DOCUMENT RESUME

ED 228 278 TM 830 153

AUTHOR Burry, James
TITLE An Introduction to Assessment and Design in Bilingual Program Evaluation.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
REPORT NO CSE-RP-1
PUB DATE 82
NOTE 53p.; Paper presented at the Title VII Bilingual Education Management Institute (Los Angeles, CA, March 30-April 1, 1981).
PUB TYPE Guides - Non-Classroom Use (055) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Achievement Gains; *Bilingual Education Programs; *Data Analysis; Decision Making; Educational Objectives; *Evaluation Methods; *Evaluation Needs; Needs Assessment; Program Effectiveness; *Program Evaluation; *Research Design; Research Methodology; Test Use
IDENTIFIERS *Elementary Secondary Education Act Title II

ABSTRACT

The most difficult problems in bilingual education evaluation are disagreement over what evaluation is and how it is done; the debate over what bilingual education is and how a program is planned and operated locally; and the nature of bilingual education itself, which creates problems in assessment and design methodology. General information is provided on three basic considerations in bilingual program evaluation: assessment, evaluation design, and data analysis. Assessment is the full range of information that might be used to make decisions about a bilingual program, including what it accomplishes for its students as well as the procedures to achieve these goals. Measures of student performance and 100 program processes, such as interviews and observations, are examined. A brief examination of major designs used in bilingual program evaluation focuses on the designs that seem to be most useful and feasible in a bilingual program setting. The most general-purpose designs for investigating program outcomes are the time-series/longitudinal designs. The exposure-to-treatment design is widely applicable for formative evaluations. Accountability designs, report information about student achievement of local objectives or national norms. Basic issues raised concern the questions an evaluation might try to answer, collection of information, and the appropriate analytic techniques. (CM)

AN INTRODUCTION TO ASSESSMENT AND DESIGN IN BILINGUAL PROGRAM EVALUATION

James Burry

CSE Resource Paper No. 1

1982

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

The work reported herein was supported in part under a grant from the National Institute of Education. However, the opinions expressed do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement should be inferred.

TABLE OF CONTENTS

AN INTRODUCTION TO ASSESSMENT AND DESIGN IN BILINGUAL PROGRAM EVALUATION*

INTRODUCTION

This paper offers general information on three basic considerations in bilingual program evaluation: assessment, evaluation design, and data analysis. By assessment is meant the full range of information that might be used to make decisions about a bilingual program, including what it accomplishes for its students as well as the procedures it follows to achieve these goals. Covered are measures of student performance and measures of program processes, such as interviews and observations. A brief examination of major designs used in bilingual program evaluation will be made, focusing on the three or four designs that seem to be most useful and feasible in a bilingual program setting. Basic issues will be raised concerning questions an evaluation might try to answer, information that might be collected to answer them, and analytic techniques appropriate to the particular questions asked and data collected.

These remarks have two purposes: (1) for the internal evaluator of a bilingual program, especially for monitoring and improving the program while in operation; and (2) for the staff of a

---

*This paper was presented at the Title VII Bilingual Education Management Institute, Los Angeles, California, March 30-April 1, 1981.

bilingual program who want an external evaluation that is technically sound as well as appropriate within the particular program's constraints. This second concern involves asking the right kinds of questions over the use of a certain kind of measure, design, or means of analysis.

The procedures to be discussed belong to a series of training workshops developed at the University of California, Los Angeles Center for the Study of Evaluation (CSE) over the last few years. Each of the techniques, designs, and procedures is covered in detail in one or more of these workshops. The broader rationale behind these workshops and the suggestions in this paper have been discussed in an earlier article (Burry, 1979).

BACKGROUND CONCERNS

General Problems in the Evaluation of Bilingual Programs

Evaluations rarely provide the most useful information about what bilingual programs are like or the range of outcomes and level of success they achieve. Among the reasons for this failure are the following:

1. Those called upon to evaluate bilingual programs, though well versed in the general area of educational evaluation, may be unfamiliar with the particular needs and characteristics of bilingual education.

2. Evaluations of bilingual programs frequently provide only general descriptions of the nature and content of the programs; such information is of limited use to decision makers.

3. Diversities among bilingual programs are great, and the extent to which these differences are related to differential development across programs is still questionable. (Agreement is needed on such basic assumptions as what constitutes a minimum bilingual education and the kinds of achievement gains and other outcomes to be expected of a particular program. Attempting to measure the effects of a bilingual program without determining whether a bilingual program actually exists or what it needs to accomplish makes it difficult to establish valid criteria for determining success or failure.)

4. Even assuming the development of valid criteria, evaluators must still explore the relationships among instructional strategies, implementation techniques, and program outcomes.

5. A shortage of adequate instruments for the assessment of students of non- and limited English proficiency (NEP/LEP), exists.

6. A problem exists over designs appropriate to the evaluation of bilingual programs. The fundamental questions regarding these programs has been to determine whether the scholastic achievement of students in bilingual programs equals or excels what it would have been had they remained in a regular, monolingual course. Two factors greatly hinder the generation of reliable findings: (a) bilingual programs are intended to provide adequate education for a broad range of NEP/LEP students, and (b) methodological problems of participant selection are likely to obscure program results.

The above comments form the background for the three most difficult problems in bilingual education evaluation. First, there is much disagreement over what evaluation is and how it is done. Second, there is much debate over what bilingual education is and how a bilingual program is planned and operated locally. Third, the nature of bilingual education, in terms of which students it should reach and help, creates problems in methodology, especially in the areas of assessment and design.

An Approach to Evaluation

Following current evaluation practice, many bilingual program evaluations appear guided by an approach that relates the methods of evaluation with those of research. This paper will distinguish between evaluation and research. This distinction is important, far beyond questions of topic of interest or techniques employed, involving fundamental differences in the kind of information generated, how it is generated, and its intended uses. Cronbach and Suppes (1969), in their discussion of modes of inquiry, cite the need to:

> ...distinguish decision-oriented from conclusion-oriented investigations. In a decision-oriented study the investigator is asked to provide information wanted by a decision-maker: a school administrator, a governmental policymaker, the manager of a project...or the like. The decision-oriented study is a commissioned study. The decision-maker believes that he needs information to guide his actions and he poses the question to the investigator. The conclusion-oriented study, on the other hand, takes its direction from the investigator's commitments and hunches. The educational decision-maker can, at most, arouse the investigator's interest in a problem. The latter formulates his own

> question, usually a general one rather than a question about a particular institution. The aim is to conceptualize and understand the chosen phenomenon; a particular finding is only a means to that end. (pp. 20-21)

In light of this distinction, evaluation can be categorized as decision-oriented and research as conclusion-oriented study. Evaluation differs from research intended to produce results of general significance; evaluation, with its decision orientation, is intended to provide results relevant for a particular program at a specific point in time. It differs from conclusion-oriented research in the reasons for conducting the work, the constraints imposed by the institutional setting, and the intended use of the information generated. Evaluation should generate information that will lead to decisions about educational questions and problems within a particular setting.

