Review of Educational Research
Summer 1992, Vol. 62, No. 2, pp. 181–203

Review of Practices and Problems in the Evaluation of Bilingual Education

Tony C.M. Lam
University of Toronto

The quality of evaluation studies of the effects of bilingual education for language minority students, similar to the quality of educational evaluation in general, has been generally poor. This conclusion is based on the finding that, for the eight studies that reviewed the literature on the effectiveness of bilingual education, the mean percentage of evaluation or research reports which were judged methodologically acceptable for inclusion in these studies was 10% (median = 6%). A review of past and present publications, and federally funded projects, shows that some national efforts have been devoted to the improvement of bilingual education evaluation but that such efforts have apparently been unsuccessful. This article argues that the lack of sound and practical guidelines and materials, which precludes adequate technical assistance, is one cause of the inferior quality of evaluation practices. Other contributing factors are incompetent program evaluators, misinformed local administration, inappropriate state and federal policies, and the complex issues involved with bilingual education itself.

It has been 22 years since the passage of the Bilingual Education Act in 1968, when direct federal grants began funding local school districts to develop Title VII bilingual/English-as-a-second language (ESL) programs designed to meet the education needs of students with limited English proficiency (LEP). Title VII bilingual grant funding supports the delivery of special instructions to a small minority of LEP students. State and local funding also provides support. The full range of instructional services commonly provided to LEP students includes ESL classes, native language support in content-area classes, sheltered English and sheltered mainstream classes, subject matter development in the native language, pullout tutorials, and combinations of these approaches. These educational services range from merely teaching English as a second language to teaching the student's native language as well as English, with the true bilingual goal of maintaining both languages and, where feasible, increasing the proficiency in the first language as well as in English.

P.L. 100–297, signed in April 28, 1988, established six types of elementary and secondary bilingual education programs. (a) Programs of transitional bilingual education provide structured English language instruction and instruction in the child's native language with the intent of facilitating English language competence. (b) Developmental bilingual education programs are full-time programs that teach English and one other language in a manner designed to develop competency in both languages. (c) Special instructional alternative programs are designed for particular

This article is based on the author's work on a project supported by the U. S. Department of Education (Contract No. 300-85-0140). The author would like to thank Kast G. Tallmadge, Nona N. Gamel, Gary Hargette, Kim O. Yap, and the three anonymous reviewers of this article for their input, as well as Marta Field and Bruno J. Bornet for their typing and editing.

linguistic and instructional needs. (d) Programs of academic excellence, which include programs of transitional bilingual education, developmental bilingual education, or special alternative instruction, have established a record of excellence and therefore can be referred to as models of effective bilingual educational practices. (e) Family English literacy programs are designed to help limited English proficient adults and out-of-school youths achieve competence in the English language, sometimes with the added intent of instructing parents in order to facilitate the educational achievement of their limited English proficient children. (f) Bilingual programs are also established that are designed to meet the instructional needs of LEP students who are also handicapped or gifted and talented.

The Title VII program is one of several federally funded programs in education that stress the importance of evaluation. It demands that every proposal include a detailed plan for demonstrating program effectiveness. Moreover, it was the first program under the Elementary and Secondary Education Act to require an independent educational accomplishment audit. Although this requirement was subsequently dropped, evaluation requirements continued to be spelled out in the 1977 and 1980 program regulations and in the 1978 and 1984 amendments to the Bilingual Education Act.

Unfortunately, in spite of this emphasis, local evaluations in bilingual education, as distinguished from large-scale evaluation research studies which focus on impact assessment for policymaking decision purposes, have been inadequate (Baker & de Kanter, 1983). Some skeptics have described them as useless—not worth the paper they are written on (Epstein, 1977). Others have concluded that local evaluation reports are of little value to decision makers, whether at the local or federal level (Government Accounting Office [GAO], 1976). In a study of the utility of Title VII evaluations for decision makers, Alkin, Kosecoff, Fitz-Gibbon, and Seligman (1974) found that local staffs rarely used the information provided by the annual reports to plan and revise programs for subsequent years.

Although data have been accumulated for many years, the poor quality of the evaluation efforts has severely hampered attempts to draw conclusions about the impact of educational interventions designed to serve LEP students (Okada, Besel, Bachelor, Glass, & Montoya-Tannatt, 1983; Rodriguez-Brown, 1980; U.S. Department of Education, 1982). Although one meta-analysis (Willig, 1985) is optimistic regarding the efficacy of bilingual education, debate continues over the merits of these programs—in particular, the issue of teaching two languages versus English only. Arguments based on limited and inadequate empirical information characterize this debate (Baker & de Kanter, 1981; Dulay & Burt, 1978; Epstein, 1977; GAO, 1976; Zappert & Cruz, 1977).

This state of confusion and ambivalence is not limited to the evaluation of bilingual programs (Campeau, Roberts, Bowers, Austin, & Roberts, 1975). In examining previous attempts to evaluate the efficacy of special education programs for mildly handicapped children, Tindall (1985) found that "serious methodological flaws in these evaluation efforts make our present knowledge in this area very weak" (p. 101). Some of the problems identified include unclear definition of treatments and students served, use of weak experimental designs, inadequate testing instruments, and poor metrics in conjunction with inappropriate statistical tests. Gold (1981) reviewed several studies which examined evaluations of other federal education programs, such as compensatory education, migrant education, neglected and delinquent,

school desegregation, and follow-through. He, too, concluded that methodological flaws found in these program evaluations preclude any conclusive statements about program effects. Cook and Gruder (1978) reviewed four projects aimed at evaluating the technical quality of summative evaluations and concluded:

> The meta-evaluation studies . . . , while not definitive, do at least justify the suspicion that the technical quality of most evaluations leaves something to be desired and that this suspicion by itself warrants attempts to improve the quality of evaluation research efforts. (p. 15)

The fact that evaluation practices are almost universally inadequate does not excuse bilingual educators and program evaluators from responsibility for deficiencies in their program evaluations. Every effort should be made to improve their quality so that the impact of bilingual education can be accurately estimated and so that sound educational practices can be identified for language-minority students. The controversial nature of bilingual education is partly due to political squabbling and partly due to the inadequate practices in the evaluation of the implementation and effects of bilingual education.

In the following pages, this article will (a) empirically appraise the quality of practices in bilingual education evaluation, (b) provide a historical overview of the efforts devoted to the improvement in the quality of bilingual education evaluation, and (c) analyze the sources of methodological flaws in bilingual education evaluation. The purpose of this article is to promote quality bilingual education evaluation by reviewing its status and the causes of its inadequacy and by using this information to begin to identify potential solution strategies. With the improvement of bilingual education evaluation, the quality of education for language-minority students also is expected to improve.

Secondary Analysis of the Quality of Bilingual Education Evaluation Reports

One way to estimate the status and quality of bilingual program evaluations is to examine the eight studies which reviewed the literature on the effectiveness of bilingual education (Baker & de Kanter, 1981; Campeau et al., 1975; Douglas & Johnson, 1981; Dulay & Burt, 1978; Okada et al., 1982, 1983; Troike, 1978; Willig, 1985; Zappert & Cruz, 1977). Each of these reviews employed methodological screening criteria for selecting evaluation and research reports for further analysis and synthesis. The screening process and its results provide a basis for inferring the state of the art in bilingual program evaluation and for reaching some understanding of the difficulties and limitations associated with such undertakings.

In an attempt to identify and describe exemplary bilingual education programs, Campeau et al. (1975) examined 175 bilingual education programs, from which eight (5%) were selected for site visitation. Most of the 167 nonqualifying programs were rejected because the evaluation methodology in their program reports was so flawed that no conclusions could be drawn regarding the outcome of the program.

In reviewing 38 research projects and 175 project evaluations, Dulay and Burt (1978) found only nine (24%) research studies and three (2%) project evaluations that were free of one or more of the following critical research design weaknesses: (a) no control for subjects' socioeconomic status, (b) no control for initial language proficiency or dominance, (c) no baseline comparison data or control group, (d) inadequate sample size, (e) excessive attrition rate, (f) significant differences in teacher qualification for control and experimental groups, and (g) insufficient data

and/or statistics reported. The 12 documents that survived the screening provided the basis for Dulay and Burt's review.

