using the average score for your state or district, the cores from a school similar to yours that does not received Title VII funding, or your whole school if you have a Title VII comprehensive school-wide grant. Now, what data should you collect from the tests?

1. Be sure you have pretest scores and posttest scores for all students, preferably from the same test administration date (e.g., spring 1995 pretest scores and spring 1996 posttest scores).
2. If you have the same test and the same test administration dates, you may collect raw scores. However, standard scores, especially NCEs, and scaled scores frequently are much easier to use. Do **not use** percentile, stanines, or grade equivalents.
3. If you are using a national norm. for your comparison group, you should use NCEs. Collect data for your students in NCE form, then compare them against the national norm average of 50.00, with a standard deviation of 21.06.

If you are not worried about actual statistical significance (i.e., whether the difference between two scores differ according to statistical formulae), you may want to consider the **gap reduction technique**. The gap reduction technique determines the amount of growth in both the comparison/norm group and the title VII project group, then calculates whether the gap between the scores of the two groups has lessened across the academic year. There are not actual statistics and there are no "rules" for determining whether the gap has been reduced sufficiently to claim success for the project students. However, the graphic presentation of the data can provide a powerful visual statement. More specifically, you will be able to answer the following sets of questions.

1. *What was the gap between the average pretest score for the comparison group and the average pretest score for the project group at the beginning of the year?* and *What was the gap between the average posttest scores at the end of the school year?* By subtracting the project average score from the comparison average score (pretests, then posttests), you will know how large the gap was at both the beginning of the project and at the end of the school year. You can use the scores from the test, or you can "standardize" each score (divide the group's average by their standard deviation.

2. *Has the gap between the performance of the two groups been reduced across the school year?* By subtracting the posttest gap from the pretest gap, you will know how much the gap has been reduced. If this number if positive (e.g., pretest gap of 10 - posttest gap of 4 = gap reduction of 6), the project students' performance has become closer to the comparison group's performance; the gap has been reduced. If this number is negative (e.g., pretest gap of 10 - posttest gap of 15 = gap reduction of -5), the project students' performance has become further away from the comparison group's performance; the gap has increased.
3. *What does the students' performance look like in a visual representation of the gap reduction?* Graph the pretest and posttest average scores for the project group and the comparison group. This visual representation of the results makes it clear whether the gap has become smaller or larger.

Using this method is one way to meet the EDGAR regulations. **Caution**: finding that the gap has been reduced across the school year does not necessarily mean that the Title VII project has **caused** this change; it is a factor to consider.

dified from KEYS TO ... testing differences between groups: gap reduction in <u>*EAC-West news*</u>, *December 1991 by the Evaluation istance Center-West/NMHU, Albuquerque, NM* *Revised 10/95*

GAP REDUCTION DESIGN

BACKGROUND OF THE DESIGN

The gap reduction design was developed to provide an easy-to-use technique that does not require a "live" comparison group of students at the school who are similar to the program students. The gap reduction design is based on comparing program students to the national norm or district average on a standardized test. Since NCE scores are recommended, the national norm is a constant 50.0 NCEs.

QUESTIONS ANSWERED

- Has the difference (gap) between the performance of the program students and the national norm of 50.00 NCEs been reduced across the school year?

- In the process of answering the above question, two others also are answered:
  - What is the gap between the performance of the program students and the national norm at the beginning of the year?
  - What is the gap between the performance of the program students and the national norm at the end of the year?

BASIC DESIGN REQUIREMENTS

1. The design originally was devised to utilize the national norm from an NRT. As a modification, the criterion on a CRT or alternative assessment could be used (e.g., compare the average % correct of the program students against the criterion of 85% correct for "mastery" of the topic) <u>or</u> a district or school average score can be used.

2. If using an NRT, the program students should be tested at the same time as the norm group was tested. If using a live comparison group, both the program and the comparison students should be tested at about the same time.

3. The same test should be used for pre- and posttest testing.

4. Data from pre- and posttest for the program group should be collected by grade level by language group/proficiency level by other appropriate demographic information.

5. A minimum of 20-25 students should be in each program subgroup of students.

MAKING SENSE OF THE DATA

- At a minimum, collect scores for students by grade level by language group by proficiency level. The original gap reduction design called for NCEs (normal curve equivalents) on NRTs. Other scores and types of tests now are used (see point 1 above). *Do not use percentiles, stanines, or grade equivalents.*

- Note that data might not be test scores in some cases. For instance, the gap reduction technique can be used to determine whether absenteeism has decreased (compare last year's numbers with this year's <u>or</u> compare average number of days absent for the school *vs* program students), number of library books checked out has increased, and so on.

- Calculate the pretest gap: comparison group's average score - program group's average score.

- Calculate the posttest gap: comparison group's average score - program group's average score.

- Calculate the gap reduction: pretest gap - posttest gap.

- A positive outcome (i.e., gap reduction = a positive number) indicates that the gap has been reduced; the program students' average score is more like that of the comparison group (or national norm). A negative outcome (i.e., gap reduction = a negative number) indicates that the gap actually has increased; the program students' average score is even less like the comparison group's score. A gap reduction of zero indicates that the project and nonproject students' average scores have remained the same distance apart. While the project students are not gaining on their peers, they are not losing ground either.

ADVANTAGES OF THE GAP REDUCTION DESIGN

- A number of different "live" and norm group comparisons are possible.
- The design meets the title VII (and other specific program) evaluation requirements. It includes a nonproject comparison group.
- The design is easy to use. Simple subtraction is the only "analysis" necessary.
- The design can be used with test data as well as nontest data.

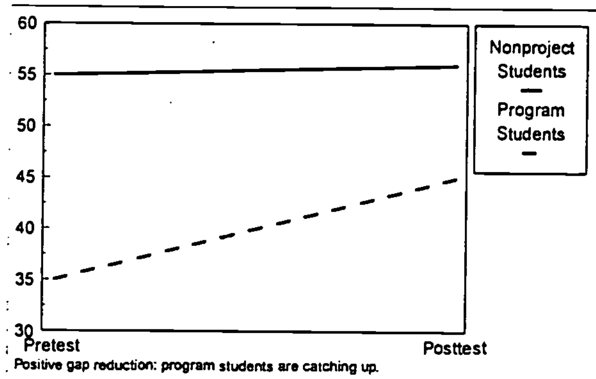
DISADVANTAGES OF THE GAP REDUCTION DESIGN

- There is no standard for what constitutes a "good" amount of gap reduction.
- Depending upon the particular data (e.g., number of students, how variable the scores are, how well the scores are distributed on a normal curve), the results may vary from what is anticipated and from what other ways of analyzing the same data may find.
- The amount of gap reduction has little meaning by itself -- average scores, standard deviations, and other information must be provided as well.
- The gap reduction technique controls for very few problems that might impact the validity of the results (e.g., history of the students, maturation, regression toward the mean).

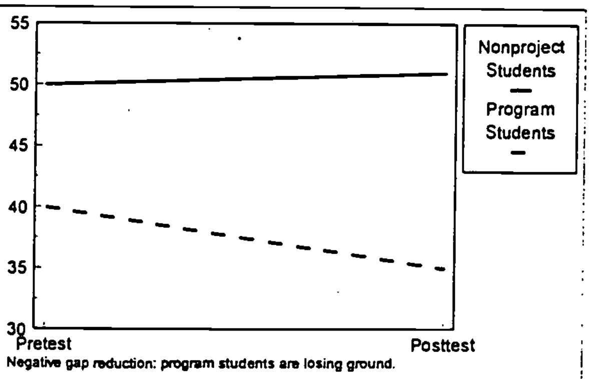
ANALYZING GAP REDUCTION DATA: Data presentation

The analyses for the gap reduction design were described above. No other analyses are possible. (Early versions of the design suggested calculating the "RGI" (relative gap index) which was a percentage that students had grown. This is no longer suggested because of various problems in its use and interpretation.)

