This paper is directed to those who are undertaking evaluation of a bilingual program for the first time or who have already struggled with the mysteries of such an undertaking. Emphasis is given to the reporting requirements of the various federal and state funding agencies. The bilingual-bicultural program structure is defined so the evaluator can see the interplay of program prototypes, student language facility, and instructional approach. The evaluation process is divided into an explication of evaluation models, evaluation design, and instrumentation. Examples of each of these process components are given. (Author)
EVALUATING BILINGUAL PROGRAMS

Alexander I. Law

ABSTRACT

This paper is directed to those who are undertaking evaluation of a bilingual program for the first time or who have already struggled with the mysteries of such an undertaking. Emphasis is given to the reporting requirements of the various federal and state funding agencies. The bilingual-bicultural program structure is defined so the evaluator can see the interplay of program prototypes, student language facility, and instructional approach. The evaluation process is divided into an explication of evaluation models, evaluation design, and instrumentation, and examples of each of these process components are given.

INTRODUCTION

Bilingual instructional programs are not a new phenomenon in the United States. Indeed, such programs existed as early as the 1890s. Those first programs were few, aimed at unique and small populations, and frequently taught in parochial schools, primarily as an attempt to maintain a particular ethnic identity. It was not until the last decade, with its attendant educational reforms and funding efforts, that bilingual programs flourished and received recognition. With this recognition came the demand for accountability and evaluation. Bilingual programs are sufficiently unique that there are troublesome problems attendant upon their evaluation. This paper will address these problems and offer some guides to a more effective evaluation effort.

HISTORICAL DEVELOPMENT

The development and full flowering of bilingual educational programs are relatively recent in American education. The melting pot theory of cultural and linguistic assimilation, which dominated education and culture in the United States until the 1960s, required instruction and reports of its results to be in English.

The massive educational reform triggered by the Elementary and Secondary Education Act (ESEA) of 1965 was the first systematic effort to identify and treat the educational deficiencies of students with problems stemming from inadequate command of the English language. The earliest such efforts were typically labeled "ESL," or "English as a Second Language." ESL's main instructional objective was the development of competence in both written and spoken English. The assessment of such programs was relatively simple, since the attainment of the objective could be readily determined through a variety of existing measures, ranging from standard vocabulary lists (for determining acquisition of sight vocabulary) to the whole set of available achievement tests (for determining reading comprehension skills in English).

These early—and often primitive—components of Title I were soon augmented by the more comprehensive thrusts of federal funding under the 1967 amendments to P.L. 89-10, which created Title VII, the Bilingual Program. Under the auspices of this act, the concepts of instructional intervention were enlarged from the ESL focus to a multicomponent program including staff development, community involvement, and development of instructional materials. This response to Title VII was the genesis of
the programmatic effort known today under the general rubric of bilingual-bicultural programs.

Friedman and Lovas (5) have identified a continuum of programs ranging from ESL to full bilingual-bicultural. Martinez and Houssen (6) trace the evolution of the various efforts, and call for a clarification of the definition of bilingual-bicultural. In addition, they propose a multidimensional evaluation framework integrating instructional approaches, program types, and student language facility. They warn that fully effective evaluation is in jeopardy until criteria mutually agreed upon by evaluators and bilingual educators are established.

As the nature and variety of program types proliferate, so do the evaluation problems—especially when information must be aggregated across programs, as in federal, state, and other large-scale endeavors. Evaluation has progressed from a relatively simple process of describing a program and judging its worth to an evaluative-research process where one is asked not only to make a statement about the worthiness of an endeavor but to contrast its effect with that of other instructional methods or programs. The American Institutes for Research (3), through a contract with U.S.O.E., searched for effective bilingual-bicultural programs for dissemination. Only about five percent of the existing programs could present data to support judgment as to their impact. While the criteria used were rigorous, they were not unreasonable; evidence from the exemplary programs had to show the following outcomes:

Evidence of bilingual program impact should be based on objective measurements obtained from sizeable pupil samples. Achievement gain measures should be estimated for program participants and for a comparable control group. Well-designed contrasts with pre-program baseline or comparison with appropriate norm reference groups are also acceptable. It is necessary that gains for program participants be significantly greater than gains for the control or comparison group.

Interpretation of the significance of the reported gains depends on customary psychometric and statistical grounds. Measurements should be reliable and valid. Tests should be of appropriate difficulty level for the groups examined. The reporting of achievement in either grade-equivalent or raw-score scales is acceptable; one scale is essentially a linear transformation of the other except at extreme ranges. Average gains for pupils in the comparison groups should be unbiased estimates of the gains for the total population of participants; that is, missing data or the effects of selection should not be great enough to cast doubt on the findings. Confidence in the generalizability and potential for replicability are also greater when results are reported for several classes and grade levels, so that unique teacher or administrator effects can be ruled out.