To estimate the impact of Title VII programs, Zappert and Cruz (1977) reviewed approximately 600 official reports prior to 1978 and accepted 18 (3%) as methodologically sound and deserving of further examination. The following criteria were used for rejection: (a) no control for socioeconomic status; (b) inadequate sample size, improper techniques, or excessive attrition rate; (c) no baseline comparison data, no control group, nonrelevant comparison data; (d) no control for initial language dominance; (e) significant differences in teacher qualifications or characteristics, or other confounding variables; (f) insufficient statistical information or improper statistical applications; and (g) for research reports, lack of immediate relevance, new data, or accessibility.

The literature review conducted by Troike (1978) was drawn in part from the survey conducted by the Center for Applied Linguistics, which

> surveyed over 150 evaluation reports as part of its work in developing the master plan for the San Francisco schools to respond to the Lau vs. Nichols decision by the Supreme Court. . . . [In that survey], only seven evaluations [5%] were found which met minimal criteria for acceptability and contained usable information. (p. 3)

Troike selected 12 reports that attested to the effectiveness of bilingual education.

At the request of the White House Regulatory Analysis and Review Group for an assessment of the effectiveness of transitional bilingual education, Baker and de Kanter (1983) examined all evaluation studies reported since those reviewed by Zappert and Cruz (1977), as well as the 18 accepted by those reviewers. Of the 176 documents studied, 137 (78%) were rejected because they had one or more of the following deficiencies: (a) failure to address the issues of English and nonlanguage subject area outcomes; (b) nonrandom assignment with no effort to control for possible initial differences between control and program groups; (c) norm-referenced design; (d) comparison of posttest scores only, with nonrandom assignment; (e) reliance on school-year gains for the program group without a control group; and (f) reliance on grade-equivalent scores. Willig (1985), in undertaking a meta-analysis of the program evaluations reviewed by Baker and de Kanter, rejected an additional five on the grounds that they were either (a) evaluations of Canadian-type projects and thus nonrelevant (three studies), (b) a secondary-source evaluation summary (one study), or (c) outliers in terms of both instructional treatment and estimated effect size (one study).

In a study designed to assess the replicability of exemplary bilingual education projects via Project Information Packages (PIPs), Douglas and Johnson (1981) used seven guidelines to rate the technical quality of 19 PIP project evaluations. The guidelines were: (a) existence of an appropriate comparison standard for establishing a no-treatment expectation, (b) use of technically adequate tests, (c) adequate description of student characteristics, (d) analysis of the match between the content of tests and curriculum, (e) proper testing and scoring procedures, (f) appropriate data analysis, and (g) reasonable interpretation of results. Out of the 19 evaluations, only one (5%) was judged to be adequate, providing acceptable evidence for the effectiveness of the PIP-based project. Despite the fact that evaluation guidelines had been provided to the projects well in advance, the PIP project evaluations were generally very low in quality.

In a more extensive attempt to synthesize evaluation and research evidence on the effectiveness of bilingual education projects funded by the Elementary and Secondary Education Act (ESEA) Title VII, the National Center for Bilingual Research (NCBR) first reviewed evaluation and research reports prior to 1979 (Okada et al., 1982) and then reviewed those submitted during the 1980–1981 academic year (Okada et al., 1983). Of the 1,411 studies conducted between 1967 and 1979, 168 (12%) were accepted for use in the synthesis. For the 1980–1981 year, 355 studies were reviewed, and 84 (24%) were accepted and included in the meta-analysis, but only 60 (17%) were consistently coded by two independent analysts. An elaborate set of primary and secondary exclusion criteria was applied in the screening process. The following is a list of these criteria reorganized and simplified by O'Malley (1984, p. 2):

- General Design Problems
  - -no outcome data
  - -posttest only, no comparison
  - -testing not related to program objectives
  - -duration of treatment less than 6 months
  - -no information on duration of treatment
  - -only pretest data
- Testing Problems
  - -nonstandardized tests only with no comparison group
  - -no core achievement data (basic skills)
  - -different pretest/posttest test levels
  - -pretest/posttest samples different by more than 50%
- Student Information
  - -LEP students not identified in the analysis
  - -no information on number of students
  - -data not by language group
  - -students not identified by grade level
- Metric Used
  - -only reported percent above a test criterion
  - -raw score data only
  - -grade-equivalent scores
- Other
  - -inadequate program description
  - -transient populations (attrition too high)

It should be noted that not all reports included in these studies were Title VII evaluations, although the majority of them were. For example, 75% of the reports reviewed prior to 1979 were official reports submitted by Title VII projects.

Table 1 summarizes the acceptance rates of the eight review studies described above. As can be seen, the mean acceptance rate was only 10% (median = 6%). The acceptance rate of each study was undoubtedly affected by the selection criteria employed and the investigator's subjective judgments when applying them. Nevertheless, the low percentage of studies identified as methodologically acceptable reflects poor quality in conducting and reporting evaluations in bilingual education in the past. The reasons for rejection suggest that the practices usually employed in conducting bilingual education evaluations are inadequate. Some of these deficiencies can be corrected easily (e.g., insufficient program information), but some cannot (e.g., lack of control group and adequate testing instruments).

One other meta-analysis study was described in a doctoral dissertation by Gold

TABLE 1
Number and percents of studies accepted for review on effectiveness of bilingual education

Study	Number reviewed	Number accepted	Acceptance rate
Campeau et al. (1975)	175	8	5%
Zappert & Cruz (1977)	600	18	3%
Dulay & Burt (1978)	213	12	7%
Troike (1978)	150	7	5%
Douglas & Johnson (1981)	19	1	5%
Okada et al. (1982)	1,411	168	12%
Okada et al. (1983)	355	84	24%
Baker & de Kanter (1983)	176*	39	22%
			Mean = 10%
			Median = 6%
			Standard deviation = 8%

*Source: K. A. Baker (personal communication, November 13, 1985)

(1981). Instead of reviewing evaluation or research reports, the author reviewed 75 proposals drawn from a sample of 25 Title VII projects funded in California from 1975 through 1978. Using 33 criteria to rate the quality and appropriateness of the evaluation designs of these proposals, Gold found "none of the criteria were fully met by the proposals studied . . . [and] evaluation designs for Title VII programs showed a consistent lack of conventional evaluation rigor" (p. vii).

Although the above-mentioned reviews were conducted prior to the early 1980s, it is my estimation (based on extensive experience in providing technical consultations on bilingual education evaluation from 1985 to 1990) that the quality of bilingual education evaluation has apparently not changed in any measurable degree over the last 2 decades. The Office of Budget, Planning, and Evaluation (OBPE) of the Department of Education recently funded a study that reviews the evaluation practices of Title VII projects. This study found that only 45% of legislatively mandated evaluation components were typically present in Title VII evaluations and the quality of evaluation reports had not generally improved from 1986 to 1990 (Hopstock & Young, 1990). Considering the amount of time and money that has been spent on bilingual program evaluations, past practices in impact assessments of bilingual education, as appraised in the above-mentioned reviews, are somewhat discouraging. In the following section, past and present efforts to improve the quality of bilingual education are reviewed.

Efforts to Improve the Quality of Bilingual Education Evaluation

Several federal initiatives aimed at improving the quality of bilingual education evaluation have been drawn up in the past but have apparently met with little success (O'Malley, 1984). It seems appropriate to review these efforts briefly in an attempt to establish the direction of future efforts.