The decision orientation toward evaluation considers both the manner and the times at which evaluations might be conducted. It points out phases in the development of a program during which various audiences might effectively use credible information. It does not assume there is a single approach to evaluation, which is universally appropriate. Rather, the approach to evaluation is guided by the kind of bilingual program to be evaluated, the context or setting in which it operates, and the real-world constraints with which the program staff must work.

The distinction between research and evaluation should not be interpreted to mean that research and evaluation are mutually exclusive categories. Evaluations often employ research methodology to enhance the generalizability of their findings. Further, many

evaluations include some research questions. For example, hypotheses regarding the relationship between certain aspects of the instructional program and student outcomes may be tested. Evaluation, thus, may encompass research and is a somewhat broader activity than classic research. Both, however, are complementary activities; in the complex world of the evaluation of bilingual programs, there is room for both pursuits.

From the above distinction between research and evaluation, the Center for the Study of Evaluation has defined evaluation as the process of selecting, collecting, and interpreting information for the purpose of keeping various audiences informed about a program. Usually these audiences will use the information to make decisions. The decisions needed to be made about a bilingual program depend on the program's stage of development. Creating a bilingual program can be viewed as occurring in four phases:

Needs assessment. The first phase in the cycle of program development is needs assessment, during which the need for the program is established. In this phase students are assessed to determine the range of English-language proficiencies that exist. This information will determine the most appropriate kind of bilingual program to meet students' needs. This phase will also include evaluating the surrounding context in which the program will have to operate, e. g., parental and community attitudes, support for the program, and staff qualifications.

Program planning. The second phase in program development is program planning. Here teachers, parents and other community members, curriculum experts, and others plan a program to meet the

high-priority goals determined by the needs assessment. Plans should also be made for evaluation of the program at this time, including consideration of the measures to be used, the most appropriate evaluation design, and means of analysis and reporting.

The first two phases are concerned with establishing the program's goals. A needs assessment sets goal priorities; program planning describes the program to reach these goals and how the program will be evaluated. Implicit in this model is one prescription about how evaluations should proceed, that is, they should address the goals and processes of the **particular program being evaluated.** The evaluator should look first at the goals and processes of the particular bilingual program and then design an evaluation appropriate to that set of goals and program processes. To ensure the widest use of the information gathered, both for people in the program and for audiences external to it, the evaluator should also plan to **describe** and **document** the program's features while in operation.

Formative evaluation. Formative evaluation requires collecting and sharing information for program improvement. While a program is being installed, the formative evaluator provides program planners and staff with implementation information to help adjust and improve it. Eventually, of course, people will want to know whether or not the program is effective; but these questions cannot be answered immediately. They must await a summative evaluation that asks about the program's overall value and effect.

While the program develops, the evaluator should provide a wide range of **program implementation** information. This infor-

mation can be used to decide if the program needs to be modified and/or to check that students are progressing.

Summative evaluation. Summative evaluation examines the program's total impact. After its developmental stage and when functioning as intended, the program is ready to be summarily described and perhaps judged. This summary evaluation, the kind of evaluation typically conducted, will include a program description and an estimation of its effect on student outcomes.

Implicit in the four phases of program development and evaluation is the notion of **program documentation.** Documentation of bilingual programs is a crucial activity. It adds to the evaluation component an element that fully describes the bilingual practices followed in the program and those events and processes that interacted to achieve outcomes. This documentation should describe the program's crucial features and be considered an integral part of program operation and evaluation. If worthwhile results are found in a program, it may be worth maintaining to test its future effects and to help others assess its relevance in a new setting. Thus, simple statement of outcome and estimation of effect is not adequate to fully demonstrate program achievements. Needed is a thorough description that, in conjunction with statements of outcome, will document the interactions of process and events that constituted the program.

The approach to evaluation described above permits consideration of a broad range of important concerns: **context characteristics** such as class size, school/district features, and the

program's time frame; **input characteristics** such as student age, linguistic background, ability, and attitude; **implementation characteristics** such as program features and need for improvement; and **outcomes**, including both anticipated and unanticipated effects. It focuses data collection where program effects are most likely to occur and gathers a wide variety of information, both process and outcome. Above all, it attempts to provide credible and useful information for a broad range of educational decisions.

The Nature of Bilingual Education

What bilingual education is, or should be, depends on who you ask. Some feel that bilingual education should be transitional; others feel it should be maintenance. Cultural, psychological, social, political, and educational issues must be considered: basic issues of language dominance versus language proficiency and differing theories of learning, development, language acquisition, and linguistics.

Four major kinds of bilingual programs exist:

Type I, Transitional Bilingual, allows pupils to adjust to school and/or subject matter until English skills are developed. In this kind of program, language development is the objective.

Type II, Monoliteral Bilingual, works to develop both languages for aural/oral skills but doesn't include literacy skills in the native language. This kind of program develops fluency in the native language and is a compromise between language shift and language maintenance.

Type III, Partial Bilingualism, aims for fluency and literacy in both languages, but literacy in the native language is restricted to certain subject matters. This kind of program emphasizes language and cultural maintenance.

Type IV, Full Bilingualism, develops skills in both languages. This kind of program is directed at development and maintenance of the native language.

Given these differences in program intention, the overall evaluation and procedures it follows should be appropriate to the nature and purposes of the program being evaluated.

Methodology Problems in Design and Testing

The minimum requirements for Title VII basic programs as amended by the regulations published in the **Federal Register**, March 29, 1979, include:

1. The evaluation should assess project progress in achieving its Title VII objectives in all components.
2. Evaluation of pupil achievement should provide for some comparison of performance on reading skills in English and in the native language, with an estimate of what performance would have been in the absence of the program. Comparisons may be based on local, regional, or national norms on standardized tests; on historical data; or on achievement scores of a comparison or control group. The limitations of such comparisons must be recognized in data analysis and interpretation.