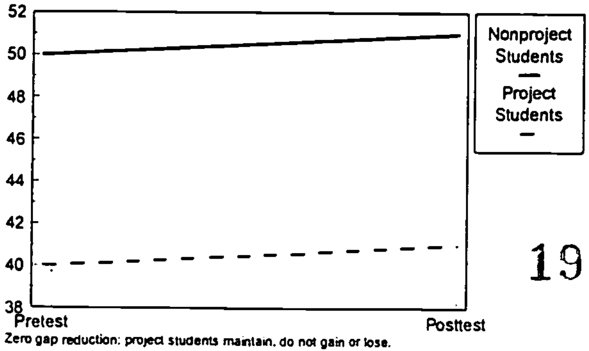
The best way to present gap reduction data is through a graph -- a visual presentation of the results. Three example graphs appear below: a gap reduction, a gap increase, and a gap showing no change. Further information is provided in the Data Presentation for Gap Reduction Design.



Positive gap reduction: program students are catching up.



Negative gap reduction: program students are losing ground.



Zero gap reduction: project students maintain, do not gain or lose.

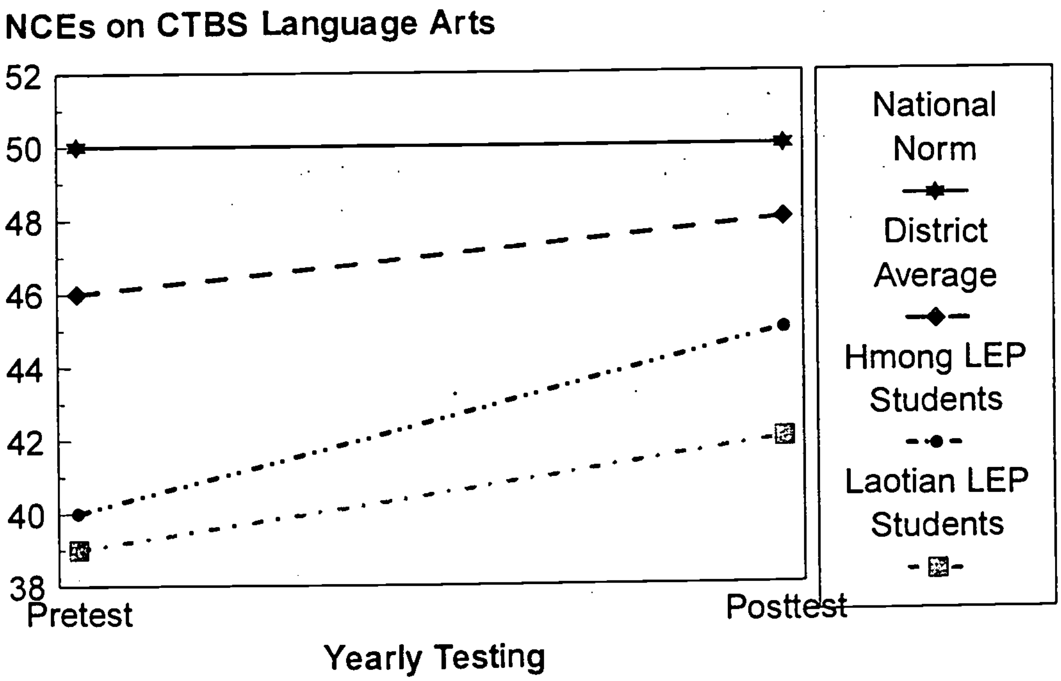
DATA PRESENTATION FOR GAP REDUCTION DESIGN

Example Table 1.
Average NCEs (and number of students) on the 5th grade CTBS language Arts Subtest

	Pretest		Posttest		Gap Reduction
	Score	Gap	Score	Gap	
Laotian LEP Students (20 students)	39	NN: 11 DA: 7	42	NN: 8 DA: 3	NN: 3 DA: 1
Hmong LEP Students (25 students)	40	NN: 10 DA: 6	45	NN: 5 DA: 3	NN: 5 DA: 3
District Average (DA) (500 students)	46	NN: 4	48	NN: 2	NN: 2
National Norm (NN)	50		50		

Note: Comparison of two LEP groups against the district average and the national norm indicates positive gap reductions; comparison of the district average against the national norm indicates a positive gap reduction.

Example Figure 1.
Gap Reduction Analysis of program students, district average scores, and the national norm



KEYS TO ... Testing differences between groups: t-TESTS

Among the regulations in the Education Department General Administrative Regulations (EDGAR) is the following:

> §75.590(b) *How students are achieving State performance standards, including comparison with nonlimited English proficient students with regard to school retention, academic achievement, English proficiency, and native language proficiency, as appropriate.*

The *nonlimited English proficient students* can be defined in several ways. One suggestion is to use a norm group comparison instead. This might be the group used by the test developer and described in the testing manual of the instrument you plan to use. This group of students is appropriate if they "look like" your students; i.e., if your students are Latino and Japanese, you would like to find that the test developer included Latinos and Japanese among the students in the norm group. Another possibility is to use the average score from all the students in your state. For instance, if your students took the test in spring 1995, use the average score for your state on the spring 1995 administration of this test. Even better, use the average score from your area of the state, your school \district, or from a school you know of that is similar to your school, but not receiving Title VII funds. If your have a comprehensive school program, the average score from all students in your school could be used as the comparison group. Of course, you still may use the national norm group if you want to compare your students' achievement against that of mainstream, majority-culture students. Now that you've identified a comparison group ("live" or "norm"), what do you do with them?

1. Be sure you have pretest scores and posttest scores, preferably from the same test administration date (e.g., spring 1995 pretest scores and spring 1996 posttest scores).

2. If you have the same test and the same test administration dates, you may collect raw scores. However, if you cannot use the same test administration dates (for instance, the national norm group was tested in March and you plan to test in May), or if the test has changed, use standard scores (preferably NCEs) or scaled scores.

3. Do not use percentiles, stanines, or grade equivalents. Points along these scales have different meanings based on the actual score. For example, the difference between percentile scores of 1 and 5 is not the same as the difference between percentile scores of 41 and 45, even though each one appears to show 4 "points" of improvement.

One of the easiest types of statistical tests to do, report, and interpret is the *t*-test. The ***t*-test** is used to determine the difference between two average scores: (1) between a pretest and a posttest to determine whether the group of students has, on the average, improved during a year or (2) between two pretest scores to determine whether the two groups differed in their average scores at the time of the pretest. If the *t*-test is significant, one of the two average scores is statistically better than the other; if the *t*-test is not significant, the two average scores are statistically equal even though the actual values may be somewhat different. More specifically, you need to answer both the following sets of questions.

1. *Are the pretest average scores of the comparison group and the Title VII project group different?* and *Are the posttest average scores of the comparison group and the Title VII project group different?* This will tell you whether or not the groups began (or ended) this school year with different average scores. Ideally, the scores should be about the same for both groups, but realistically this doesn't usually happen. Use the independent *t*-test to test your students against a local or state comparison; use the one-sample *t*-test to test your students against a larger state or national norm group.