The U.S. Department of Education has long been concerned with providing technical assistance in evaluation to Title VII projects. This concern is evidenced by the support centers maintained by the Office of Bilingual Education and Minority Language Affairs (OBEMLA) nationwide. The Evaluation, Dissemination, and

Assessment Centers (EDACs) were funded by OBEMLA to provide support services to bilingual education programs and to bilingual teacher education training programs in the assessment, evaluation, and dissemination of relevant materials. Although the centers' primary focus was on the production and distribution of materials (Rodriguez, Sherman, Pelavin, & Hayward, 1984), numerous workshops on evaluation were offered, and voluminous evaluation materials were published by these centers. The Bilingual Education Multifunctional Support Centers (BEMSCs) were also responsible for providing technical assistance in evaluation to federally funded local bilingual education projects. Since 1985, the Evaluation Assistance Centers (EACs) have held all responsibility for the evaluation assistance function. In addition to the supportive services provided by these centers, OBEMLA has periodically sponsored management training institutes designed to familiarize Title VII project directors with current rules, regulations, and evaluation methodologies. A few projects aimed at advancing the state of the art in bilingual education evaluation have also been funded by the federal government.

The Bilingual Evaluation Technical Assistance (BETA) project was awarded to UCLA's Center for the Study of Evaluation (1980) by the National Institute of Education (NIE). The purpose of this project was to develop a series of modular workshops to train practitioners and community members in the evaluation of bilingual programs. A series of five texts designed to accompany workshop instruction was developed and field tested. The dissemination of the BETA's modules was limited by the lack of funding. Compared to others of its kind, the project was comprehensive in providing hands-on information about conducting evaluations in bilingual education. Another federal effort to develop evaluation and data gathering models for bilingual projects was carried out by InterAmerica Research Associates, which described the recommended practices in *A Handbook for Evaluating ESEA Title VII Bilingual Education Programs* (Perez & Horst, 1982). The handbook provides numerous forms and instructions for describing and documenting program operations and for identifying areas for program improvement. It also describes procedures for analyzing outcome data to determine student performance levels.

OBEMLA also attempted to improve bilingual evaluation practices by developing validation procedures for demonstration projects (programs of educational excellence). In one project (National Clearinghouse for Bilingual Education [NCBE], 1983), a panel of bilingual evaluators was formed to devise more relevant alternative procedures for validating bilingual education project success than those adopted by the Department of Education's Joint Dissemination Review Panel (Tallmadge, 1977). The task force presented a list of criteria for determining the effectiveness of demonstration projects and suggested potential solutions for problems commonly encountered in bilingual education evaluations. As a follow-up to this effort, OBEMLA contracted for the design of a comprehensive system to identify and validate effective bilingual programs and to disseminate information about these programs. The study's funding period was from January, 1984, to June, 1985.

A concern closely related to the evaluation of bilingual programs is the development of an effective student placement system. Two federally funded projects have been undertaken. The first project, conducted by the Southwest Regional Laboratory for Educational Research (SWRL) under contract to the U.S. Department of Health, Education, and Welfare (HEW), was completed in 1980. It produced a comprehensive set of resources for developing a student placement system for

bilingual programs. The size of the documents (several thousand pages), unfortunately, is intimidating to both practitioners and evaluators, who usually want quick answers to their questions. This deficiency may account for the fact that the materials are no longer available for dissemination.

The second project, entitled Selection Procedures for Identifying Students in Need of Special Language Services, was conducted by Pelavin Associates in the mid-1980s, under contract with the Department of Education's Office of Planning, Budget, and Evaluation. The purpose of the project was to identify procedures and criteria for placing LEP students in and exiting them from bilingual and other special programs. The study had very limited dissemination.

In addition to a number of articles written about bilingual education evaluation in general (e.g., De George, 1981; De Mauro, 1983; Garcia, 1980; Gezi, 1981; Gold, 1979; Hubert, 1982; Lam, 1986, 1989; Law, 1977; Martinez & Housden, 1975; Oller, 1978; Spolsky, 1978; Tucker & Cziko, 1978), several guides aimed at improving evaluation practices in bilingual education have also been published. The following is a selected list of these publications:

- Some of the papers in the Bilingual Education Paper Series—such as, the two papers by Burry (1981, 1982) on assessment and evaluation design, and on program documentation—produced by the Evaluation, Dissemination, and Assessment Center, that were written in response to local concerns;
- The Bilingual Education Teacher Training materials developed by the Center for the Development of Bilingual Curriculum in Dallas (Spencer, 1982);
- *Guidelines for Preparing the Annual Progress Report for Title VII Projects in Bilingual Education* (Evaluation, Dissemination, & Assessment Center, 1983b); and
- *Guide to Bilingual Program Evaluation* (Ulibarri, 1983).

The SWRL Education Research and Development Center published two evaluation guidebooks:

- *Program Impact Evaluations: An Introduction for Managers of Title VII Projects* (Bissell, 1979); and
- *Guidelines for the Evaluation of Bilingual Education Programs* (Cardoza, 1983).

The *Program Impact Evaluations* booklet, because of its nontechnical presentation, has been well received and widely distributed.

The Midwest BEMSC developed a training module for bilingual education evaluation designs (Secada, 1983), but only in outline form. The dissemination of this training module is very limited, however. The BUENO Center BEMSC "initiated a study of evaluation models and processes . . . in an effort to facilitate standardization of evaluation practices for Title VII projects" (Georgetown BESC and BUENO BEMSC Services, 1985). However, it is not clear which products were generated from this study.

Most of the aforementioned guidebooks contain a component that deals with language assessment, a key element in bilingual education evaluations. Many articles have also been written about this issue, and a number of booklets have been written evaluating the various language tests available in the field (e.g., Locks, Pletcher, & Reynolds, 1973; Northwest Regional Educational Laboratory, 1978). Articles specifically written about strategies for selecting tests for bilingual

programs have also been published (e.g., De George, 1983; Impink-Hernandez, 1984; Walker & Cabello, 1980).

From 1985 to 1987, under the auspices of the Office of Budget, Planning, and Evaluation of the Department of Education, a system for the evaluation of bilingual programs referred to as the Bilingual Education Evaluation System (BEES) and described in a user's guide of three volumes (Tallmadge, Lam, & Gamel, 1987a, 1987b; Lam & Gamel, 1987) was developed. The three volumes have had only limited distribution and virtually no marketing.

The *User's Guide* (Tallmadge et al., 1987a, 1987b) was a concerted effort to consolidate and build on previous work on bilingual program evaluation methods. It proposes the application of the gap-reduction design to measure project participants' performance and describes a careful measure of program implementation for the purpose of explaining and determining program effects. Although the *User's Guide* represents a significant technological advance in bilingual program evaluation, much work is still needed to further develop BEES by searching for better solutions to methodological problems. Also, to make the BEES more prescriptive and practical and thus more useful for practitioners, supplementary materials are needed. The two EACs have the opportunity and responsibility of accomplishing that goal. In addition to supporting training and providing technical assistance on evaluation issues, the federal government also has the obligation to support the development and dissemination of sound evaluation materials.

It is clear from the preceding overview that substantial efforts have been made to improve the quality of evaluation in bilingual education. It is important to note, however, that a number of these efforts took place in the 1980s at a time prior to, or concurrent with, completion of the aforementioned eight reviews of bilingual education effectiveness. Therefore, as this article draws its conclusions regarding the quality of bilingual education evaluation from the findings of these eight review studies, it would be inappropriate to attribute poor evaluation practices to inadequate improvement efforts in the 1980s. However, it is my contention, based on my professional experience and the findings from a recent study by Hopstock and Young (1990), as previously mentioned, that bilingual education evaluation continues to suffer from deficiencies and flaws similar to those identified by the eight review studies. This observation appears to suggest a negligible impact of improvement efforts that have continued through the 1980s.

Sources of Methodological Problems in Bilingual Education Evaluation

The apparent ineffectiveness of the national efforts discussed above can be attributed to the nature and dissemination of the materials produced by these efforts. First, evaluation guides have been disseminated to project directors (O'Malley, 1984), who are expected to pass them on to their evaluators. This delivery system has failed to ensure the full and proper use of the materials. Second, the dissemination of the materials has been largely limited and unsupported by a technical assistance system. Third, the documents themselves have tended to be cumbersome, poorly presented, and redundant. Finally, and most compellingly, the contents are generally nonprescriptive. They elaborate on the necessity for certain evaluation practices, assuming that readers already have or will learn the skills needed to understand and implement the recommendations.