Statistical significance should be demonstrated so that one may confidently conclude that the results showing superior program effect did not occur by chance; that is, results showing significant program effect when in fact there is none, should occur no more than five percent of the time. In addition, mean gain differences between program and control groups must be educationally relevant whether reported as grade equivalents or as relative within-group standard deviation units. For example, mean differences of the order of one-half grade equivalent, or one-half standard deviation, are meaningful, as is a large positive shift in mean percentiles between pre- and posttests when program outcomes are compared to those of norm-reference groups. [Pp. 8-9]

While these criteria are acceptable for judging the impact of exemplary programs, one could argue that their rigor is not necessary for unique local efforts. However, the frustration of federal authorities with the seeming lack of demonstrably successful efforts has culminated in the issuance of new regulations regarding the conduct and evaluation of programs carried out under the aegis of Title VII. The regulations issued in April 1976 state in part:

(iii) A description of the evaluation design of the proposed program. Such evaluation design shall include provisions for assessing the applicant's progress in achieving the objectives set out in its application for assistance. In the case of an application to carry out the activities described in §123.12(a), the evaluation design shall also include the following:

(A) Provisions for comparing the performance of participating children on tests of reading skills in English and in the language other than English to be used in the proposed program with an estimate of what such children's performance would have been in the absence of the program. Where the applicant chooses to base such estimate on the performance of nonparticipating but similar children on such tests, the evaluation design shall include a description of the methods used to identify nonparticipating but similar children for such purpose;

(B) A description of instruments of measurement to be used by the applicant in evaluating the performance of participants in the program, the rationale for selecting these instruments, and procedures to be followed in their use; and

(C) Provisions for reporting pre-test and post-test results on reading tests for all participating children (and, where their performance is compared with the performance of nonparticipating but similar children for all such nonparticipating children) using mean scores, standard deviations and appropriate tests of statistical significance. No application which fails to include the elements of an evaluation design described in this paragraph will be approved for assistance under this subpart. [P. 14990]

Clearly, the direction from the federal authorities is toward more rigor and toward an evaluation-research approach.

Compounding the programmatic difficulties, both state and local, is the recent Lau v. Nichols decision by the
Supreme Court, which, in essence, states that a conventional ESL approach is no longer sufficient, but that the instruction must, where necessary, be conducted in the language of the student. The Lau remedies require, as a minimum, that:

1) Schools systematically and validly ascertain which of their students are linguistically different;
2) Schools systematically and validly ascertain the language characteristics of their students;
3) Schools systematically ascertain the achievement characteristics of their students; and
4) Schools match an instructional program to the characteristics as ascertained.

These remedies, taken together, provide the general framework on which to build an acceptable evaluation plan.

THE BILINGUAL INSTRUCTIONAL PROGRAM

The first requisite of an evaluation plan is an explication of what is to be evaluated. Typically, the bilingual instructional program has been poorly defined. To assist the evaluator, Martinez and Houssden (6) have conceptualized a multidimensional framework. The dimensions are:

<table>
<thead>
<tr>
<th>Program</th>
<th>Student Language Facility</th>
<th>Instructional Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transitional</td>
<td>Non-English Speaking</td>
<td>Translation</td>
</tr>
<tr>
<td>Monoliterate</td>
<td>Limited English Speaking</td>
<td>Preview-Review</td>
</tr>
<tr>
<td>Partial Bilingual</td>
<td>Fluent English Speaking</td>
<td>Concurrent</td>
</tr>
<tr>
<td>Full Bilingual</td>
<td></td>
<td>Back-to-Back Language-Other-Than-English Immersion Eclectic</td>
</tr>
</tbody>
</table>

The program prototypes can be considered as a continuum from emphasis on instruction in the dominant language of the surrounding culture to equal competence in both languages of the student. The types of programs are described as follows:

**Transitional.** The native language is used in the early grades only to facilitate the mastery of the subject matter, so that the child may eventually be phased into a curriculum totally reflecting the second language.

**Monoliterate.** These are programs that address aural-oral competency in the native language, but focus on the attainment of literacy only in the second language.

**Partial Bilingual.** The goal is to attain aural-oral competency and literacy in both languages, but restrict literacy to subject matter relevant to cultural heritage, i.e., social science, literature, and art.

**Full Bilingual.** The goal is to attain aural-oral competency and literacy in both languages in all content areas (including mathematics and science).