2. *Did the Title VII project group improve their average scores during the school year; i.e., is the posttest average score different from the pretest average score?* The answer to this question will tell you whether or not your title VII group of students changed their average achievement level during the school year. We would like the score of the posttest to be better than the pretest and usually these scores do show improvement. Use the dependent *t*-test to test whether this group of students' pre-to-posttest scores are different.

Using this information is one way to meet the EDGAR regulations. **Caution**: finding that the students had better cores on the posttest does not necessarily mean that the Title VII project **caused** this change; however, it is one factor to consider. Also, you must be careful when interpreting differences between your students and a norm group – they may be very different students.

lified from KEYS TO ... testing differences between groups: t-tests *in* EAC-West news, *December 1991 by the Evaluation Assistance* *nter-West/NMHU, Albuquerque, NM* *Revised 10/95*

BASIC NONPROJECT COMPARISON GROUP DESIGN

BACKGROUND OF THE DESIGN

The nonproject (or non-equivalent) comparison group design is a form of experimental or quasi-experimental design (that is, it requires a control that is similar to [or, ideally, exactly like] the students in the program being evaluated). In its simplest form, the design allows the comparison of the two groups of students before the program starts and at the end of the program; in addition, the growth of each group should be monitored from preprogram to postprogram.

QUESTIONS ANSWERED

- Are there differences between the pretest average scores of program students and nonproject comparison group students (i.e., are they different when the program begins)?

- Are there differences between the posttest average scores of program students and nonproject comparison group students (i.e., are they different when the program ends)?

- Did the program students' average scores increase significantly from pretest to posttest (i.e., do they show achievement gains)?

- Did the nonproject comparison group's average scores increase significantly from pretest to posttest (i.e., do they show achievement gains)?

BASIC DESIGN REQUIREMENTS

1. The students identified for the program and the non-project comparison groups are as similar as possible: same language(s), ethnic/racial groups, language proficiency, grade level, SES, and so on.

2. The two groups are pre- and posttested at about the same time.

3. The same test should be used for each group. If this is not possible, standard scores (e.g., NCEs) might be used for comparison purposes.

4. Data from pre- and posttest should be collected by grade level by language group by other appropriate demographic information.

5. A <u>minimum</u> of 10 to 15 students should be in each subgroup that will be analyzed (e.g., Spanish home language, LEP, 3rd grade girls). More subjects are preferable.

6. Pretest average scores should not be significantly (or substantially) different between the two groups of students.

MAKING SENSE OF THE DATA

- At a minimum, collect scores for students by grade level by language group by language proficiency. The scores should be NCEs (normal curve equivalents), percent correct, raw score correct, or a form of standard score (Scale Scores, Standard Scores, etc.). *Do not use percentiles, stanines, or grade equivalents*; these scores cannot be used in calculations.

- Calculate average scores for each subgroup of students (such as described in point 5 above). Present the data in a table, such as the Example Table, that includes the number of students in each group.

- Calculate a comparison between the Title VII students and the nonproject comparison group. Create a graph that visually demonstrates the achievement levels of each group; see the Example Figure.

- Perform statistical analyses to answer the four QUESTIONS listed above. Ideally, the answer to the first question should be "no, the two groups are not significantly different;" the answer to the second questions should be "yes, the program group is performing better than the nonproject comparison group;" and the answer to the third question should be "yes, the program group's average posttest score is higher than their pretest average score." The answers to the other question is less important.

ADVANTAGES OF THE NONPROJECT COMPARISON GROUP DESIGN

1. The design meets the Title VII (and other specific program) evaluation requirements. It includes a nonproject comparison group.

2. The design eliminates some of the reasons not related to the project that children might increase their achievement scores. For instance, the same assessments must be used across time -- this eliminates the problems that might occur by changing instruments.

3. The design controls for problems such as history of the children, maturation of the children, and mortality -- all because the nonproject comparison group is similar to the program students.

DISADVANTAGES OF THE NONPROJECT COMPARISON GROUP DESIGN

1. It can be difficult to locate a nonproject comparison group that is similar to the program group, especially for programs that serve special groups such as culturally and linguistically diverse students.

2. Costs can be relatively higher than with other designs because of the number of students who need to be tested.

3. The design does not control for all problems that may occur. Thus there may be other reasons for the differences found (or not found) between the two groups of students.

ANALYZING NONPROJECT COMPARISON GROUP DATA: Analyses, data presentation, and summarizing

The easiest way to analyze the data for a grade cohort design is to compare the average scores of students through the use of multiple *t*-tests. Use dependent *t*-tests to analyze the pre-post data of the program students (look for significance) and the pre-post data of the nonproject comparison group students. Use independent *t*-tests to analyze the pre-pre data of the two groups (look for nonsignificance) and to analyze the post-post data of the two groups.

In addition to the statistical analyses, tables and graphs should be prepared. See the example table and figure in Data Presentation for NonEquivalent Comparison Group Design. These show the numerical data and its visual representation for a hypothetical group of 8th grade students. From these it is obvious that both groups of students' achievement increased across the school year from pre- to posttest. However, it also can be seen that the program students' knowledge increased at a greater rate that the NONPROJECT students'. The fact that the pretest average scores of the two groups is similar indicates that they were, indeed, similar students.

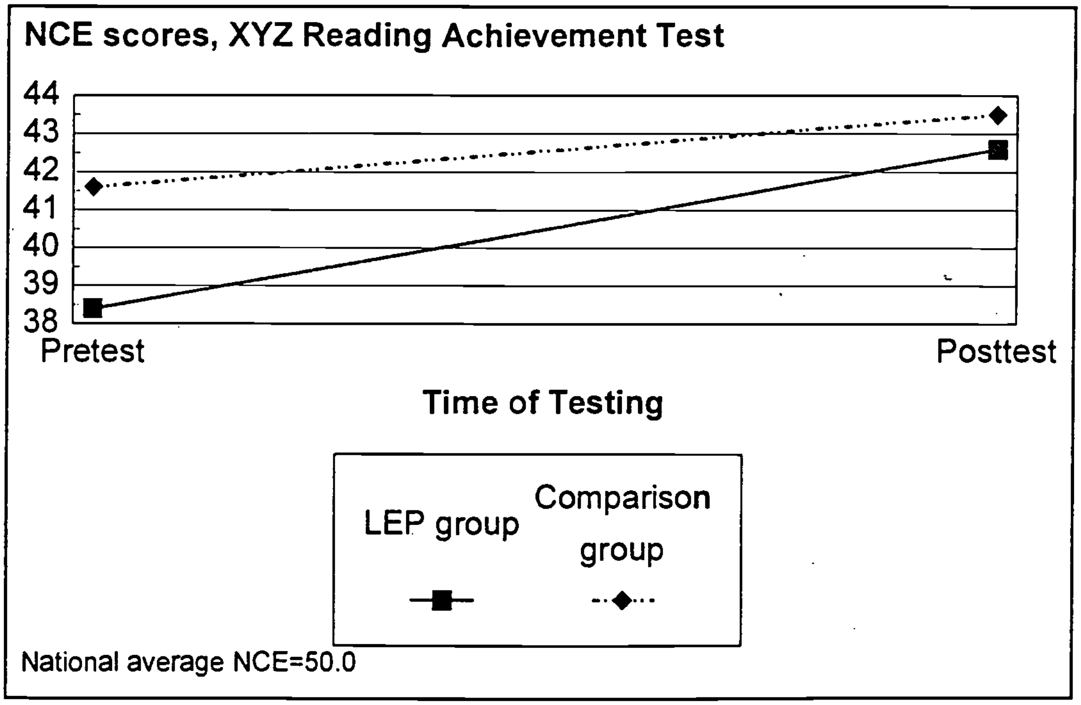
DATA PRESENTATION FOR NONEQUIVALENT COMPARISON GROUP DESIGN

Example Table 1.
Average scores (and standard deviations) for 8th grade Russian-speaking LEP students and district-wide comparison group