3. Instruments to evaluate student performance should be described, as should the rationale for their selection and procedures for use.
4. Pre- and posttest results should be reported on all participating students (and comparison students, if used) using means, standard deviations, and appropriate tests of statistical significance.
5. Procedures should be established for determining when students no longer need assistance in developing English proficiency; this includes an individual evaluation of each student enrolled in the program for two years to determine if the student should remain in the program.

A major consideration in planning a bilingual program evaluation is to consider the establishment of an appropriate standard of comparison and the selection or development of measures appropriate to the desired means of comparison. These decisions should be influenced by the particular program's features and relevant federal, state, and local requirements. While becoming familiar with the program, the evaluator must keep in mind the basic requirements of an acceptable evaluation and the design and testing issues associated with them.

Federal regulations that provide for transitional bilingual programs call for comparisons based on local, regional, or national norms; on historical data; or on achievement scores of a comparison or control group. It is generally impossible to establish a control group to provide a close estimate of how well students would have done without the program. However, the evaluator

should attempt to make the best estimate available within the particular program's constraints, noting any limitations stemming from the comparison used as the estimate's basis.

Once the standard of comparison has been established, tests and other measures appropriate to that type of comparison can be selected or developed. Tests, whether selected or developed, should possess acceptable technical properties and match the program's curriculum and the language of instruction. Given the current limitations of both norm-referenced and criterion-referenced tests, and because all aspects of a bilingual project cannot be assessed by achievement measures, a broad data base should be developed. This base may rely on standardized and criterion-referenced tests as well as other measurement techniques such as questionnaires and observations. A broad data base, in addition to serving the need of program documentation, may help to alleviate some of the more troublesome problems discussed below.

In general, program evaluations are conducted to provide evidence of a program's success. They try to answer the basic question, "Did the students in bilingual programs achieve as well or better than if they had remained in a regular, unilingual program?"

As mentioned earlier, two factors pose design problems for evaluating bilingual programs: (1) the broad range of NEP/LEP students, and (2) the methodological problems of participant selection.

Many designs have been proposed based on comparison groups. Yet bilingual programs can rarely be studied experimen-

tally because legislation or program administrators may prohibit withholding bilingual services from students entitled to them. Because of legal and ethical considerations, students who might benefit from bilingual services cannot be randomly assigned to a non-treatment comparison group. Further, bilingual programs do not consist of a single isolated treatment that can be evaluated experimentally. Many employ multiple compensatory efforts that overlap, so student progress might be due to any combination of these treatments.

For these reasons, evaluations frequently use other than randomized controlled experimental designs. Since these nonexperimental comparisons are confounded in various ways, reliable baseline data will provide information that helps the evaluator determine whether or not a program has significantly affected student performance. Gathering baseline data requires collecting information prior to the program's implementation and comparing that data to the program's results. For example, it may be necessary to consider sources of student ability differences. Issues to explore include the student's initial linguistic status, the period in the student's development in which the second language is introduced, and the length of time the student has been in the program. Such factors could either be controlled when evaluating program impact or examined as to the effect they have on the bilingual student's education.

Initial language dominance and/or proficiency must be considered when evaluating a program's effects. Language dominance and proficiency are often the primary criteria for selecting students

to participate in bilingual programs. In addition to its application in student placement, initial diagnosis of language skill is also important for understanding test relevance and interpretation of subsequent test performance. Previous bilingual evaluations have revealed that the continuum of relative fluency in both English and the native language can be extremely large among bilingual program students. Thus, proper generalizability about the students served depends on an approximate account of the heterogeneity of the dual language skills present in programs. By contrasting similar language groups, it may be possible to provide much more detailed information about program effect. By estimating students' educational gains in terms of their particular linguistic groups, the gain derived is also likely to be more reliable.

The appropriate time for introduction of a second language, of reading, and the length of time students will be in a program before final assessment are issues that may demand longitudinal (more than a one- or two-year time span) evaluation. Longitudinal studies are critical to determining how two languages should be used within an entire program curriculum and what ages are most conducive to language learning and achievement. Because of the number of variables involved, bilingual program evaluation should be based on: (1) multiple questions (not just whether students educationally benefited more by program participation as opposed to remaining in regular classes), and (2) multiple analytic designs (rather than the traditional cross-sectional design associated with end-of-year summative evaluations).

Beyond the question of design, lack of adequate instruments for the assessment of NEP/LEP students presents another problem for evaluators. Much has been said about the inadequacy of standardized tests for assessment of culturally and linguistically different students. These limitations are magnified when norms are used, since NEP/LEP students are generally underrepresented in the field testing and norming sample and separate norms for these groups are not available. Moreover, even when adequate field testing has been conducted among NEP/LEP students, the possible linguistic and cultural biases of many standardized instruments undermine their validity and reliability. The problems are particularly acute with respect to English language measures, but are often equally pervasive in instruments that are simply translations from English-language versions.

There is a serious shortage of technically sound instruments that are culturally and programmatically appropriate for non- and limited English speakers. Test scores obtained on instruments currently in use can vary considerably based on the congruency between the curriculum and test content.

An alternative to standardized tests is the use of criterion-referenced tests. These may potentially indicate the extent to which students have mastered their instructional program's objectives. They may serve particularly well for diagnostic and prescriptive information at the local level. But criterion-referenced tests also have limitations. Sparse technical information on them is available, and how appropriate it is to use them across projects is questionable. Further, because of the varying

lack of difficulty of items and objectives, it is difficult to compare an achievement score on one test to that of another. The preceding design and testing issues should be kept in mind while the evaluator examines the program's characteristics.

IMPLICATIONS AND SUGGESTIONS FOR PRACTICE

Assessing Title VII Program Gains

Title VII evaluations often intend to provide information for use at local, state, and federal policy levels; and they are often more concerned with broad accountability than with program-specific information. In addition, Title VII audiences tend to ask for information that can be used to compare students in the program with other students. Because of accountability and comparison concerns, Title VII evaluations often rely on the collection and analysis of pupils' scores on a norm-referenced test. This kind of test has a specific purpose and measures certain kinds of goals, and its use as the sole measure for Title VII program evaluations is subject to question.

A norm-referenced test draws its content from a general body of subject matter. It is designed to give information that can be used to compare a pupil's performance with the performance of other pupils assessed on the same test. It demonstrates the differences among pupils on the basis of their test scores rather than showing exactly what they have learned.

A norm-referenced test can compare pupils because it is used to place the pupils on a scale of performance from highest to lowest on the basis of their test scores. Because a norm-referenced

test ranks students, it might not accurately show what a Title VII program has actually accomplished for the participating pupils. To rank students on the basis of their test scores, pupils' scores must be changed to another kind of score, such as a percentile. This kind of score does not show **how much** or how much of **what** a pupil has learned because the test it comes from measures general goals instead of specific program objectives.