The language facility of bilingual students is necessarily an essential point of focus. Three language facility categories are sufficient, given the current state of the art of assessing language facility of bilingual students. They are:

**Non-English Speaking.** A student who is incapable of appropriately reacting to statements or directions given by a teacher in the English language because of the inability to decode verbal English language messages and because of the inability to cognitively relate an idea in a language other than his/her primary language is considered to be a non-English speaking student.

**Limited-English Speaking.** A student who has not developed English language skills of comprehension, speaking, reading, and writing sufficiently to benefit from instruction only in English and who comes from a home where a language other than English is spoken is considered to be a limited-English speaking student.

**Fluent-English Speaking (Bilingual).** A student who can learn equally well through use of the English language as through his primary language is considered to be a fluent-English speaking student.

The six instructional approaches are:

**Translation.** Lessons are presented in English then translated to a second language. These may be done simultaneously, at a later time during the day, or even on another day.

**Preview-Review.** Students receive instruction in two languages in any specific lesson or subject area. A review is presented in one language, followed by a lesson in the second language. Finally a review may be done either in both languages, or only in the language of the preview. Usually two language-model instructors are used in this format. Students in one group may be presented in one language on the content or context of a subject and the presentation is reviewed in the primary language by the appropriate adult model.

**Concurrent.** Both languages are used simultaneously in the instruction of any specific lesson. The objective is to teach concepts in both languages, avoiding trans-
Back-to-Back. A designated portion of time, such as in the morning, is set aside for instruction in one language and another portion of the day is devoted to instruction of the same curriculum content in the other language. The student receives instruction in two or more languages, but at different times during a day.

Language-Other-Than-English Immersion. A language other than English is used for instruction in academic areas with concentrated English-as-a-Second-Language development component.

Scriven for his formative-summative distinction and Stake for the describing and judging focus. Again, these models achieve these goals. Stufflebeam further states:

Selecting an Evaluation Model

There are three steps the evaluator should take following the explication of the nature of the program: 1) select an evaluation model, 2) establish an evaluation design, and 3) select the appropriate instruments. The first, a model, will provide a framework which will assist the evaluator by providing a coherent course of action.

A useful summary of recently proposed evaluation models can be found in Worthen and Sanders (12). Each has its proponents. Many evaluators will select from this assortment of models those parts that seem most appropriate to their particular situations, and then construct their own. For example, Stufflebeam's (10) CIPP model will direct the evaluator to consider the context, input, process, and product evaluations. One can turn to Tyler (11) to assure that a focus is given to determine objective attainment. The evaluator should be generally familiar with other models, particularly those of Scriven (8) and Stake (9). Scriven for his formative-summative distinction and Stake for the describing and judging focus. Again, these models will be useful in providing a framework but should not be adopted intact in lieu of creating the unique plan necessary by the nature of a bilingual evaluation.

The models are general statements designed to guide the evaluation process; to help the evaluator systematically plan, identify critical questions to be answered, and gather and analyze the data to answer these questions.

It would be instructive to take one of the models, the CIPP model by Stufflebeam, and discuss its application. This is the most comprehensive model, considering four evaluation types or approaches which, when taken together, form a single model.

The Context evaluation (the C of CIPP) should be considered, although it is often overlooked in the total evaluation scheme. Among its many characteristics noted by Stufflebeam (10) are descriptions and analyses of the system to be evaluated, descriptions of goals and objectives, and a focus on the factors known to be important for achieving these goals. Stufflebeam further states:

Context evaluation provides a basis for stating change objectives through diagnosing and ranking problems in meeting needs or using opportunities, and it analyzes change objectives to determine the amount of change to be effected and the amount of information grasp available for support. Thereby, it provides an initial basis for defining objectives operationally, identifying potential methodological strategies, and developing proposals for outside funding. [Pp. 219]

Following the Context evaluation, the evaluator is directed to the Input evaluation. Like the Context, it has several facets. Generally it can be summarized as

identifying and assessing 1) relevant capabilities of the responsible agency, 2) strategies for achieving program goals, and 3) designs for implementing a selected strategy. This information is essential for structuring specific designs to accomplish program objectives. [Pp. 222-3]

The Context and Input evaluations are in a sense idealized. Frequently the evaluator comes to a situation where the objectives have been stated, the instructional framework set, and resources committed. The previously discussed instructional framework should be sufficiently comprehensive that the evaluator can infer from it the instructional strategies. These strategies will determine the specific designs to be used to meet the program objectives.