	# of Students	Pretest Average NCE scores (& SD)	Posttest Average NCE scores (& SD)
Program students	20	38.4 (5.6)	42.6 (4.5)
Comparison group	50	41.6 (7.3)	43.5 (6.9)

For pretests: t = 1.85, nonsignificant
For posttests: t = 1.96, nonsignificant
For program students, pre-post: t = 2.55, significant
For comparison students, pre-post: t = 2.01, significant

Example Figure 1.
Average pre- and posttest scores for 8th grade Russian-speaking LEP students and district-wide comparison group



ADVANCED NONEQUIVALENT COMPARISON GROUP DESIGN

Note: Various analyses are possible for this design. None is "best" in all cases. The various factors to be included in the analyses will determine which analysis should be done in any given evaluation. You may wish to read only the introductory paragraphs of this handout, then to share it with the program evaluator. Analyses are presented in the ▸ paragraphs below.

FACTORS TO CONSIDER WHEN ANALYZING DATA

Statistics are calculated on average scores for groups of students. The greater the number of students, the better the chance of finding a significant difference between their pretest scores and their posttest scores -- indicating that their scores have increased (we hope they have not decreased!) significantly across the year of the program. The basic requirement for a program evaluation is that a group of program students be compared with a group of students not in the program. If this is the only comparison planned, the *t*-test described in "Basic NonEquivalent Comparison Group Design" is appropriate. However, it may be helpful to disaggregate the program students into smaller groups of interest -- boys and girls, different grade levels, different language groups, and different language proficiencies. This would allow a determination regarding the efficacy of the program for all types of students. In addition, it may be of interest to analyze several scores simultaneously (e.g., the four modalities of language proficiency) or to collect longitudinal data (across several years, rather than just one year). Analyses for these options are described briefly below.

COMPARISONS TO ANALYZE

1. Compare the test scores of students, disaggregating the data by gender, English language proficiency levels (including native English speakers), handicapping condition, grade, and language/ethnic group (these types of data are required within IASA Title VII and Title I programs). There should be at least 10 students in each group (a group might be the 3rd grade LEP non-handicapped girls whose home language is French). This comparison would allow a careful determination of the effectiveness of the program for each group of students. If the program is differentially effective, modifications in the educational program might be necessary.

   ▸ If all independent variables are categorical or semi-continuous (not continuous such as age in months), an analysis of variance (ANOVA) will allow a determination of whether there are significant differences among the various groups. If several grade levels, or other semi-continuous variables, are included, it might be important to do a trend analysis to see if the changes in scores follow a specific type of pattern (linear increase, up-and-down, or another). Of course, if several related scores are to be analyzed simultaneously (e.g., reading, writing, and oral proficiency scores -- nonrelated constructs such as mathematics and oral proficiency should be analyzed separately), a multivariate analysis of variance (MANOVA) should be computed.

   ▸ If some or all independent variables are continuous, multiple regression should be used. It usually is preferable to leave continuous variables in their raw form, rather than to categorize scores into arbitrary groupings. If multiple dependent variables (i.e., related scores of some type) are to be analyzed, multivariate multiple regression should be utilized.

   ▸ If both pretest(s) and posttest(s) are to be analyzed, two options are available, depending upon the information that is sought. (1) If the purpose is to show that students' scores changed across time (i.e., from the pretest to the posttest), a type of repeated measures analysis should be used. Most typically these will be the within subjects analysis (repeated measures) if only the pretest and posttest will be analyzed. If independent variables are to be analyzed as well, split-plot factorial analyses should be utilized. (2) If the purpose is to look more closely at the changes in student achievement or proficiency since the pretest was administered, the analysis of covariance is most appropriate. This analysis can be used within an ANOVA, a multiple regression, or various multiple dependent measure designs.

2. Compare students longitudinally, looking across 2, 3, or more years of data at the same time. This would allow a determination of how well the program "works" across time -- do some portions of the program work better than others? do students continue to progress in a linear fashion, or does their progress move

up and down? do several years of the program have a stronger effect than merely adding the effects of individual years of the program?

- Most appropriately a time-series type of analysis should be conducted. However, it is rare that enough data on enough students can be collected to make this feasible. Instead, various repeated measures designs, allowing several dependent variables for the same students, as well as several independent variables, to be used.

3. Compare students who have been in the educational program with those who were not in the program; use several achievement and/or language proficiency dimensions simultaneously. Such an analysis would determine whether there is an overall effect of being in the program; it would help to show that the program has increased students' skills on several levels.

- A discriminant analysis will allow the comparison of several dependent measures at once -- but only utilizing one independent variable (e.g., program status: student, nonstudent or students participating in the program for different numbers of years and nonstudents). The analysis not only will determine significance among the groups, but also will use the scores on the dependent measures to predict group membership. The analysis then can determine its success in predicting group membership. This is a various powerful analysis requiring at least 100 students.

Note: Other analysis types are possible, and appropriate, as well. These analyses are fairly easy to do using good statistical software. Consider the expertise of the evaluator and the staff as well as the needs of the evaluation before determining which analysis to do.

Appendix V

Instructions for the annual performance report
Evaluation outline
Guidelines for presenting data
KEYS TO ... Reporting evaluation results to different audiences
Types of evaluation conclusions and recommendations
Evaluation for IASA Title VII
Methods for integrating findings
Suggested form for summarizing report results
IASA Title VII reporting procedures
Evaluation design checklist

INSTRUCTIONS FOR THE ANNUAL PERFORMANCE REPORT

Each Title VII grant recipient should receive the following information from OBEMLA near the end of each grant year. It is provided here to ensure that all are aware of the requirements for the annual performance report. Materials received from OBEMLA also will include a budget form (OMB Nº 1880-0532, authorized through 7/31/98) and an optional form for reporting Parts 1 and 5. The bulk of the information can be provided in any "reasonable format." If you have questions, contact your program officer at OBEMLA or your regional Comprehensive Center.

OBEMLA estimates that the time required to complete this information collection will be an average of 20 hours, including the time to review instructions, search existing data resources, gather and maintain the data needed, and complete and review the information collection. If you have any comments concerning the accuracy of the time estimate(s) or suggestions for improving the form, please write to: US Department of Education, Washington, DC 20202-4651. If you have any comments or concerns regarding the status of your individual submission of this form, write directly to:

[insert program sponsor/official]
US Department of Education
600 Maryland Avenue, SW
Washington, DC 20202-______

To receive a continuation award, recipients of discretionary grants must submit an annual performance report that establishes substantial progress toward meeting their project objectives. The instructions for the annual performance report have been designed to provide the Department with the information that it needs to determine whether recipients have done so. (See §75.118, 75.253, and 75.590 of the Education Department General Administrative Regulations [EDGAR].) Do not use these instructions to prepare the final performance report after the project is completed.