Because a norm-referenced test covers broad goals, it may not adequately measure a school's Title VII program's particular instructional objectives. If the validity of a test can be challenged, then statements about its reliability or consistency have little meaning.

Due to these problems, criterion-referenced tests have been proposed instead of, or in association with, norm-referenced tests. A criterion-referenced, or objectives-based, test is not meant to rank or compare pupils on the basis of their test scores. Instead, it is intended to show **how much** pupils have learned in terms of the specific objectives of the program they are in. This kind of test often has a standard of performance or cut-off score used for making decisions about the **individual** pupil's learning, independent of what the **other** pupils have learned.

While criterion-referenced tests are intended to provide specific performance information about individual pupils, they should not be seen as a cure for all testing problems. Some issues must still be solved in the development and use of these tests. For example, a criterion-referenced test might not provide a base that can be used to interpret what pupil achievement of specific objec-

tives actually means. That is, on the basis of the test scores alone, it might be difficult to judge the educational significance of attaining a set of instructional objectives, to decide if that attainment is important or trivial. Methods to overcome this problem will be covered later in this paper.

Norm-referenced tests, beyond any general limitations, may be even more limited when used to evaluate educational programs like Title VII. A norm-referenced test is usually tried out with a national sample of students to see how well it functions. These students are called the **norming sample** because their test scores are used to interpret the scores of other students. Since pupils who are eligible for Title VII programs differ considerably from the norming sample, comparing them with a national sample is not the best way to determine the quality of their performance.

There are other problems in the use of norm-referenced tests for Title VII program evaluations. Title VII evaluations usually try to state what the program accomplished for its pupils. One way to make this kind of statement is to compare the performance of the pupils in the program with identical pupils who did not get the program. But, as mentioned earlier, establishing this kind of comparison group, consisting of students who are eligible for the program but do not get it, is difficult. Thus, for Title VII evaluations, where it may only be possible to get information on participating pupils, the need is for a method of obtaining information interpreted in terms of pupil progress compared to national student standards.

The two most frequently used strategies for establishing comparison on the basis of norm-referenced test scores have problems because of the assumptions they make about Title VII pupil learning. Proponents of one approach believe that pupil learning, rather than being cumulative, will follow a straight line. If a pupil's score on the pretest is known, that pupil's posttest score can be estimated. If a pupil scores higher than was estimated on the posttest, then the program is said to be successful. With the other approach, the assumption is that if a pupil scores at a certain level on the pretest, at a certain percentile, for example, then that pupil will score at the same level unless educated in a Title VII program. If pupils score higher on the posttest, then the program is presumed to have been beneficial.

It is not clear whether the first procedure provides information that will make the program look better or worse than it actually was. But the second procedure may make the program look as if it accomplished less than it did. This is because pupils in programs like Title VII will not score the same on pretest and postttest if they do not get the program. Rather, they tend to fall further and further behind and score even lower on the posttest unless they are provided with additional instruction. So if a Title VII program simply preserves a pupil's relative standing from pretest to posttest, the program may have been more effective than the scores would suggest since the pupils did not fall further behind.

Both approaches also assume that norm-referenced tests are accurate for measuring the effects of specific instructional pro-

grams. This assumption is questionable and suggests the possibility of using criterion-referenced tests, even though this kind of test could still be improved.

To summarize to this point, there are problems in the use of norm-referenced tests for Title VII evaluations and there are difficulties in setting reasonable performance standards for the pupils in these programs; so it is hard to judge the educational significance of the gains accomplished by a given program. Therefore, there is a need for further work on: (1) the identification and use of tests that are accurate for measuring the effects of Title VII programs, and (2) setting standards of performance appropriate for students eligible for these programs.

These problems highlight the need for a broader approach to evaluation that uses not only achievement tests but also other techniques such as observations, interviews, questionnaires, and other information that can be used to see what the program accomplished. These kinds of measures can be used to provide a background of information for interpreting what the test scores actually mean, whether these tests are norm- or criterion-referenced. This expanded approach to evaluation is necessary not only for the interpretation it offers but also because of the contribution it can make to help determine exactly which components of the Title VII program are most effective.

Title VII Program Components and Pupil Performance

From the above, it should be clear that Title VII evaluations can be improved if information is provided that will help inter-

pret the meaning of pupil scores. Such evaluations will be collecting information for decisions about the worth of the local program. The local program will essentially be one version of a larger externally-funded program since each local school site often decides what its program will look like and what it will do to achieve its objectives. Because of variation from school to school, it is difficult to find out which kinds of instructional materials and strategies lead to pupil gains.

Local school district Title VII programs are not all the same, since each program tries to be most beneficial for its particular pupils. Evaluation should therefore help clarify local program intentions and operations, help operators understand the program they have implemented, and whether it is proceeding on target or needs to be improved. Using a broad set of measures is appropriate to this kind of assistance, if these measures are selected or developed so that they match the educational objectives of the local school program and fit the setting in which it operates.

For example, since schools develop their own objectives and expectations, the accuracy of a norm-referenced test for a given school's program must be questioned. It is possible that the test used is more or less appropriate for one program compared to another, and so there are questions about the accuracy of the test scores provided and what they mean for the individual school. If one school's objectives match better with what the test measures than those of other schools, this school will possibly score higher on the basis of this match, no matter what takes place in

its classrooms. If a school's objectives do not match the test well, this school's test scores would possibly be low even though it was benefiting the pupils.

Each school planning and developing its own instructional program may rely on a wide variety of materials and strategies. Therefore, it is important to know how well a certain set of instructional practices fits the test used as the principal means of evaluation. Schools can also differ in how and how much they use resource teachers, aides, teaching assistants, and volunteers. The extent to which such resources can affect program impact needs to be verified; their effect cannot be determined on the basis of test scores alone.

In short, current evaluation practice makes it difficult to determine the degree of pupil gains and which particular school efforts led to them. Evaluation practice must be broadened if we are to judge the educational significance of the gains growing out of Title VII programs. New techniques for describing and documenting local school practices must be explored so that we can determine which specific Title VII instructional features succeed and why.

No matter which test or kind of test is used for assessing Title VII pupil outcomes, the first evaluation priority is a careful analysis of the school's Title VII curriculum. This information can help determine the match between the school's curriculum and the objectives the test measured. From this, over time, descriptions of Title VII instructional features and the kinds of educational gains they lead to can be developed.