The lack of sound and practical evaluation guidelines and materials, which reduces the effectiveness of any available technical assistance, represents only one source of problems in bilingual education evaluation.

Based on the relevant literature (e g., Baca, 1983, 1984; Burry, 1979, 1981; Cohen & Laosa, 1976; Evaluation, Dissemination, & Assessment Center, 1983a; Gezi, 1981; NCBE, 1983; Piper, 1984; Rodriguez-Brown, 1980; "Some Common Pitfalls," 1980; Yap, 1984), the strengths and weaknesses of bilingual education evaluation practices can be attributed to four other major sources: (a) the competence and knowledge of evaluators, (b) local administrative practices, (c) state and federal policy, and (d) characteristics of the bilingual education programs themselves. Each of these sources is discussed briefly below.

Evaluator Knowledge and Competence

Some of the deficiencies in bilingual evaluations are directly attributable to the lack of knowledge and inadequately developed skills of the individuals who conduct the evaluations. Shortcomings—such as, presentations including insufficient data and/or statistics, lack of control for initial language proficiency and socioeconomic status, use of inappropriate test scores, sample sizes not reported, no information on program description and implementation, and so on—which were observed in the eight review studies presented above, can be avoided if evaluators are properly trained in evaluation methodology. In a needs assessment survey conducted by the National Dissemination and Assessment Center in Los Angeles ("Bilingual Project Evaluators," 1978), eight (7%), out of the 123 bilingual project evaluators responding, specified evaluation and research as their area of concentration; two (2%) indicated specific preparation in bilingual education. Ninety-one percent of the evaluators surveyed were not trained in either evaluation or bilingual education.

Some of the problems inherent in bilingual education (e.g., high attrition rates) are beyond the control of the evaluator. It is also true that evaluators are usually restricted by insufficient funds and/or lack of administrative support. These points will be discussed later. Nevertheless, inappropriate analyses, inadequate reporting, and failure to point out threats to the validity of findings are probably attributable to a lack of evaluator competence. Bilingual education evaluations are plagued with so many other formidable impediments that "these specific difficulties in program evaluations should be resolved so the attention can be directed to some of the more difficult challenges in evaluations of instructional programs for LEP students" (O'Malley, 1984, p. 2).

What are the important skills and knowledge an evaluator should have? This question has been addressed in the evaluator-training literature. Anderson and Ball (1978) developed a list of 32 evaluator competencies and submitted it for review to a group of distinguished evaluation experts. The review panel added 34 more competency areas, although some of them overlapped with the initial list. Another list of evaluator competencies was produced, in several iterations, by a task force of the American Educational Research Association (Glass & Worthen, 1970; Millman, 1975; Worthen, 1975). In the last formulation (Worthen, 1975), the list consisted of 25 tasks requiring some 82 skills and/or areas of knowledge. Worthen described the list as incomplete. A list of six global evaluator competencies was offered by Ricks (1976). Another article, specifically written for bilingual educators ("Towards Selecting," 1980), discusses the roles of formative and summative evaluators, the pros and

cons of employing internal as opposed to external evaluators in conducting formative and summative evaluations, and the use of independent auditors or consultants to add credibility to an evaluation. Using internal evaluators, Cook and Shadish (1986) perceived the failure of mandated self-evaluation as attributable to the following three causes:

> First, project managers rarely want systematic information based on social science methods and instead prefer ammunition to help with their project's public relations. Second, in-house evaluators tend to have little power and multiple responsibilities, and are named the "evaluator" only because someone has to have this title and they know something about methodology. Finally, in-house evaluators are sometimes seen as allies of project management. (p. 201)

The competencies described in the various articles listed above clearly suggest that evaluations should be conducted by persons well trained in methodologies, skilled in interpersonal relations, knowledgeable in the areas in which the evaluation is to be conducted (e.g., bilingual education), and familiar with the projects they are evaluating. Needless to say, finding all of these attributes in a single individual may not be possible. Thus, it is often necessary to employ an evaluation team that coordinates the knowledge, skills, and experience of all its members. To select evaluation team participants, Bissell (1979) offers two guiding principles:

> Principle No. 1: The evaluators should have enough independence to be objective, but should be thoroughly familiar with all aspects of the project. They should be perceived as members of the project team, fully accessible to the rest of the project staff.
>
> Principle No. 2: Effective evaluation requires a variety of skills. The evaluator or evaluation team should include individuals with the collective range of expertise necessary to evaluate all project objectives, to accurately document the complexities of the project's school and community context, and to consider the sociolinguistic patterns and characteristics of the student participants. (p. 3)

Related to these principles is Bissell's suggestion that an evaluation monitoring team should include administrators, teaching staff, secondary level students, school board members, and district testing and evaluation staff. The responsibilities of such a team will consist of reviewing, commenting on, and facilitating all evaluation activities performed by the evaluator or evaluation team. Granted, the formation of such an evaluation monitoring team is probably feasible only in large school districts with large projects.

Even where evaluation monitoring teams are impractical, the quality of evaluations can be improved if key personnel—such as, principals, project directors, resource persons, and teachers—can sensitize the evaluator to the setting in which the program operates. This contextualization of summative evaluation is a much-needed improvement in bilingual education evaluations ("Towards Selecting," 1980). Cohen (1980) suggests several ways in which teachers and project directors can assist evaluators to ensure accurate assessment of their programs. In order for the evaluators to capitalize on these working relationships, it is just as important for them to know about the project as it is for the directors to know about evaluation. Only then can the two sides communicate effectively and complement each other's expertise in the production of adequate project evaluations. Thus, it may be necessary to

enhance not only the general level of evaluators' competence but also the project directors' knowledge of evaluation.

Although the competencies of evaluators who conduct national or large-scale evaluation studies usually exceed those of local evaluators, they may still be deficient. Cook and Gruder (1978) emphasize one of the reasons for low quality evaluation:

> Most evaluation research is conducted by profit-making, or not-for-profit, contract research agencies . . . [and] according to Bernstein and Freeman (1975), contract research agencies are rewarded for writing and winning contracts, and not for doing work that is at the level of the state of the art. Also, few mechanisms exist for punishing firms when the quality of their work falls below that of the state of the art. (p. 479)

Administrative Practices

Although evaluators are apparently guilty of misdirecting the evaluation process and thus of ultimately producing inadequate reports, administrators who supervise the evaluators must share the blame. Local administrators and project directors often do not appreciate the importance of evaluation and hence consider such appraisals to be meaningless bureaucratic exercises required by their funding agencies. Their cooperation in adjusting school routines to accommodate evaluation activities is therefore negligible. In addition, project directors often treat evaluation reports as public relations documents (Rodriquez-Brown, 1980) that have no real bearing on the program operations, or they misuse program evaluation in other ways (see Suchman, 1972, for a list of misuses of program evaluation). Consequently, project directors often are not motivated to formulate the clear program goals and objectives (Horst et al., 1980) necessary for adequate evaluation of the project.

It is also often the case that evaluators are pressured by their contractors or employers to repress negative findings and/or to avoid measures or analysis procedures that might produce them (Berman & McLaughlin, 1974). The time and financial constraints under which evaluations are conducted contribute further to the problem. Evaluators commonly are hired only after the project is underway and sometimes toward the end of the project year. This practice invariably—and understandably—seriously undermines any attempts to evaluate processes. To compound the problem, the money allocated for evaluation is rarely sufficient for even the most competent evaluator to do an adequate job. For example, classroom observation is crucial in documenting program implementation, but it is almost always beyond the evaluation budget. Usually, 3% of a project's total budget is allowed for evaluation (N.C. Gold, personal communication, November 5, 1985). For small projects in particular, this funding level is clearly deficient. Budget for program evaluation is often affected by policy, which is another determining factor of evaluation practices. The policy impact is discussed next.