Parts 1-3 and 5 of these instructions request from recipients the information that EDGAR requires to permit the Secretary to make decisions on whether or not to make continuation awards. Part 4 of these instructions requests a summary of new information that may bear on the direction of future activities. This information is requested to help the Department to monitor grant activities and provide technical assistance to recipients.

Submit an original and one copy of the annual performance report. The Department will notify recipients of the due date for submission of the performance report, which will be as late as possible in the project's current budget period.

For those programs that operate under statutes or regulations that require additional (or different) reporting for performance or monitoring purposes, the Department also will inform recipients whether any other (or different) reporting is necessary, and when this additional reporting should be made.

I. COVER SHEET

1. Recipient name and address. Unless changed, repeat this from Block 1 on your last "Notification of Grant Award."

2. PR/Award Number (e.g., T158A20021-95). See Block 4 on your last "Notification of Grant Award."

3. Project title. This should be identical to the title of the approved application.

4. Contact person - name and title. Please provide the name of the project director or other individual who is most familiar with the content of the performance report.

5. Project telephone number and FAX number.

6. Internet address.

7. Performance reporting period. This is the time frame that is requested in Parts 3 and 4 of the performance report for information nonproject status and supplementary information/changes.

   a. For projects that are operating in their first budget period, this period covers the start of the project through 30 days before the due date of this report.

b. For projects that are operating in interim budget periods, *and that submitted a non-competing continuation grant application in the prior budget period,* this period covers the date of submission of that application (unless the Department establishes another beginning date) through 30 days before the due date of this report.

c. For all other projects that are operating in interim budget periods, this period covers the end of the reporting period for the annual performance report that the recipient submitted to receive its previous continuation award, through 0 days before the due date of this report.

8. Current budget period. See Block 5 of your last "Notification of Grant Award."

The cover sheet also must contain the name, title, and signature of the authorized representative of the grantee.

II. PROJECT SUMMARY. One or two paragraphs

III. PROJECT STATUS.* Report your progress in accomplishing the objectives of the project. In doing so, for each project objective, describe the project activities, accomplishments, and outcomes since the submission of the last performance report, or, if you are currently in the first budget period, since the start of the project. Also reference the page numbers and sections of the approved application that address the planned activities or anticipated accomplishments and outcomes. Where it is possible to do so, information on current activities, accomplishments, and outcomes should be quantified.

If a planned objective was not attained, or a planned activity was not conducted as scheduled, explain why, what steps are being taken to address the problem, and the schedule for doing so.

If performance indicators for evaluating your project have been established for your program, or were approved as part of a project evaluation plan contained in your project application, provide information on your project's performance using those indicators.

IV. SUPPLEMENTAL INFORMATION/CHANGES.* As a result of actual performance, recipients often gain additional information (beyond that provided in their initial applications) that affects their future grant activities and/or strategies for accomplishing their approved scope of work. If this is the case for your project, please provide a summary of this information (quantified, where possible) and any change in project strategies, activities, or project outcomes.

V. BUDGET REPORT.*

1. For the current budget period, provide for each approved budget category the total amount of project funds obligated as of 30 days before the due date of the performance report. (See Blocks 9.A-L of the reporting form.) For projects that require recipients to provide matching funds or other non-federal resources, also provide the total of all non-federal contributions as of 30 days before the due date of the performance report. (See Block 10 of the reporting form.)

2. Indicate whether the project expects to have any unobligated grant funds at the end of the current project period. (See Block 11 of the reporting form.

REMEMBER: Recipients must request authorization to carry over funds that were unobligated in one budget period for use in the following budget period. If unobligated funds are needed to complete activities that were approved for the current budget period, §75.253 of EDGAR permits the Secretary to add the amount of these funds to funds that will be awarded through a continuation award for use in the following budget period. Conversely, if any unobligated funds are NOT needed to complete activities that were approved for the current budget period, §75.253 permits the Secretary to deduct the amount of these unobligated funds from the amount of funds that will be awarded for use in the following budget period.

*Note for Parts 3, 4, and 5: Most projects submit with their applications a single budget form, and have a single approved budget for each budget period. However, if your project has multiple components, and was required to submit for approval a separate budget form for each component, please ensure that the information that you provide in Parts 3, 4, and 5 of the performance report reflects activities or expenditures for each of these components.

Evaluation Outline

I. EXECUTIVE SUMMARY

1. Project Design
2. Methodology
3. Findings
4. Conclusions and Recommendations

II. INTRODUCTION

1. Background
   - a. Community
   - b. District
   - c. School
   - d. Students/Participants
2. Screening Procedures and Needs Assessment
3. Project Design
   - a. Goals, objective, and activities for each project component
   - b. Limitations and constraints of the project
4. Summary of Previous Years' Findings and Program Modifications

III. METHODOLOGY

1. Evaluation Design
2. Data Collection Forms and Instruments
3. Validity and Reliability of Assessment Instruments/Procedures
4. Data Collection and Processing Procedures
5. Data Analysis Activities
6. Limitations and Constraints of Methodology

IV. FINDINGS

1. Program Context
   - a. Climate, management, and resources
   - b. Relationship of Title VII to other programs serving LEP students
2. Program Implementation
   - a. Curriculum and instruction, staff, administrators, parents
   - b. Management and effectiveness, professional development activities
   - c. Language of instruction
3. Student Outcomes
   - a. Student descriptions
   - b. Learning strategies and self-concept, cultural-ethnic identity, motivation
   - c. How students are achieving State performance standards
   - d. Comparison of Title VII students to nonlimited English proficient students with regard to school retention, academic achievement, language proficiency
   - e. Effect of program on traditionally underrepresented groups
4. Success in Meeting Goals and Objectives
   - a. Goals and objectives met, not met
   - b. Plans to ameliorate
5. Unanticipated Results/Findings

V. CONCLUSIONS, DISCUSSION, AND RECOMMENDATIONS

1. Summary of Findings and Problems: Context, Implementation, Outcomes, Goals/Objectives
2. Discrepancy Between Expected and Obtained Findings
3. Relationship Between Program Implementation, Context, and Outcomes
4. Recommendations for Program Improvement

Created by the Evaluation Assistance Center-West/NMHU, Albuquerque, NM (Revised 12/95)

GUIDELINES FOR PRESENTING DATA

Numbers are an essential part of any annual progress report or biennial evaluation report; in order to show *how much* students have improved, we must quantify the information. Too many numbers can make a report difficult to read. Here are some suggestions (modified from the *Publication Manual of the American psychological Association*, 4th edition) on dealing with numbers and whether to include them in the text or in tables or figures.

NUMBERS IN TEXT

The general rule is to limit the use of numbers in text. Instead, put the numbers in tables or figures, then provide interpretations of the numbers in the text. When numbers are placed in the text, use figures to express numbers 10 and above and words to express numbers below 10. More specifically, use figures to express

- all numbers 10 and above;
- all numbers below 10 that are grouped for comparison with numbers 10 and above (e.g., *in the 5th, 8th, and 11th grades* ... or *of the 25 words, there were 8 verbs, 12 nouns, and 5 adjectives* ...);
- numbers that represent statistical or mathematical functions, percentages, percentiles, and quartiles;
- numbers that represent time, dates, ages, sample or population size, and scores or points on a scale; and
- numbers that denote a specific place in a numbered series (e.g., *grade 3, page 71, Table 10*).

Use words to express

- numbers below 10 that do not represent precise measurements and that are not grouped for comparison with numbers 10 and above;
- any number that begins a sentence, title, or heading (if possible, reword the sentence to avoid this problem);
- common fractions (e.g., *one-fifth, two-thirds*); and
- universally accepted usage (e.g., *the Fourth of July, the Ten Commandments*).