At the same time, accurate and cost-effective means of documenting or describing Title VII instructional practice at the school and classroom levels should be investigated. By thoroughly describing the total instructional setting, a picture will emerge of the specific school/classroom objectives; how these objectives were established; instructional materials and practices used; kinds of students for whom the instruction is most effective; setting and manner in which instruction and assessment take place; extent to which instructional practice was modified on the basis of ongoing use of assessment information; kinds of modifications they led to; and manner and extent of use of human and material resources. In turn, identification of successful and unsuccessful Title VII program operation will be enhanced.

Given the previously described problems with norm-referenced tests, use of such tests should be subject to further investigation. Should a norm-referenced test be used, it should be accurately established how well it fits with the instructional objectives and practices of the program evaluated. Scores obtained would then be interpreted in terms of how well the test measures the individual school's objectives.

Because a norm-referenced test is intended to show the total picture, it might best be used on some sampling basis. That is, not every Title VII pupil needs to be tested on a norm-referenced test for the broader picture to emerge. Sampling is attractive in that it will reduce the burden of testing time. It will also provide a background for comparison that will enhance possible Title VII use of criterion-referenced tests, which will assess spe-

cific instructional objectives instead of the broad program features.

A good criterion-referenced test should be tied in to a specific set of educational objectives. As with norm-referenced tests, the fit between the criterion-referenced test and the program's objectives and instructional practices must be determined. This fit, incidentally, may involve not only changing the test, but also restating school-level objectives where necessary.

Over time, these criterion-referenced tests should provide a more precise picture of pupil achievement in terms of specific instructional objectives rather than broad educational goals. A norm-referenced test, however, might continue to provide part of the information, especially that which looks at the program as a whole, with a criterion-referenced test providing information about the specific components and objectives of the school or project.

Since some audiences will continue to ask for ways to interpret criterion-referenced test scores so that they provide some normative or comparison basis, it may be necessary to devise methods of equating criterion-referenced scores with norm-referenced scores. With this kind of strategy, norm-referenced and criterion-referenced scores are part of the overall evaluation effort; and since two kinds of pupil scores will be available, the possibility of more meaningful assessment is increased. This technique helps overcome the problem of interpreting the meaning of criterion-referenced test scores mentioned earlier.

The strategies suggested above can reduce the amount of testing in Title VII schools while increasing its decision-making value for different audience levels, and, consequently, a more accurate demonstration of the impact of the program on the participating pupils.

In Baker et al., 1980, procedures were described by which project staff can examine/select norm-referenced tests in terms of their technical properties and their match with the program's objectives. The paper also discusses procedures for developing/selecting criterion-referenced tests.

Describing the Program that Led to the Gains

In the procedures described above, information on what students actually received in the program must be collected so that statements can be made about what parts of the program led to which outcomes. A documentation system is needed that will provide valid and reliable information about program implementation. The three basic approaches to documentation consist of information gathered directly from program participants, examination of program records, and observation. Information gathered from program participants can consist of staff reports, questionnaires, and interviews. Examination of records can consist either of record-keeping systems designed specifically for the program's documentation or they can consist of records that naturally evolve during the life of the program. Observations, which can be informal or systematic, usually take the form of checklists, coded reports, or delayed reports.

These kinds of documentation procedures, in conjunction with accurate assessment of pupil performance, should lead to Title VII program evaluations, which offer an accurate picture of the program's achievement, the educational significance of that achievement, and the particular program components contributing to achievement. This approach obviously relies upon the use of multiple measures and data-gathering techniques that let us assemble converging data. To the extent that it is **appropriate** and **feasible**, a combination of interviews, records, and observations can be used to generate information that supports or qualifies the picture of the program gained by each single approach.

Elsewhere (Burry, 1981), I have described the pros and cons of each of the above techniques, how they tie in with the larger evaluation effort, and the program situations in which one technique is more appropriate than another. I also offer some tips for designing, constructing, and testing each of the documentation techniques.

Selecting an Appropriate Evaluation Design[1] and Analysis Technique

An evaluation design describes **from whom** evaluation information will be collected, **with what** measuring device, and at **what times** in the life of the program. When an evaluator plans a design, he or she has to decide on: (1) groups or units from whom information will be collected, (2) measurement instru-

[1]The suggestions for design are adapted from a workshop (Winters et al., 1980) dealing with the planning, design, and conduct of an active bilingual program evaluation.

ments to be used, (3) times when instruments will be administered, and (4) appropriate procedures for analyzing data.

Evaluation design is a management tool for organizing data collection activities. The design specifies the questions to be answered by the evaluation as well as the information that best addresses these questions and incorporates techniques to make this information as credible as possible, given constraints imposed by setting or time. "Design" does not only refer to the statistics used for data analysis, although data analysis procedures do affect the credibility of evaluation results. Rather, design is a term for all the organizing activities related to information gathering: asking the right questions, identifying the information that will answer them, selecting or developing appropriate instruments, and applying appropriate data analysis procedures. Part of the analysis consists of interpreting the information, drawing conclusions, and making recommendations. The evaluation design should pay attention to the political context in which the program operates, the theories and assumptions of the program participants, and the program itself. It is essential to know what kinds of data gathering activities, from observation and interviews through achievement tests, will yield information most relevant to the questions guiding an evaluation. Also necessary for the credibility of the evaluation is a familiarity with alternative data analysis procedures and the inferences they support.

The need for credible information arises in several kinds of bilingual evaluations. As mentioned earlier, decision makers may be interested in program context, inputs, implementation, or out-

comes. The courts may be concerned with whether the program promotes effective participation in the educational process. Funding agencies need reliable data related to funds allocation (program inputs), program installation (implementation), and student achievement (outcomes). School administrators are concerned with student achievement and effective allocation of educational resources. Project staff, in addition to having the above concerns, also need information for use in on-going program improvement. Parents, whose participation is required in the planning and implementation of bilingual programs, are concerned with student outcomes and the school environment.

Each of the four evaluation concerns--context, input, implementation, and outcomes--generates different sets of questions to guide the evaluation. If decision-makers need information about program context, they may have some of the following questions in mind:

1. Why was the program installed?
2. Who is interested in the program and why?
3. What are staff and commmunity attitudes toward the program?
4. What bilingual education theories guide the program, and do staff members concur in their theories?

These questions get at the "climate" in which the program is operating, possible conflicting interests in program results, and relevant conditions existing prior to program implementation.