State and Federal Policy

State and federal policies impact on the quality of evaluations in much the same manner that local administrative practices do. According to Horst et al. (1980), "most bilingual program evaluation designs are affected by local policies and conditions and by legal and funding agency regulations. In combination, these constraints may completely preclude any accurate assessments of program impact" (p. 60). In other words, evaluation practices encouraged or required by federal policy often

impose constraints that do not take into account the real conditions and expectations of bilingual programs. The ineffectiveness of federal policy is reflected in the quality of evaluation plans in approved Title VII applications (Gold, 1981) and in the quality of Title VII reports (Hopstock & Young, 1990).

Although federal regulations for bilingual education evaluations exist, they provide no specific instructions with respect to the ways in which data should be collected, analyzed, and presented. As discussed earlier, even the few guidebooks that have been developed for bilingual program evaluations (e.g., Bissell, 1979; Center for the Study of Evaluation, 1980; De George, 1980; Perez & Horst, 1982) are generally nonprescriptive regarding procedures for assessing program effects. This lack of technically sound and practical standards for conducting evaluations is undoubtedly a contributing factor to poor practices (Yap, 1984).

In addition to conforming to specific guidelines, bilingual education project proposals should be routinely reviewed by competent specialists. Active, sustained monitoring of the quality of evaluation and research in bilingual education can assure improvement. To avoid stakeholder bias, it has been recommended that evaluations not be monitored by the same office which funds the program (Cook & Gruder, 1978; L.M. Laosa, personal communication, October 25, 1985). Although some have argued that the real goal of bilingual programs is to provide bilingual education per se, and not to provide data for research on their effectiveness (Cooper, 1978), the improvement of services clearly depends on being able to identify those practices that facilitate the achievement of program objectives by different target groups. It seems reasonable to urge local educators and administrators to use the majority of the evaluation budgets for formative purposes—that is, to document and guide full implementation of the program design, including the analysis of problems arising when the school's capacity to actually implement the proposed program is being developed. Major evaluation research studies, of course, can concentrate on summative and impact analysis.

The recently instituted federal policy allowing a 12-month start-up period for new Title VII projects is a prime example of how policy may be changed to significantly improve the quality of bilingual program evaluations. Policy options for improving local evaluations were proposed by O'Malley (1984). They include: "(1) coordination among Federal, State, and local efforts, (2) developing a standardized reporting system, (3) strengthening LEA [Local Education Agency] use of evaluations, and (4) using LEA evaluation data at the aggregate level" (p. 6). Other examples of potential impact of policy on the quality of bilingual education evaluations are state and federal projects, such as the national improvement efforts discussed previously, which were designed to seek improved practices in program evaluation.

Characteristics of Bilingual Education Programs

The three factors discussed above (evaluator competence, administrative practices, and state and federal policy constraints) are modifiable through policy changes and training. A fourth factor that affects the quality of evaluation practices, the inherent characteristics of bilingual education programs, cannot be altered. These characteristics, discussed below, significantly restrict what can be done in program evaluation.

Differing cultural backgrounds of LEP students. The most salient and obvious feature of a bilingual program is that all LEP students served are limited in English

proficiency and that their native languages and cultural backgrounds are different from those of the mainstream population. This feature means that available affective and cognitive achievement instruments are usually not well suited for use with LEP populations. Very often, pretreatment achievement data cannot be obtained because students do not know enough English to take a test. When they are tested, their scores are likely to be quite unreliable (Baker & Pelavin, 1984; Lam, 1987). The resulting lack of sound baseline data makes it impossible to generate credible treatment–effect estimates. An additional complication is that it may not be possible to test children in their native languages. Suitable instruments may not exist, and LEP students' native language literacy skills may be inadequate for existing tests (Lam, 1991).

Because one major goal of bilingual education is to develop LEP students' proficiency in English, measurement of this skill is crucial both for placement and for outcome assessment purposes. Currently, the most popular oral language proficiency tests are the Bilingual Syntax Measures (BSM), the Language Assessment Battery (LAB), the IDEA Proficiency Test (IPT), and the Language Assessment Scales (LAS). Unfortunately, "all of these [instruments], according to the office of Bilingual Education of the California Department of Education, suffer serious psychometric defects" (Piper, 1984). The major criticism is that skills measured by these language tests do not adequately represent the English language proficiency construct (Willig, 1985). One related problem is that, while federal regulations use the term *English proficiency* to include all language skills, most English proficiency tests measure only oral language skills.

Remedial LEP services required by law. The legal requirement that all LEP students must receive special instructional services effectively eliminates any possibility of employing a true experimental design with random assignment of students to treatment and control groups or even of employing a nonequivalent comparison group design. Baker and Pelavin (1984) have suggested that one way a control group might be obtained is by delaying service to some students while serving others. Such a delaying strategy may be attractive from a research perspective, but it would be certain to draw strong protests from the bilingual education community—especially if the delay were long enough to guarantee that treatment effects could be reliably measured. With a delay of at least a year, even the more intensive special help described by Baker and de Kanter would be perceived as inadequate to compensate students for the loss of time.

A variation on the random assignment theme involves conducting true experiments with less needy LEP students (Balasubramonian, 1979). However, it seems inappropriate and hazardous to generalize the results from studies of less needy children to the population of those LEP students with more severely limited language skills who are the main targets of bilingual education. Without a control group composed of such students, it is virtually impossible to establish a valid no-treatment expectation, without which measurement of program effects is very difficult.

Because random assignment is apparently not feasible, an alternative strategy is to seek out a preexisting intact group of LEP students not participating in a bilingual program to use as a standard of comparison. Unfortunately, the legal requirement to serve all LEP children makes the existence of suitable comparison groups extremely unlikely. On the other hand, it may be feasible to find a comparison group which is receiving bilingual services that vary from those of the experimental group and to

compare the relative effectiveness of the different treatments. Even this possibility is remote, however, because the meaningfulness of the comparison would hinge on the two groups' being virtually identical in all attributes except the treatment. And "if the two groups are not matched on key variables, it will not only invalidate the results [but] will also produce very misleading information that can do great harm" (McConnell, 1983, p. 4).

Another way of deriving some estimate of treatment effects is to utilize a historical record approach in which achievement measures collected prior to students' entry into the program can be contrasted to posttreatment measures obtained on children at the same age or grade level (see, e.g., McConnell, 1982). Unfortunately, the number of situations in which it is possible to compile the data needed for this type of assessment is probably limited. Still other quasi-experimental designs are at least theoretically possible, although the unique characteristics of bilingual education programs typically cause nontrivial implementation problems.

An approach to assessing program effects without empirical no-treatment expectations is the theoretical logical deduction analysis (Lam, 1990), which can be implemented in three ways. First, if the program design is well grounded in sound theory and research findings, data showing success in program implementation can be used as evidence for effective programming. Second, the level of correspondence between the expected pattern of outcomes based on program theory and design and the pattern actually obtained is used to suggest the extent to which observed outcomes can be attributed to program operations. Finally, no-treatment expectations can be posited based on theory, well-developed logical arguments, research findings, needs assessment data, or a combination of these. For example, Peleg (1978) argues that students with no, or limited, English proficiency will not progress in their academic achievement without special bilingual instruction; therefore, the no-treatment expectations for these students are zero or negative growth. The theoretical logical deduction analysis approach to program effects assessment, although intuitively attractive, has nevertheless yet to be systematically studied, tested, and used.

Mobility characteristic of bilingual students. Many LEP students are either recent immigrants whose families are still in transition or migrant students who relocate seasonally. The resulting high rates of transiency, attrition, and accretion in bilingual programs result in data sets characterized by large amounts of missing data, widely varying exposure to treatment, and diverse student-by-treatment interactions. All of these problems combined make it hard for evaluators to assess program effects and to generalize evaluation findings.

Slow emergence of bilingual program effects. Ovando and Collier (1985) reviewed several studies, including Cummins' (1980) article, and concluded that the cumulative effects of bilingual programs on increasing achievement and IQ scores are not apparent until the fourth, fifth, or sixth years of bilingual instruction. One proposed strategy for evaluating program effectiveness is to determine how successfully reclassified LEP students function in mainstream classrooms or in society "in terms of personality disorder" (Paulston, 1977, p. 100). The mobility problem reduces the size of the usable data base and makes follow-up and longitudinal research or evaluation nearly impossible. Piper (1984) reported that only 10% of the bilingual students in his evaluation sample had complete data over a 3-year period. If sample sizes are small to begin with, meaningful data analyses are not possible in longitudinal evaluations.