Sometimes it is necessary to combine figures and words to express rounded large numbers (e.g., *about 3 million people*) or with back-to-back modifiers (e.g., *ten 7-point scales, twenty 6-year-olds*).

TABLES and FIGURES

Tables and figures can be complicated and time-consuming to create. For this reason, they are best reserved for important data directly related to the content of the report. However, a well-constructed table or figure can be economical in that the author, by isolating the data from the text, enables the reader to see patterns and relationships of the data not readily discernible in the text. The following guidelines should be considered when creating both tables and figures.

1. Avoid references to "the table above/below" or "the figure on page 32;" position and page number may change when finally printed. Use statements such as "see Figure 6" or "in Table 8."

2. Use Arabic numerals or number all tales and figures; e.g, Table 3, Figure 4. Do not use letters such as Table A or Figure 2A. However, if these are placed in an appendix, identify them with capital letters (e.g., Table C). Number tables consecutively, and number figures consecutively; do not mix the two numbering systems.

3. Give every table and figure a brief but self-explanatory title; for example, "Average NCE Scores of LEP Students by Grade Level."

4. Make sure the table/figure supplements the descriptive information in the text of the report. However, discuss only the highlights in the text; if every item is discussed, the table or figure becomes unnecessary.

TABLES

The easiest way to present data visually is in tabular form. A well-constructed table can summarize and group data in an efficient and precise form. Information can be presented, numerical or verbal, in rows or columns so that relationships and trends can be identified easily. In developing tables, consider the following procedures.

1. Determine the amount of information (data) necessary for addressing the objective or issue under discussion. Omit peripherally related or extremely detailed data.

2. Make sure the table can be interpreted without reference to the text. Use a descriptive title and notes or footnotes to explain any abbreviations.

3. Do not list two identical columns of figures in two tables. When two tables overlap, consider combining them to form one table.

4. When developing a table, use the headings to list information such as the sex of students, grade levels, the school, school district, and the language groups.

5. The body of the table contains the data. Do not include columns of data that can be calculated easily from other columns.

6. Use the same number of decimal places throughout a given table.

FIGURES

Figures are another efficient way to present comparisons, relationships, or concepts. However, figures are more time consuming and more expensive than presenting data in text or table form. When developing figures, consider the following suggestions.

1. Ensure that the figures in the report are consistent and prepared in the same style as similar g\figures in the report by keeping the lettering, size, and typeface the same.

2. Keep the lines clean and simple, and eliminate all extraneous details.

3. Present the data on the horizontal and vertical axes from small to large and in comparable units of measurement.

4. Plot independent variables (e.g., grade level, schools, testing period) on the horizontal (X) axis and dependent variables (e.g., frequencies, percents, NCE scores) on the vertical (Y) axis.

5. Explain all abbreviations and symbols in a legend or caption; provide an explanatory title or caption.

KEYS TO ... Reporting Evaluation Results to Different Audiences

The information herein has been updated (December 1995) to reflect the IASA statutes and the current EDGAR regulations regarding evaluation.

Sometimes EAC-West staff are asked, "Who really reads these Title VII evaluation reports?" Let's rephrase that question to "Who needs to know what our Title VII program is doing?" Each of those listed below should be kept in mind as you, and your evaluator, write the progress and evaluation reports.

- **PARENTS** If you have an active Parent Advisory Committee that meets regularly, you may have been providing them with on-going information that the report includes. The report lets parents know how their children are benefitting from the program and how the school institutionally is responding to their needs. The evaluation report can be a tool of parent empowerment.

- **YOUR STAFF** Show them the results of their efforts, document their work, highlight their achievement, and show where they need to develop. The report can provide the basis of discussions for program improvement.

- **PRINCIPALS** Principals need and want to know what is going on within their facility. Giving principals their own copy of the report, and reviewing it with them, shows how the bilingual program fits in with the school's other services. It also increases their sense of ownership, hence support.

Outside your immediate school, but within the district, are others who are interested in the outcome of your Title VII project. These include many decision-makers.

- **SCHOOL BOARD** Remember, Title VII grants are intended as seed money to be used to build local capacity -- School Boards have to make that commitment. The report should make clear to the Board what the program is all about, and exactly what they are being asked to commit to when they institutionalize the program.

- **SUPERINTENDENT** School Boards make policy, but usually on the basis of the superintendent's recommendations. The superintendent needs to know how many LEP students the district has, what their needs are, what successful practices the Title VII project has demonstrated, and what gains students have made. With this information, the superintendent can propose a district-supported program and ensure it is carried out.

Outside your immediate educational community are others who can impact Title VII, your school, and your particular Title VII project.

- **MEDIA** The news media frequently report the results of standardized tests; they comment on many aspects of what is happening in the local educational community. The news media are a powerful group with regards to public sentiment.

- **OBEMLA** Annual progress reports must be submitted on a timely basis to assure your grant's continued funding. The biennial evaluation report provides similar information, but on a much more in-depth basis. These reports also let OBEMLA know that you are meeting the conditions of your grant and provide sources of data for policy planning at the federal level.

Most of the potential audiences for the evaluation report are local because that is where the decisions will be made to bring about program improvement and institutionalize program services. The reports, then, should summarize the demographic information about your program students as well as provide results of assessments, and school context and program implementation information.

TYPES OF EVALUATION CONCLUSIONS AND RECOMMENDATIONS

IASA Educational projects are based on broad goals and more specific objectives. The purpose of the evaluation is to determine whether the objectives have been met. When the objectives have been met, the project is deemed "successful." When the objectives have not been met, it is especially important to determine why this has occurred. These "whys" lead to recommendations that should help the project to meet its stated objectives and goals. In this example, the **objective** being evaluated is:

> *The LEP students who attend summer school will show greater achievement gain in English, as measured by program developed assessment instruments, than those LEP students who do not attend summer school.*

The assessment was designed to be used in the fall and spring of the regular school year to measure all students' progress in English. For this project, the spring administration became the pretest for summer school and the following fall's administration became the posttest for summer school. Since all students were tested, the LEP students could be divided int "summer school" and "no summer school" groups. Analysis of the data indicated that both groups of students performed similarly on the assessments. What can the evaluator say about these results?

1. **State major findings matter-of-factly as descriptions of results.**
   The achievement gains of students attending summer school were equal to, but no greater than, those of similar students who did not attend summer school. This objective was not met.

2. **Categorize findings to highlight those that require action.**
   Major findings requiring action in order to meet objective:
   The achievement gains of students attending summer school were equal to, but no greater than, those of similar students who did not attend summer school.

3. **State findings that require action in terms that indicate the necessary action.**
   Without revisions, the summer school curriculum and schedule are not effective in producing achievement gains for summer school attenders that are greater than the gains made by nonattenders.

4. **State options that should be considered.**
   The achievement gains of students attending summer school were equal to, but no greater than, those of similar students who did not attend summer school. In order to meet this goal, the following options should be considered:
   - lengthen the summer school term,
   - make structural changes such as matching the curriculum more closely to the regular school year curriculum, eliminating field trips and other activities that are not directly related to the content of the class, and match summer teachers with their regular students, and/or
   - eliminate the summer sessions and redirect funds to regular year activities that have been proven effective.