Funding agencies, administrators, and project staff often want to know about the quality and quantity of program inputs in order to assess the level of program implementation or the rela-

tionship of these inputs to program outcomes. Some input questions are:

1. Which languages are spoken by students and how well?
2. What is the home language used by students?
3. How many bilingual/ESL teachers are available, and what is their competence to teach in bilingual programs?
4. What kinds of materials are used?
5. How much money is spent on salaries, materials, aides, etc.?

These questions deal with the kinds and amounts of resources used in the program.

Program implementation is of special concern to the evaluator. One must know that the program really exists, what its goals are, and what it looks like before data collection can be planned. Program implementation data are also very useful for substantiating theories of bilingual instruction, assessing future program needs, and examining the relationship between program participation and student achievement. Implementation evaluation focuses on such questions as:

1. Is there a bilingual program in operation?
2. What are its major features?
3. How much time is being spent in various activities?
4. What are teachers' preferred teaching styles?
5. How are materials being used?
6. What are the patterns of teacher-student language use?
7. Is the program complying with legal guidelines?
8. Does the program need to be improved and, if so, how?

A final evaluation focus is on program outcomes--summative evaluation. This perspective, as I mentioned earlier, is perhaps the most familiar and certainly the one that often comes to mind when you think of "evaluation." The focal point of outcomes evaluation is on student achievement. As seen earlier, federal regulations require student achievement data. Evaluators, then, may be requested to provide information about:

1. How well the program promotes student achievement **vis-à-vis** a norm group.
2. How students in the program compare to those in a similar bilingual program.
3. How students compare to students receiving no special bilingual instruction.
4. The pattern of student achievement over time.

To summarize, bilingual evaluators need designs that provide credible information for a wide range of questions. A variety of information should be gathered in bilingual program evaluations because there are many diverse audiences with interests in program processes and outcomes. There is little baseline information available for decision makers to know what the program should be when fully implemented, what desirable and normal patterns of language growth in the native language and English should be, and what instructional strategies and materials are most positively associated with desired student outcomes. In addition, different audiences have different information needs. The bilingual program evaluator, therefore, must provide a variety of background, descriptive, process, and outcome information to augment subsequent program planning and evaluation efforts.

Differences between bilingual and monolingual evaluations and the threats to information validity posed by sample, instruments, and extraneous factors in bilingual settings might suggest that it is impossible to conduct a credible bilingual evaluation. But there are three powerful concepts that help counter many threats to validity. These concepts are: comparison, controlled assignment, and multiple designs.

Comparison. One way to avoid many of the threats to validity is to gather data for making comparisons between the program group and some external standard. The notion of comparison is powerful and deals with the threats to validity contained in sample, instrumentation, and extraneous factors.

Under ideal circumstances, the evaluator would find a group of students who were alike in all relevant educational characteristics, such as abilities and parental and home characteristics, assign a sample to a special program, or "treatment," while other samples would continue in the regular educational program. For example, the group under study may participate in a new program designed to improve reading comprehension. Other than participation in the new reading program, everything about the two groups would be the same.

The evaluator might ask if students in the reading program show greater growth in reading comprehension than students in the regular program. To answer this question, the evaluator might give both groups a reading comprehension test at the beginning of the year and then administer the test again at the year's end. If the control students' scores average 23 on the pretest and 29 on

the posttest, and the reading program students' scores went from 23 to 40 on the posttest, the differences in scores between the two groups would be evidence for the reading program's value.

It is, however, never possible to find identical groups of students who differ only with respect to a single educational treatment. Yet if judgments about the relative performances of two groups are to be fair, the groups must be as similar as possible. One way to enhance the comparability of the groups is via controlled assignment.

Controlled assignment. Since evaluators cannot really find groups with identical characteristics, they may use random assignment to ensure that both groups have similar characteristics and resemble, with only chance variations, the general diversity commonly occurring in the population they represent. The notion behind random assignment is that any member of the population (such as second graders in a particular school) has an equal chance to be selected for one of the groups (either the control or the treatment group).

Randomization in a school setting is not easily achieved. Students are not ordinarily assigned to classrooms randomly. Even within classrooms, they may be assigned to special treatments on the basis of need rather than randomization. The situation is further complicated in most bilingual programs if all students eligible for the program must be served, effectively precluding use of a control or comparison group.

But the notion of controlled assignment does address the problem of finding an appropriate comparison group for the evalua-

tion. If the process by which groups are formed is known, for example, then actual differences between the groups can be controlled or explained; here the evaluator can use naturally occurring equivalent groups for comparison. An example of naturally occurring equivalent control groups is a comparison between students enrolled in a bilingual program in one school and students enrolled in a similar bilingual program in another school in the same district (note that the similarity addresses both students and their program). Or, you might find an "equivalent" comparison group by locating another classroom in another school or district whose students have the same language proficiency levels, native language background, and socio-economic status as those enrolled in the bilingual program. Randomization, matching, and selection of comparison groups on the basis of a cutting score are all instances of controlled assignment. Any controlled assignment procedure can be applied to different kinds of groups, to students, to classrooms (without regard to how the students in them got there), to schools, or to districts. If the process by which the comparison groups were formed is known, and the potential systematic differences between them are documented (using the kinds of procedures mentioned earlier), you will be able to reach fairly good conclusions about student achievement based on these comparisons.

Multiple designs. The third concept for improvement of design is the use of multiple designs. The information needs of bilingual programs are extensive, and they require several kinds of information that probably cannot be addressed in any single de-

sign. Since design refers to preselected units of observation measures, the timeline for data collection and data analysis plans, it is unlikely that any one particular combination of timeline, sample, instruments, and analysis would be appropriate for answering several different evaluation questions. For example, the evaluation design for gathering information on the question, "Is the program being implemented as planned?" will differ from one that asks "How well are students who learn to read in their native language doing when compared to those instructed only in English?" In the first instance, it will be necessary to get observation and questionnaire data about teachers collected at frequent intervals. The data analysis will consist, in part, in matching what was discovered happening in the program to the official course description as stated in the funding proposal and/or course of study. Needed for the second question, which may be asked during the same evaluation, will be achievement test data on English reading and reading in the native language as well as some of the descriptive information described above. These data will probably be collected at the beginning and end of the year from students enrolled in the bilingual program as well as NEP/LEP students not receiving reading instruction in their native languages. Data will be analyzed by testing for significant differences between group scores on the posttest.

Several data collection plans can be incorporated into the evaluation, each chosen to collect information from a particular group by a particular method at a specified time.. Some of these designs may include random assignment within the program in order

to answer research questions. Other designs will involve neither comparison nor controlled assignment because they are intended to gather information to be used only within the confines of the particular program.