The loss of data due to transiency also casts some doubts on the representativeness of the sample. If the scores of those who exit the program early, enter the program

late, or enter and exit the program repeatedly differ systematically from those who remain in the program, the results can be generalized only to the population of nonmobile LEP students. Two other potential sources of bias are absenteeism at the time tests are given (Piper, 1984) and retention of students in the program after they should have been exited. Students for whom test data are available may differ systematically from the true target group. If this were indeed the case, it would be inappropriate to generalize from students with complete sets of test scores to the target population.

Highly variable LEP student characteristics. LEP students may differ from the mainstream student population in ethnicity, country of origin, language, length of residence in the United States, language proficiency, prior school experience, and socioeconomic status (NCBE, 1983). These characteristics also vary within the LEP student population to such an extent that students clearly have different needs. For example, a refugee from Vietnam who has missed 3 years of schooling will require a very different instructional strategy from that of a recent Mexican immigrant who has missed no schooling. Since various background characteristics can influence how rapidly the students will learn English and achieve in school, it is very important to document, control, and/or otherwise account for these characteristics in order to enhance the interpretability of evaluation findings. This point will be discussed in greater detail later.

Nonstandardized bilingual program component applications. Bilingual education treatments traditionally include instruction, curriculum development, staff development, and parent and community involvement components. The implementation of these components varies from project to project, depending on local needs and feasibility. For the instructional component, which is common to all projects, the degree of implementation may vary not only among but also within projects. "Indeed, variations occur between schools within the same project, between classrooms within the same school, and between students within the same classroom" (Piper, 1984).

There are many reasons why the implementation of treatments is not uniform. First, as mentioned before, different students have different needs. Second, implementing bilingual projects in school districts is very difficult. "A large degree of organizational change and mutual adaptation is required to successfully implement a bilingual education project. Local capacity building and strong commitment supported by a well-planned in-service program are also needed" (Yap, 1984, pp. 1–2). As discussed earlier, local administrators' attitudes toward bilingual education, especially those of the school principals and the mainstream teachers, play important roles in determining the level of staff cooperation in adopting the bilingual program in their schools. Thus, the degree of program implementation varies, depending on how often a project encounters these obstacles and how successfully it overcomes them. Because of the complex difficulties that bilingual educators face in implementing the programs, it becomes imperative that the degree of program implementation be assessed (Bissell, 1979; Burry, 1982).

Other factors that affect levels of program implementation are the qualifications of the bilingual teaching staff and the availability of teaching and learning materials for LEP students. With regard to staff, at the time of the initial bilingual program implementation, there appears to have been a shortage of bilingual teachers having the qualifications specified in the 1980 Title VII Rules and Regulations (Brown, 1979; Ortiz, 1979). Without well-prepared bilingual teachers and aides, of course, bilingual instruction cannot be provided as planned. Teaching practices are also affected by the availability of instructional materials. Unfortunately, very few native language materials (except those in Spanish) are available on the market. Because of the difficulties created by these two factors in achieving the instructional goals, the implementation of instructional components has been uneven across different sites.

The final reason for treatment variation is the recent origin of bilingual education programs. Very often a program changes and evolves over time as it adapts to local conditions and to improvements in design. Program designs may be modified due to practical constraints or research input. Instructional strategies and materials are tried, abandoned, adopted, or adapted to meet the demands and the needs of the students and the school. As long as the program is in a state of flux, impact evaluation is difficult, if not impossible (Horst, et al., 1980), and the need is correspondingly greater for implementation evaluation.

Small number of students served by each bilingual program. Where schools are small, treatments may be implemented in only one or two classrooms per grade level. The typical Title VII project nationwide serves some 200 to 400 LEP students in three to four schools across several grades (N. C. Gold, personal communication, November 5, 1985). This situation results in small sample sizes (possibly preventing detection of small treatment effects). In addition, it means that unusually effective (or ineffective) teachers or schools with outstanding (or inept) leadership can have a marked influence on the results of the evaluation (Horst, 1982). To combat small sample size, one solution is to aggregate data across projects, but care must be taken that any such aggregations deal appropriately with any differences in children served, settings, and treatment characteristics.

Based on the preceding review of the difficulties inherent in evaluating bilingual programs, the following conditions are needed to correct current deficiencies in local evaluation and to increase validity of evaluation findings:

1. Technical skills in planning, collecting, processing, and analyzing data;
2. Measurement and/or documentation of program implementation and student or setting characteristics that may interact with the program;
3. Processes for selecting and/or developing reliable and valid assessment instruments and procedures;
4. Evaluation designs with adequate internal validity that do not require a randomized control group;
5. A system of comparing and aggregating data across projects; and
6. A system for utilizing setting, student, and process information in outcome evaluations.

Summary

The preceding discussion is summarized in Figure 1, which represents the presumed causal relationships between the various influencing factors and the technical quality of evaluation practices. The arrows show the direction of causal influences.

The extent of evaluation guidelines and technical assistance is affected by state and federal policies, the knowledge and competency of evaluators, and the quality of evaluation practices. These guidelines and technical assistance should in turn have some impact on evaluators' knowledge and competency and the quality of evaluation.

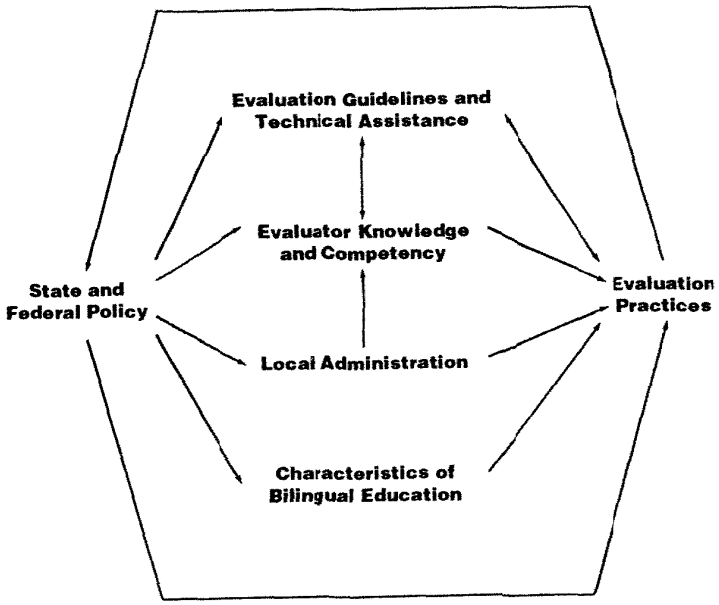

FIGURE 1. *Causal relationships among various factors and the technical quality of evaluation practices*

Local administrators and project directors can affect evaluation practices by the level of excellence they demand of the evaluation and by the extent and quality of their cooperation with the evaluator.

The unique characteristics of bilingual programs—affected by the variety of student groups served, by policy and guideline constraints, and by the political space within which the program operates—necessitate careful consideration of extraneous influences and tailoring of evaluation designs. State and federal policies restrict local administrative operations through funds allocated for evaluation and through deadlines and regulations. Such policies may also contribute to the hiring of incompetent evaluators because such policies set no standards. In addition, state and federal regulations can have a direct effect on evaluation practices by, for example, failing to provide adequate proposal reviews. On the other hand, state and federal policies are shaped (or should be) by the extent to which these policies are indeed workable or practical at the local level. Also, actual evaluation practices can affect federal policy, as evidenced by federal initiatives to improve bilingual program evaluations.