5. **Recommend a specific action to be taken.** Be careful of this one as it can backfire. Without really knowing the project, the evaluator may recommend an action that cannot be taken due to restrictions on time, finances, staff, or something else. However, if the evaluator works closely with the project staff, one strong recommendation may be appropriate.

   The achievement gains of student attending summer school were equal to, but no greater than, those of similar students who did not attend summer school. The objective was not met. Because summer school has proven consistently to be ineffective in increasing achievement, it should be terminated. The funds can be redirected to regular school-year activities that have been effective in the past.

Revised by the Evaluation Assistance Center-West/NMHU; Albuquerque, NM (Revised 12/95) from G. Ligon & E.E. Jackson (Austin [TX] Public Schools), Who writes this junk? Who reads evaluation reports anyway? *Paper presented at the annual meeting of the American Educational Research Association, March 1989.*

EVALUATION FOR IASA TITLE VII

Within the IASA, evaluation will occur every two years. 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 addition, there will be an annual progress report, due towards the end of the fiscal year. This report will

- provide information on the progress of the project toward meeting its goals,
- list problems and corrective actions, and
- trigger funds for the following fiscal year.

The evaluation components will include

1. how students are achieving the State student performance standards, if any, including comparing limited English proficient children and youth with non-limited English proficient children and youth with regard to
   - school retention,
   - academic achievement, and
   - gains in language proficiency (English and; where appropriate, native language);

2. program implementation indicators that provide information for informing and improving program management and effectiveness, including data on appropriateness of
   - curriculum in relationship to grade and course requirements,
   - program management,
   - program staff's professional development, and
   - language of instruction;

3. program context indicators that describe the relationship of the activities funded under the grant to the overall school program and to other Federal, State, or local programs serving limited English proficient children and youth; and

4. such other information as the Secretary may require, including
   - objective, quantifiable data;
   - valid, reliable, and fair evaluation and assessment procedures; and
   - the effect of the project on persons being served including any traditionally underrepresented groups such as racial/ethnic minority groups, women, handicapped, elderly, and any students enrolled in private schools.

NOTE: IN DUAL LANGUAGE PROGRAMS, THE GRANT MAY BE TERMINATED IF STUDENTS ARE NOT LEARNING BOTH LANGUAGES. IN A SCHOOLWIDE OR SYSTEMWIDE PROGRAM, THE GRANT MAY BE TERMINATED IF STUDENTS ARE NOT BEING TAUGHT TO AND MAKING ADEQUATE PROGRESS TOWARD ACHIEVING CHALLENGING STATE CONTENT STANDARDS AND PERFORMANCE STANDARDS.

METHODS FOR INTEGRATING FINDINGS

Why?
A professor is called upon to testify before Congress as to whether "pull-out" educational programs work. A policy maker faces the challenge of restructuring the public schools. A program officer is asked to describe the success of different bilingual education models. How do they determine the "right" course of action? According to Hunter, Schmidt, and Jackson (1982), there are "two steps to the cumulation of knowledge: (1) the cumulation of results across studies to establish facts and (2) the formation of theories to place the facts into a coherent and useful form" (p 10). The "cumulation of results" is a determination of an overall pattern of results from earlier studies, evaluations, research, and projects. Further, there is an underlying assumption that new studies, or projects, will incorporate and improve upon the lessons learned in earlier work. The synthesis is the intermediate step between past and future work. The professor, policy maker, and program officer will rely upon such evidence to make their decisions about education in the future.

How?
Three methods generally exist for synthesizing the findings of studies, research, evaluation.

1. ***Narrative review*** In this method, the reviewer collects studies, then provides an overall description of the findings. Frequently, this amounts to combining the conclusions sections of various studies. The resulting information is presented serially, with an overall synopsis. (Study A says ... Study B says ... Study C says ... and so on to Study Y says ... In summation, ...) Critical comments are rarely made in a narrative review.

   Benefits: Easy to do in that the reviewer usually chooses some of the many possible studies to review. Must have enough statistical knowledge to have an overall feel for what the study purports to find.

   Disadvantages: Is subjective--there are few formal rules. An inefficient way to extract useful information--especially when the number of studies being reviewed is large. Difficult to mentally juggle the relationships among many variables, within many studies, and have a meaningful synthesis. Dependent upon studies selected for review.

   Example: In 1974 Munsinger examined a group of studies on children who were adopted and concluded that environmental effects are small: "Available data suggest that under existing circumstances heredity is much more important than environment in producing individual differences in IQ" (p 623). In 1978, Kamin reviewed the <u>same</u> set of studies and reached the <u>opposite conclusion</u>. Baker and de Kanter (1981) reviewed 28 studies on bilingual education. They concluded that the case for bilingual education was weak (see meta-analysis, below).

2. ***Vote-counting*** This method is a refinement of the narrative review. In this case, the reviewer collects the studies, then lists each study and its conclusions. The results are then tabulated as *positive result*, *no result*, or *negative result*. For instance, if looking at a study of bilingual/ESL differences, the results might indicate *higher score for ESL classroom*, *higher score for bilingual classroom*, or *no difference between classrooms*. The reviewer then counts the number of studies supporting the various views and chooses as "successful" the view with the most "votes."

Benefits: Generally easy to do. Provides an overall picture of results for several studies, regardless of the number of studies. More easily allows the combination of different variables from different studies.

Disadvantages: Does not consider some of the important research features--how big were the differences between groups? was the design of the project appropriate? how many people were included in the study as subjects? Still dependent upon studies selected for review.

Example: The *Review of Hawaii's Title VII Part A 1990-91 Annual Reports* is an example of the vote-counting method for summarizing the Title VII reports for the four projects within the state of Hawaii. The results were then summarized into a brief report on the accomplishments and the challenges for the Title VII projects.

3. ***Meta-analysis*** Meta-analysis often is referred to as the "analysis of analyses" or the "statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating findings." In addition to looking at the specific results, meta-analysis allows the reviewer to statistically analyze such features as the "goodness" of the design, assessment instruments, size of subject group; the size of the difference between groups; the year of the study, the type of study (published, dissertation, etc.); and other features of the overall study.

Benefits: Allows the consideration of many facets of the research project. Is extremely objective. Meta-analyses can be considered research studies themselves, requiring a great deal of work.

Disadvantages: Heavily dependent upon the studies selected for inclusion. Requires a great deal of research and statistical knowledge and abilities. While purporting to be objective, may be biased in selection of studies to be included. May oversimplify and mislead. Some statistical procedures still being defined. Qualitative research cannot be included.

Example: Willig (1985) conducted a meta-analysis using the studies selected by Baker and de Kanter (see narrative review example). She eliminated 5 of them (i.e., programs in Canada, programs in special schools). When statistical controls for methodological inadequacies were employed, participation in bilingual education programs consistently produced small to moderate differences favoring bilingual education in the areas of reading, language, mathematics, writing, social studies, listening comprehension, and attitudes toward school or self. Programs characterized by instability and/or hostile environments showed lower effects.

Now what?

In reviewing evaluations of Title VII bilingual programs, the narrative review is difficult due to the number of reports and types of programs that are available; the meta-analysis is difficult due to the time and statistical expertise necessary. The vote-counting method appears to be the best, but still has some inherent biases and problems. At EAC-West, we suggest that reviewers of annual reports carefully define the purpose of their review, then specify the parameters of the review. Once this is done, a checklist type of form can be created to assist in codifying the information.