In planning the evaluation, one must consider what will be asked about the program. Will comparisons need to be made between program students and students not in a bilingual classroom? Between program students and a norm group? Must the trend of language acquisition be shown or the parts of the program contributing to student achievement be identified? Is it necessary to know how people feel about the program and why? Each of these information requirements leads to different statistical analysis procedures.

Generally, both descriptive and inferential statistics should be reported. Both instruments and the kinds of data they provided should be described before making judgments about program effects. This preliminary data analysis will include instrument statistics such as inter-rater reliability for all constructed response measures, internal consistency, and perhaps test-retest (pre- and post) reliability for achievement tests, decision or classification consistency for criterion-referenced measures, item difficulties, as well as group means and standard deviations.

In addition to descriptive statistics about your instruments, you will want to provide intercorrelations among dependent variables as an indication of their similarities and differences. You will also examine all the data collected in order to answer the following:

1. What general patterns emerge that can be examined with inferential statistics?

2. Are there missing data, unusual frequency distributions, small sample sized, or restricted variance that will affect the interpretation of evaluation results?

3. Do the data to be used for making inferences meet the requirements for inferential statistics?

Deciding on which kind(s) of statistical analysis technique to use is often seen as an overwhelming problem--beyond the scope of the project staff. Perhaps this is because of the esoteric language often used in discussion of statistics. I'm not saying that statistical analysis is easy, but it need not be bothersome. If project staff do not have expertise in statistical analysis, then they should look for expert help. If they do look for expert help, then it is critical that the staff have some background in the use of statistics so that they will know the right questions to ask of the expert, to assure the most appropriate statistical analysis, given the individual project, its constraints, and its information needs.

The selection of an appropriate statistical technique is governed by: (1) the data the project is able to collect, and (2) the variables they are examining.

A project can normally collect three kinds of data: nominal, ordinal, or interval. Nominal data means the information names something--it doesn't measure, it only gives names. The information is classified into categories with no necessary relationship between those categories. For example, we might say that a classroom appears to have some happy versus unhappy children.

Ordinal data means the information has been ordered according to rank, with a categorization of these things in terms of **more than** or **less than.** For example, we might have a measure that rank-orders degrees of student self-esteem.

Interval data not only tell the order of things, but also tell the interval or difference between the judgments. For example, if one pupil scores 87 and another scores 77, the first student has not only performed better, but has also scored better by 10 points. Rating scales and achievement tests are examples of devices that give interval data.

Data from one scale can, if necessary, be converted **downwards.** That is, interval data can be legitimately converted to ordinal or nominal. Data, however, should not be converted **upwards,** for example, from ordinal to interval.

An evaluation might cover any of the following variables: independent, dependent, or moderator. An independent variable is the stimulus variable or input. It is the thing that is examined to see its relation to something (e. g., a test score) that we observe. Program and educational interventions are examples of independent variables.

A dependent variable is the response variable or output. It must be observed and measured to find the effect, or the accomplishments, of the independent variable such as a program component. Student performance on an achievement test is an example of a dependent variable.

A moderator variable is a special kind of independent variable that may modify the relationship between the independent and

dependent variables. Student socio-economic status is an example of a possible moderator variable. (For statistical purposes, moderator variables should be considered to be independent variables.)

Statistical analyses can be parametric or nonparametric. Parametric analyses, in short, make certain assumptions about the data and make strict demands. For example, use of a parametric technique might be based on the assumption that student scores were drawn from a normally distributed population, that is, the normal, bell-shaped curve, or that both sets of scores were drawn from a population having the same variance, that is, the same spread of scores. Using a parametric technique assumes the existence of interval data. Nonparametric techniques, on the other hand, do not require that data be normally distributed or that the sample variances be equal.

Frequently used parametric techniques include the t-Test, parametric correlation (Pearson product-moment), and analysis of variance. Frequently used nonparametric techniques, corresponding in use to the three parametric techniques mentioned above, are the Mann-Whitney U-Test (a kind of nonparametric t-Test); Spearman rank-order correlation; and Chi square.

These kinds of tests are intended to show the statistical significance of something; for example, the difference between a pre- and a post-measure of a program component. They show whether the differences seem to be the result of chance or whether they are more likely to be the result of the program. The greater the

level of statistical significance, the more likely it is that the differences are the result of the program.

Selection of the appropriate statistical technique will depend on the kind (nominal, ordinal, interval) and number of dependent variables we have and the kind and number of independent variables. Since the mix of variables and kind of data sometimes indicates only one appropriate technique, and sometimes offers alternatives to select from, decisions about the whys and wherefores of analyses should involve discussion between staff and evaluator.

The Range of Designs[2]

Figure 1 lists a series of possible evaluation designs, the kinds of comparative data they generate, and the requirements that cut across designs. Figure 2 shows the kinds of threats to validity for which each design is intended to compensate.

Designs Recommended for Bilingual Programs

The following are some suggested designs suitable for gathering information related to the major questions most often addressed in bilingual evaluation. By examining the strengths and weaknesses of these designs as they relate to the particular evaluation concerns at hand, they can be adapted and/or combined as needed.

---

[2]Use of these designs is amplified in Winters et al., 1980.

Figure 1

REQUIREMENTS OF TYPICAL DESIGNS

Design	Comparative Data Obtained
Pretest/Posttest Control Group	Equivalent Group
Posttest-Only Control Group	Equivalent Group
Single Cohort Comparison	Equivalent Group
Multiple Cohort Comparison	Equivalent Group
Nonequivalent Control Group	Nonequivalent Group
Regression Discontinuity	Nonequivalent Group
Regression Projection	Nonequivalent Group
Normed Pretest/Posttest	National Norm Group
Normed Pretest, Local Pretest/Posttest	National Norm Group
Minimum Competency	Prespecified Achievement Standard
Time Series/Longitudinal	Withen Group
Multiple Time Series	Withen Group
Exposure to Treatment	Program to Itself

Requirements for all Models

1. Theory/knowledge of developmental rate of skill being examined.
2. Evidence that the program existed.
3. Description of program and sample "mortality" with reasons, if possible.
4. Documentation of activities that may produce competing explanations of results.
5. Valid and reliable instruments.
6. Explicit criteria by which program "success" will be judged.