The preceding review highlights the fact that there are serious methodological flaws in bilingual education evaluation and research reports. It should be noted at the outset that, while some of these deficiencies and problems are relatively simple to resolve (e.g., through greater methodological rigor), others are not. As discussed above, except under certain conditions, deriving a valid estimate of how participants would have performed without the program appears to require groups of LEP students, with needs similar to those of program participants, who do not participate

in bilingual projects—a situation that is expressly prohibited by the legislative requirement that the neediest students be served. Furthermore, some of the difficulties encountered in local evaluations are not the same as those encountered in national or large-scale impact evaluations. For example, insufficient resources are often the origins of problems found in the former but not the latter type of evaluation effort. In any event, the major obstacles that must be overcome in order to obtain valid impact assessments of bilingual education are the same for local, state, and national evaluations.

As can be inferred from the above discussion, the amelioration of difficulties in evaluating bilingual education programs that have led to inferior evaluations requires more than developing evaluation guides and providing technical assistance. A concerted and broad-based effort must be made to effect policy and to enhance the competence and commitment of evaluators and evaluation teams. For the promotion of equality of educational and economic opportunities for LEP children, evaluation of bilingual education must be taken seriously for the instrumental purposes that it is designed to serve—program accountability and improvement. The refinement of current practices and the reduction of problems in the evaluation of bilingual education must be made a top priority if the LEP student population is to be adequately served.

References

Alkin, M., Kosecoff, J., Fitz-Gibbon, S. C., & Seligman, R. (1974). Evaluation and decision-making: The Title VII experience. *CSE Monograph Series in Evaluation, 4.*

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

Baca, R. (1983). *Notes on bilingual program evaluation.* Los Angeles: California State University, Multifunctional Support Service Center.

Baca, R. (1984). *Bilingual education evaluation: An overview.* Los Angeles: California State University, Multifunctional Support Service Center.

Baker, K. A., & de Kanter, A. A. (1981). *Effectiveness of bilingual education: A review of the literature.* Washington, DC: U.S. Department of Education.

Baker, K. A., & de Kanter, A. A. (1983). Federal policy and the effectiveness of bilingual education. In K. A. Baker & A. A. de Kanter (Eds.), *Bilingual education* (pp. 33–86). Lexington, MA: Lexington Press.

Baker, K. A., & Pelavin, S. (1984, April). *Problems in bilingual evaluation.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Balasubramonian, K. (1979). *Measurement, evaluation and accountability in bilingual education programs.* Rosslyn, VA: National Clearinghouse for Bilingual Education.

Berman, P., & McLaughlin, M. W. (1974). *Federal programs supporting educational change: Vol. 1. A model of educational change.* Santa Monica, CA: Rand.

Bernstein, I., & Freeman, H. E. (1975). *Academic and entrepreneurial research.* New York: Russell Sage Foundation.

Bilingual project evaluators: An overview. (1978). *Bilingual Resources, 2*(1), 32–34.

Bissell, J. S. (1979). *Program impact evaluations: An introduction for managers of Title VII projects.* Los Alamitos, CA: Southwest Regional Laboratory for Educational Research and Development.

Brown, H. C. (1979). Why the bilingual education shortage? *Bilingual Resources, 2*(3), 21–23.

Burry, J. (1979). Evaluation in bilingual education. *Evaluation Comment, 6,* 1–14.

Burry, J. (1981). An introduction to assessment and design in bilingual program evaluation. *Bilingual Education Paper Series, 5*(5).

Burry, J. (1982). Evaluation and documentation: Making them work together. *Bilingual Education Paper Series, 5*(9).

Campeau, P. L., Roberts, A. O. H., Bowers, J. E., Austin, M., & Roberts, S. J. (1975). *The identification and description of exemplary bilingual education programs*. Palo Alto, CA: American Institutes for Research.

Cardoza, D. (1983). *Guidelines for the evaluation of bilingual education programs*. Los Alamitos, CA: National Center for Bilingual Research.

Cohen, A. D. (1980). *Describing bilingual education classrooms*. Rosslyn, VA: National Clearinghouse for Bilingual Education.

Cohen, A. D., & Laosa, L. M. (1976). Second language instruction: Some research considerations. *Curriculum Studies, 8*(2), 149–162.

Cook, T. D., & Gruder, C. L. (1978). Meta-evaluations research. *Evaluation Quarterly, 2*, 5–51.

Cook, T. D., & Shadish, W. R., Jr. (1986). Program evaluation: The worldly science. *Annual Review of Psychology, 37*, 193–232.

Cooper, R. L. (1978). Research methodology in bilingual education. In J. E. Alatis (Ed.), *Georgetown University round table on languages and linguistics 1978* (pp. 66–74). Washington, DC: Georgetown University Press.

Cummins, J. (1980). The entry and exit fallacy in bilingual education. *Journal of the National Association for Bilingual Education, 4*(3), 25–60.

De George, G. P. (1981). Bilingual program evaluation: The need goes on. *Bilingual Journal, 5*(1), 15–16.

De George, G. P. (1983). Selecting tests for bilingual program evaluation. *Bilingual Journal, 7*(2), 22–28.

De George, G. P. (Ed.). (1980). *Improving bilingual program management: A handbook for Title VII directors*. Cambridge, MA: Laney College, Evaluation, Dissemination and Assessment Center.

De Mauro, G. E. (1983). Models and assumptions for bilingual education evaluation. *Bilingual Journal, 7*(2), 8–12.

Douglas, D., & Johnson, D. M. (198[illegible], April). *An evaluation of Title VII evaluations: Results from a national study*. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles.

Dulay, H., & Burt, M. (1978). *Why bilingual education? A summary of research findings* (2nd ed.). San Francisco: Bloomsbury West.

Epstein, N. (1977). *Language, ethnicity, and the schools. Policy alternatives for bilingual-bicultural education*. Washington, DC: George Washington University, Institute for Educational Leadership.

Evaluation, Dissemination, and Assessment Center. (1983a). *Guide to bilingual program evaluation*. Austin, TX: Author.

Evaluation, Dissemination, and Assessment Center. (1983b). *Guidelines for preparing the annual progress report for the Title VII projects in bilingual education*. Austin, TX: Author.

Garcia, G. N. (1980). Plans, designs, and reports: Evaluation is the common denominator. *FORUM, 3*(3), 4–6.

General Accounting Office. (1976). *Bilingual education: An unmet need (Report to the Comptroller General of the United States)*. Washington, DC: U.S. Government Printing Office.

Georgetown Bilingual Education Support Center and BUENO Bilingual Education Multicultural Support Center Services. (1985). *FORUM, 8*(5), 5.

Gezi, K. (1981). Effective evaluation in bilingual education. *Educational Research Quarterly, 6*(3), 104–111.

Glass, G. V., & Worthen, B. R. (1970). *Essential knowledge and skills for educational research and evaluation*. (Tech. Paper No. 5, Task Force on Research Training). Boulder, CO: American Educational Research Association.

Gold, N. C. (1979, May). *Improving evaluation of bilingual education projects*. Paper presented at the Annual Meeting of the National Association for Bilingual Education, Seattle.

Gold, N. C. (1981). *Meta-evaluation of selected bilingual education projects*. Unpublished doctoral dissertation, University of Massachusetts.

Hopstock, P., & Young, M. (1990). *A review of local Title VII project evaluation plan and evaluation reports* (U.S. Dept. of Education Contract No. LC 89023001). Arlington, VA: Development Associates.

Horst, D. P. (1982). *ESEA Title I evaluation and reporting system: Evaluation of the models at the project level*. Mountain View, CA: RMC Research Corp.

Horst, D. P., Johnson, D. M., Nava, H. G., Douglas, D. E., Friendly, L. D., & Roberts, A. O. H. (1980). *An evaluation of project information packages (PIPs) as used for the diffusion of bilingual projects: Vol. III. A prototype guide to measuring achievement level and program impact on achievement in bilingual projects* (Report No. UR-460). Mountain View, CA: RMC Research Corp.

Hubert, J. A. (1982, October). *Conceptualizing outcome evaluation in bilingual education programs*. Paper presented at the Annual Meeting of the Northeastern Educational Research Association, Ellenville, NY.

Impink-Hernandez, M. V. (1984). *Language proficiency assessment test selection* (Sept. 4, 1984, memorandum to Title VII project directors). Arlington, VA: National Clearinghouse for Bilingual Education.