SUGGESTED FORM FOR SUMMARIZING REPORT RESULTS

Purpose of the synthesis (be as specific as possible) ______________________________

____

____

Type of program
□ Develop & Implementation □ Enhancement
□ Comprehensive Schoolwide □ Systemwide
□ FLAP □ Other Title VII______________

Number of languages served
□ One □ Two □ Three
□ 4-5 □ 6-10 □ 10-25 □ More than 25

Languages
Check the general type of language, then list the name of each language.
□ Western European ______________________
□ Eastern European ______________________
□ Asian/Pacific Island ______________________
□ Native American/Alaskan ______________________
□ Other ______________________

Grade-level
□ preK/K □ 1-3 □ 4-6 □ 7-8
□ 9-12 □ Adult □ Other ______________________

Number of participants in grade
□ Less than 10/grade level
□ 10-20/grade level □ More than 20/grade level

Number of participants in program
□ Less than 50
□ 50-75 □ 76-100
□ 101-150 □ 150-250
□ 250-1,000 □ More than 1,000

Type of assessments used
□ Standardized (stdz) NRTs
□ Stdz language proficiency □Altemative assessments, academic achievement
□ Affective tests □ Altemative assessments, language proficiency
□ Combination of standardized and altemative assessment
□ Other ____

Results for ***program*** students
□ Gains in Engish proficiency □ Gains in L1 proficiency
□ Gains in 1-2 content areas □ Gains in 3-5 content areas
□ Affective gains □ Other gains: ______________
□ Losses in ____

Combine the listed factors with the results of a summary of the information from other checklists created to measure the effects of programs included in this project. Depending upon the purpose of the synthesis, different factors might be more/less important. Results should be compared only for programs that are similar. For instance, a program using only standardized tests is not comparable to a program using only alternative assessments; two-way developmental programs with ESL programs.

Developed by the Evaluation Assistance Center-West/NMHU; Albuquerque, NM (Revised 12/95)

IASA TITLE VII REPORTING PROCEDURES

Under the new IASA Title VII, two types of reports will be sent to OBEMLA:
(1) annual progress report (2) biennial evaluation report.

Annual Progress Report
To receive a continuation award (monies for the next year), Title VII program directors will need to submit an annual progress report. Specific directions and guidelines will come from OBEMLA. In general, the annual progress report will follow the criteria set forth in the *Education Department General Administrative Regulations (EDGAR)*. Besides basic information (name of project, school/school district, responsible person[s], and so on), OBEMLA anticipates requesting the following:

Progress toward accomplishing objectives of the project. For each objective, describe the activities, accomplishments, and outcomes for the current year. Actual page numbers and sections of the approved grant application should be referenced. Quantify this information wherever possible. If specific performance indicators have been established, provide information about how the program is doing relative to those indicators.

If a planned activity or objective was not attained or was not conducted as planned, explain why and explain what steps have been taken to remedy the situation. Include a timeline.

Supplemental information should be provided as necessary and appropriate. For instance, preliminary performance outcomes may suggest modification of future grant activities or strategies for accomplishing the objectives. If this occurs, provide a summary of the information (quantified wherever possible) and any modification in project strategies, activities, or anticipated outcomes.

The budget report will provide OBEMLA with the information about monies expended and/or obligated for the budget period as well as monies that were unobligated as of the end of the budget period. (There are some circumstances under which monies may be carried over into the next budget period.)

OBEMLA will notify grant recipients regarding the due date for the annual reports. They will be scheduled as late in the current budget period as possible--the report will be required before the following year's funds can be approved. It is anticipated that much of the information can be provided through charts and tables, with less text required.

Biennial Evaluation Report
The biennial (every two years) evaluation report will include information from the previous annual progress report(s). It will differ in that more detail is needed, and more text will be required to provide the necessary material. According to §7123 of the *Improving America's Schools Act* and §75.590 of EDGAR, the following will be required:

Progress toward accomplishing objectives of the project. For each objective, describe the activities, accomplishments, and outcomes for the appropriate years. Actual page numbers and sections of the approved grant application should be referenced. If a planned activity or objective was not attained or was not conducted as planned, explain why and explain what steps have been taken to remedy the situation.

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 appropriate, native language) proficiency.

Program implementation indicators that provide information for informing and improving program management and effectiveness, 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.

Program context indicators 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.

The effect of the project on persons being served by the project, including any persons who are members of groups that have been traditionally under represented, such as members of racial or ethnic minority groups, women, handicapped persons, and the elderly.

If private school students are participating, they should be described, as well as their progress towards meeting the objectives of the project.

Any other information the Secretary may require.

The purpose of both **annual progress reports** and **biennial evaluation reports** is to (1) improve the program, (2) further define the program's goals and objectives, and (3) determine program effectiveness. Make sure the reports are useable for program director, staff, evaluator, administrators, and families.

EVALUATION DESIGN CHECKLIST

Type of Grant
(Note: Not all grant types are funded each year.)

□ Program Development and Implementation
□ Program Enhancement
□ Comprehensive Schoolwide
□ Systemwide Improvement

EDGAR is identified by section (§) numbers; all else is from the IASA Title VII statutes. When using this checklist to review a grant application, remember that the evaluation plan may be weighted differently depending upon the grant type.

Evaluation Item	**Check Rating**			**Comments**
	Not Found	Sparse	Ade-quate	
EDGAR §75210(b)(6)(I) Evaluation is appropriate to the project.				
Evaluation every two years.				
Evaluation can be used for program improvement.				
Evaluation can be used to further define goals and objectives.				
Evaluation can be used to determine program effectiveness.				
EDGAR §75210(b)(6)(ii) Evaluation is objective and quantifiable.				
Evaluation plan ensures accountability in achieving high academic standards.				
Evaluation and assessment procedures are valid, reliable, and fair.				
EDGAR 34 CFR §75.590(a) Grantee is progressing toward meeting approved objectives.				
EDGAR 34 CFR §75.590(b) Effectiveness in meeting purposes of the [Title VII] program				
How students are achieving State performance standards including ✓ comparison with nonlimited English proficient students with regard to ✓ School retention, ✓ Academic achievement, ✓ English proficiency, and ✓ Native language proficiency, as appropriate.				

Evaluation Item	Check Rating			Comments
	Not Found	Sparse	Ade-quate	
Program implementation indicators for informing and improving ✓ Management & effectiveness, ✓ Appropriateness of curriculum for promotion/graduation, ✓ Appropriateness of management, ✓ Appropriateness of professional development activities, and ✓ Appropriateness of language of instruction.				
Program context indicators that describe the relationship of title VII to other federal, state, or local programs serving LEP students.				
EDGAR 34 CFR §75.590(c) Effect of the program on persons served by the project, including				
EDGAR 34 CFR §75.590(c)(1)(I-v) Any persons who are members of groups that have been traditionally underrepresented ✓ Racial or ethnic minorities, ✓ Women, ✓ Handicapped persons, and ✓ Elderly.				Note: It is unlikely that women or elderly will be involved, except as parents or extended family.
EDGAR 34 CFR §75.590(c)(2) If the program statute requires private school participation, students who are enrolled in private schools.				Note: Title VII does not require private school participation.

No particular evaluation design is required or recommended. Some which may be appropriate, depending upon the manner in which the objectives are written, include

- gap reduction,
- pre/post,
- nonproject comparison group (only needed in EDGAR 34 CFR §75.590 (), and
- grade cohort.

Dynamic (on-going, more fluid) evaluation will be helpful for an in-depth view of the progress of students.

FL025124

NOTICE

REPRODUCTION BASIS

□ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☑ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").