The key in this process is to determine and get agreement on the objectives, the "what" that is to be assessed and judged. The Context and Input evaluation yields information that sharpens the focus of the process and product evaluation phases. It provides a background that goes beyond the instructional objectives to help the evaluator define other areas of interest—costs, benefits, attitudes of participants, involvement of parents, and whatever other important objectives have been identified.

The Process evaluation is perhaps more familiar to the reader. Process evaluation provides the feedback of information about the program as it evolves. Stufflebeam posits three main objectives for Process evaluation:

Process evaluation has three main objectives—the first is to detect or predict defects in the procedural
design or its implementation during the implementation stages, the second is to provide information for programmed decisions, and the third is to maintain a record of the procedure as it occurs. [P. 229]

The Process evaluation strategy is one of flexibility—not a fixed or rigid process. It is more a sensing process than strictly formal measurement. Ideally, the evaluator is independent of the program or project staff, yet is in constant communication with them. The instruments used can be less formal than objective tests (though objective tests have an important role if used judiciously), typically consisting of interviews, questionnaires, school records, parent contacts and reactions, relationships with community agencies, records of utilization of instructional materials, and so on. Reporting may be formal or informal but it must be continuous. The most critical phase is early in the implementation of the program.

Product evaluation is the process that is most familiar to the reader. It is, however, broader than typically conceived. It occurs not only as the terminal assessment and analysis process but also during the implementation of the program, as certain previously determined check points are reached. Scriven (8) makes the distinction between instrumental accomplishments at an intermediate level and consequential (the terminal assessment of fundamental attainments).

Reviewing some of the important concepts in bilingual education and placing them in the CIPP framework will give the reader a sense of direction. The following questions are appropriate for each type of evaluation:

Context: What are the values and goals held by the system as related to bilingual instruction? What are the desired and actual conditions in the environment, for example, how many students need a bilingual program? Of what type? What information is needed or exists about the system, and state and federal regulations and guidelines? What is to be the role of the evaluator?

Input: What existing capabilities does the system have; for example, faculty capability in a second language? What instructional framework is appropriate? What overall design for instruction and evaluation is desired and feasible?

Process: Is the program proceeding as scheduled? What problems exist in implementation? Are instructional materials adequate and in place? What are the attitudes of key people—parents, student, teachers, and administrators? Are initial instructional units effective? (Note: This may also be a Product evaluation.)

Product: What are the attainments of the pupils? Have the objectives been met? What decisions can be made?

Establishing the Evaluation Design

The models give general direction: the design yields a specific plan for data collection and analysis. It would not be appropriate to call this process an experimental design; that label would be accurate only in rare situations. True experimental designs call for the establishment of treatment and control groups, random assignment, and the like. Such luxuries rarely occur in typical public or special programs. Should the exceptional case arise, a variety of classical experimental designs is available.

I believe the very nature of bilingual-bicultural education often mitigates against a rigorous experimental-control, random assignment design. The population is unique, culturally and linguistically, preventing a readily available comparison group. The instructional program is typically loosely defined, and the measurement problems are acute. The evaluator must be ingenious in constructing a design that is both feasible and capable of yielding evidence of the program’s impact or lack of it.

Realistically, the best approximation available is the series of designs proposed by Campbell and Stanley (2). The most common of these is a pre-post assessment with some type of comparison group. The comparison group could be a group of students who are not in the program but who have characteristics similar to those of the participants.

The interrupted time series design is another possibility. Most simply, this design calls for a series of measurements of student performance before a treatment, intervention of the treatment, and then measurements of performance after the treatment. Schematically it looks like this:

\[(\text{Measurement} - M - M - (T)\text{treatment} - M - M - M)\]

Historical measure: could be derived from existing school records.

An alternative is to use the participants as their own control, by establishing an expected level of attainment and contrasting the actual, or obtained, level with this expected level. Such a process gives an indication of the attainment of objectives but provides only a minimum of statistical evidence.

Popham (7) provides an excellent discussion of the Campbell and Stanley quasi-experimental designs.

Selecting the Appropriate Instruments

The most difficult aspect of any bilingual evaluation is assessing pupil progress in the instructional components of reading or language acquisition. External funding sources require some quantitative estimate of pupil growth both in English and the native or home language. There are few, if any, appropriate norm-referenced tests in languages other than English. Limited English-speaking students, in some cases, may be assessed using available norm-referenced English-language tests. The lack of appropriate norm-referenced instruments clearly suggests that any quantitative assessment must rely heavily on data collected from locally constructed or criterion-referenced instruments or items.

The use of such instruments, without extensive statistical treatment, precludes normative comparisons. Although the absence of normative comparison does not...
C.3
REFERENCES