Lam, T. C. M. (1986, April). *Local bilingual program evaluation: Its function and some directions for improvement*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Lam, T. C. M. (1987, November). *Testability: A critical issue in testing minority students with standardized achievement tests*. Paper presented at the Minority Assessment Third Annual Conference, Tucson.

Lam, T. C. M. (1989, May). *What is the gap-reduction design, really?* Paper presented at the Annual National Association of Bilingual Education Conference, Miami.

Lam, T. C. M. (1990, October). *A two-stage approach to assessing program effects*. Paper presented at the Annual Meeting of the American Evaluation Association, Washington, DC.

Lam, T. C. M. (1991). Testing of limited English proficient children. In K. Green (Ed.), *Educational testing: Issues and applications* (pp. 125–167). New York: Garland.

Lam, T. C. M., & Gamel, N. N. (1987). *Bilingual education evaluation and reporting system: Users' guide, abbreviated recommendations for meeting Title VII evaluation requirements*. Washington DC: Department of Education.

Law, A. I. (1977). *Evaluating bilingual programs* (TM Report No. 61). Princeton, NJ: ERIC Clearinghouse on Tests, Measurement, and Evaluation.

Locks, N. A., Pletcher, B. A., & Reynolds, D. F. (1978). *Language assessment instruments for limited-English-speaking students*. Washington, DC: U.S. Department of Health, Education and Welfare.

Martinez, J., & Housden, J. L. (1975, November). *Critical issues in the evaluation of bilingual education*. Paper presented at the Annual Meeting of the California Educational Research Association, San Diego.

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

McConnell, B. B. (1983). Needed: A new technology for evaluating bilingual education programs. *Bilingual Education Paper Series, 7*(1).

Millman, J. (1975). *Selecting educational researchers and evaluators* (TM Report No. 48). Princeton, NJ: ERIC Clearinghouse on Tests, Measurement, and Evaluation.

National Clearinghouse for Bilingual Education. (1983). *Information and technical assistance needs of Title VII demonstration projects: Validation and evaluation*. Washington, DC: Office of Bilingual Education and Minority Language Affairs.

Northwest Regional Educational Laboratory. (1978). *Assessment instruments in bilingual education: A descriptive catalogue of 342 oral and written tests*. Los Angeles: California State University, Evaluation, Dissemination, and Assessment Center.

Okada, M., Besel, R., Bachelor, P., Glass, G. V., & Montoya-Tannatt, L. (1983). *Synthesis of reported evaluation and research evidence on the effectiveness of bilingual education: Basic projects, final report: Tasks 7–8.* Los Alamitos, CA: National Center for Bilingual Research.

Okada, M., Besel, R., Glass, G. V., Montoya-Tannatt, L. & Bachelor, P. (1982). *Synthesis of reported evaluation and research evidence on the effectiveness of bilingual education: Basic projects, final report: Tasks 1–6.* Los Alamitos, CA: National Center for Bilingual Research.

Oller, J. W. (1978). The language factor in the evaluation of bilingual education. In J.E. Alatis (Ed.), *Georgetown University round table on languages and linguistics* (pp. 410–422). Washington, DC: Georgetown University Press.

O'Malley, J. M. (1984). *Options for improving local evaluations. Focus, 15*, 1–8.

Ortiz, F. I. (1979). The administration of bilingual education programs. *Bilingual Resources, 3*(1), 2–7.

Ovando, D. J., & Collier, V. P. (1985). *Bilingual and ESL classrooms*. New York: McGraw-Hill.

Paulston, C. B. (1977). Viewpoint: Research. *Bilingual education: Current perspectives* (Vol.2). Arlington, VA: Center for Applied Linguistics.

Peleg, Z. R. (1978). Impact assessment in the evaluation of bilingual programs: Is it feasible? *Educational Technology, 18*, 19–23.

Perez, R. S., & Horst, D. P. (1982). *A handbook for evaluating ESEA Title VII bilingual programs* [Draft]. Rosslyn, VA: InterAmerica Research Associates.

Piper, R. (1984). *Effective bilingual education evaluation: Is it possible?* Los Alamitos, CA: National Center for Bilingual Research.

Ricks, F. A. (1976). Training program evaluators. *Professional Psychology, 7*, 338–343.

Rodriguez, B. R., Sherman, J. D., Pelavin, S. H., & Hayward, B. J. (1984). *An evaluation of the Bilingual Education Evaluation, Dissemination, and Assessment Center.* Washington, DC: Pelavin.

Rodriquez-Brown, F. V. (1980). Do's and don'ts of bilingual program evaluation. *Bilingual Education Paper Series, 3*(6).

Secada, W. (1983). *Evaluation designs for bilingual education programs: Looking at program outcomes.* Arlington Heights, IL: Midwest Bilingual Education Multifunctional Support Center.

Some common pitfalls in the evaluation of bilingual education programs. (1980). *Bilingual Resources, 3*(3), 49–51.

Spencer, M. (1982). *Bilingual education teacher training packets, Series A: Bilingual program planning, implementation, and evaluation.* Austin, TX: Evaluation, Dissemination, and Assessment Center.

Spolsky, B. (1978). A model for the evaluation of bilingual education. *International Review of Education, 28*(3), 347–360.

Suchman, E. A. (1972). Action for what? A critique of educative research. In C. H. Weiss (Ed.), *Evaluating action programs* (pp. 52–84). Boston: Allyn & Bacon.

Tallmadge, G. K. (1977). *Ideabook: The Joint Dissemination Review Panel*. Washington, DC: U.S. Department of Health, Education and Welfare.

Tallmadge, G. K., Lam, T. C. M., & Gamel, N. N. (1987a). *Bilingual education evaluation and reporting system: User's guide, Vol. I—Recommended procedures.* Washington DC: Department of Education.

Tallmadge, G. K., Lam, T. C. M., & Gamel, N. N. (1987b). *Bilingual education evaluation and reporting system: User's guide, Vol. II—Technical appendices.* Washington DC: Department of Education.

Tindall, G. (1985). Investigating the effectiveness of special education: An analysis of methodology. *Journal of Learning Disabilities, 18*(2), 65–128.

Towards selecting a bilingual project evaluator. (1980). *Bilingual Resources, 3*(2), 42–47.

Troike, R. C. (1978). Research evidence for the effectiveness of bilingual education. *Journal of the National Association for Bilingual Education, 3*(1), 13–24.

Tucker, G R., & Cziko, G. A. (1978). The role of evaluation in bilingual education. In J. E. Alatis (Ed.), *Georgetown University round table on languages and linguistics* (pp. 423–446). Washington, DC: Georgetown University Press.

Ulibarri, D. M. (1983). Documenting classroom program implementation. In *Guide to bilingual program evaluation* (pp. 77–96). Dallas, TX: Evaluation, Dissemination and Assessment Center.

University of California, Center for the Study of Evaluation. (1980). *Program documentation, planning for students' assessment, planning for management, selecting evaluation designs.* Los Angeles: Author.

U.S. Department of Education. (1982). *The condition of bilingual education in the nation, 1982: A report from the Secretary of Education to the President and the Congress.* Washington, DC: Author.

Walker, C., & Cabello, B. (1980). Features to consider when selecting a test for a bilingual program. *Bilingual Resources, 4*(1), 25–27.

Willig, A. C. (1985). A meta-analysis of selected studies on the effectiveness of bilingual education. *Review of Educational Research, 55*, 269–317.

Worthen, B. R. (1975). Competencies for educational research and evaluation. *Educational Researcher, 4*(1), 13–16.

Yap, K. O. (1984, April). *Standards for Title VII evaluation: Accommodation for reality constraint.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Zappert, L. T., & Cruz, B. R. (1977). *Bilingual education: An appraisal of empirical research.* Berkeley: Bay Area Bilingual Education League/Lau Center, Berkeley Unified School District.

Author

TONY C. M. LAM is Associate Professor, Faculty of Education, University of Toronto, Ontario M5S 2R7, Canada. He specializes in program evaluation and applied measurement.