Figure 2

RELATIVE ADVANTAGES OF EVALUATION DESIGNS
(Threats to Validity Countered by Design)

	Sample			Instrumentation			Extraneous	
	Selection	Maturity	Mortality	Validity	Reliability	Admin. Procedures	Concurrent Programs	Hawthorne Effects
Pretest/Posttest Control Group								
Posttest Only Control Group								
Single Cohort Comparison								
Multiple Cohort Comparison								
Nonequivalent Control Group								
Regression Discontinuity								
Regression Projection								
Normed Pretest, Posttest								
Normed Pretest, Local Pretest/Posttest								
Minimum Competency								
Time Series/ Longitudinal								
Multiple Time Series								
Exposure to Treatment								

The four design types suggested for bilingual programs are summarized in Figure 3. Each type was selected to address a different, but frequently encountered, bilingual evaluation purpose. The designs are presented in order of utility, from general to specific. The most general-purpose designs are the time-series/longitudinal for investigating program outcomes. The exposure-to-treatment design is widely applicable for formative evaluations. Accountability designs, for reporting information about student achievement, either in terms of local program objectives or national norms (when required), are also presented.

For each design, the figure also specifies the comparison implied by the design as well as the requirements that need to be met if the design is to be used.

CONCLUSION

As evaluators of bilingual programs begin to plan, they must consider the various agencies--federal, state, and local--who direct the program. The legitimate differences that exist among programs make it inappropriate to apply one set of design or test criteria to all bilingual programs. The evaluator must consider the particular needs of the program evaluated, recognize the inherent difficulties in design and testing, and plan an evaluation appropriate to the program and feasible within its constraints.

It is critical that the goals and objectives of a bilingual education program be made explicit so that valid criteria for determining program success or failure can be established. The goal

Figure 3

RECOMMENDED DESIGNS FOR BILINGUAL PROGRAMS

Evaluation Purpose	Design	Implied Comparison	Requirements of the Design
All-purpose design to provide documentation of program development, implementation, and student achievement over time.	Time Series/ Longitudinal	Within group or age/ grade cohorts	1. Way to trace/document departure of students from bilingual or comparison groups to explain "mortality." 2. Tests/instruments that provide comparable data over time, e.g., same test with many levels and alternate forms, observation scales that remain the same across the period of the program evaluation, or highly correlated
Identifying program strengths and weaknesses in relation to student achievement.	Exposure to Treatment	Parts of the program with each other	1. Identified theory of what variables contribute to bilingual student achievemen 2. Achievement instruments that measure program goals. 3. Data gathering techniques that identify qualities of program implementation such as time on task, amount of curriculum covered, etc.
Accountability to school or district administration.	Minimum competency	To prespecified minimal performance standard and/or to students in district outside the bilingual program.	1. Basic skill/minimal competencies that bilingual students are responsible for completing. 2. Reliable and valid minimal competency test. 3. Defensible minimal competency standard of excellence.
Reporting required norm-referenced test data for Title VII programs.	Normed pretest, Local pretest/ posttest	National norm group. Other students taking local pretest and posttest.	1. Norm-referenced instrument appropriate for measuring bilingual program goals. 2. Reliable and valid local criterion-referenced test that matches program goals. 3. (Optional) Local norms for criterion-referenced test.

statements should include outcomes beyond those specifically concerned with student achievement.

Evaluators working in a bilingual setting must also become familiar with a variety of program features frequently absent in a unilingual setting. They must also build into the evaluation plan a variety of formative tasks. These tasks should include examination of program implementation to suggest areas of program improvement and careful documentation of the implemented features and their relationships. Evaluators must also attempt to improve their summative evaluation activities, for example, in their estimations of which program features contribute to program outcomes.

Evaluators of bilingual programs must ensure that the appropriate program features or variables are selected for evaluation. The evaluation itself--design, measures, analyses, and reporting--should be technically sound, acceptable under existing regulations, and sensitive to the needs and constraints of the program under examination.

These aims can be facilitated by: reviewing the operant regulations; examining the program plan or description of the program (if such a plan or description exists); determining (in consultations with program participants) which features of the program need to be evaluated; deciding upon the means to evaluate these features; selecting methods to ensure ongoing documentation of their implementation, relationships, and cumulative effects; and considering appropriate means of reporting evaluation information to various audiences.

Evaluation planning will be more feasible and effective the earlier the evaluator is involved in the program. If possible, evaluator involvement should occur before program implementation so that the evaluation plan is an integral part of the program's operation. The kind of evaluation planning described here also relies upon a program plan or description. Such program descriptions may not provide sufficient information to permit adequate evaluation planning without discussions with program staff to elaborate intentions, achievement strategies, means of implementation, and expected outcomes. In short, the kind of evaluation planned will be influenced by the time at which the evaluator enters the program and the existence of observable and measurable program processes and outcomes.

How smoothly a bilingual program's evaluation is managed depends on the care devoted to its planning. Information use will be enhanced to the extent that it is reported quickly to the appropriate decision makers.

Bilingual education programs are extremely complex and address a great variety of needs and methods. Evaluation should provide greater documentation of bilingual programs in terms of their intentions and their implementation. It must clarify exactly what a particular program is to accomplish and how the program intends to meet its goals. Tests and other measures must be selected according to technical properties and relevance to the individual bilingual program. Test results should be interpreted in light of the program's objectives. In evaluation design, variables must be specified and controlled. Design features

should allow for information to improve the program (formative evaluation) and to show program effect (summative evaluation).

These procedures, if implemented at the level of the local project, will lead to more useful evaluations. Over time, they should lead to an evaluation strategy consisting of: (1) information on pupil performance before entering the program; (2) information on how much instruction they received in the program, the manner in which they received it, and the context in which it was provided; and (3) information on pupil performance after the program. Gains at the end of the program could then be attributed to instruction of a certain kind.

REFERENCES

Baker, E. L., L. P. Polin, James Burry, and C. Walker. **Making, Choosing, and Using Tests: A Practicum on Domain-Referenced Testing.** Los Angeles, California: Center for the Study of Evaluation, University of California, Los Angeles, 1980.

Burry, James. **Evaluation and Documentation: Making Them Work Together.** Los Angeles, California: Center for the Study of Evaluation, University of California, Los Angeles, 1981.

_____. "Evaluation in Bilingual Education," **Evaluation Comment**, VI, No. 1 (1979), 1-14.

Cronbach, L. J., and P. Suppes, eds. **Research for Tomorrow's Schools: Disciplined Inquiry for Education.** New York: Macmillan, 1969.

Federal Register. Washington, D. C.: Government Printing Office, March 29, 1979.

Winters, L., L. Spooner-Smith, and D. Elvenstar. **Planning for Evaluation Design.** Los Angeles, California: Center for the Study of Evaluation, University of California, Los Angeles, 1980.