DOCUMENT RESUME

ED 321 557 FL 018 677

AUTHOR Tallmadge, G. Kasten; And Others
TITLE The Evaluation of Bilingual Education Programs for Language-Minority, Limited-English-Proficient Students: A Status Report with Recommendations for Future Development.
INSTITUTION RMC Research Corp., Mountain View, Calif.
SPONS AGENCY Department of Education, Washington, DC.
PUB DATE Sep 87
CONTRACT 300-85-0140
NOTE 232p.; For related documents, see ED 291 650 and FL 018 675.
PUB TYPE Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC10 Plus Postage.
DESCRIPTORS *Bilingual Education Programs; Elementary Secondary Education; English (Second Language); *Evaluation Methods; *Limited English Speaking; *Measurement Techniques; *Program Evaluation; Research Design; *Research Methodology; Statistical Analysis; Test Validity

ABSTRACT

A discussion of the evaluation of bilingual education programs focuses on building a comprehensive framework for local efforts at evaluation. The discussion begins with an introduction to the legislative history of bilingual education programs and the evolution of their evaluation. This is followed by a review of the literature on current practices and problems in program evaluation, looking at the kinds of inferences that can be drawn from program evaluations and the threats to the validity of those inferences. Strategies for reducing threats to validity are examined. A chapter is devoted to treatment, student, and setting variables that have been identified as potentially interactive on the basis of either theoretical formulations or empirical findings. Lists of the variables and methods for obtaining and documenting relevant information are presented. Four sources of systematic error associated with simple measurements of growth are discussed, and types of tests and other measures for assessing bilingual education program impact are examined. Eight evaluation designs reported in the literature are described, and problems associated with them are reviewed. Several approaches to the measurement of outcomes on a common scale are evaluated. A bibliography of over 400 items is included. (MSE)

RESEARCH CORPORATION

Phase I Report

The Evaluation of Bilingual Education Programs for Language-Minority, Limited-English-Proficient Students: A Status Report with Recommendations for Future Development

G. K. Tallmadge
T. C. M. Lam
N. N. Gamel

September 1987

RMC Research Corporation
2570 West El Camino Real, Mountain View, CA 94040

Prepared for the U.S. Department of Education
Washington, D.C.

THE EVALUATION OF BILINGUAL EDUCATION PROGRAMS FOR LANGUAGE-MINORITY, LIMITED-ENGLISH-PROFICIENT STUDENTS: A STATUS REPORT WITH RECOMMENDATIONS FOR FUTURE DEVELOPMENT

G. K. Tallmadge
Tony C. M. Lam
Nona N. Gamel

September 1987

Prepared for:

U.S. Department of Education

The research reported herein was performed pursuant to Contract No. 300-85-0140 with the U.S. Department of Education. Contractors undertaking such projects under government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official U.S. Department of Education position or policy.

TABLE OF CONTENTS

List of Figures and Tables

ACKNOWLEDGMENTS

The first draft of this report was reviewed by eight "external" readers:

William H. Angoff - Educational Testing Service
Karen E. Banks - Oregon Total Information System
Thomas D. Cook - Northwestern University
Norman C. Gold - California State Department of Education
Luis M. Laosa - Educational Testing Service
Robert L. Linn - University of Illinois
Beverly B. McConnell - Private Consultant
Ann C. Willig - University of Texas at Austin

Other reviewers included, from the U.S. Department of Education,

James J. English (Project Officer for the study)
Janice K. Anderson
Keith A. Baker
John Chapman
Edward J. Fuentes
Robert L. Kastner

and, from the Council of Chief State School Officer's Committee for Evaluation and Information Systems,

Gerald E. De Mauro

Their comments and suggestions were extremely helpful and are reflected in this, the third revision of the report.

There were exactly 100 significant, substantive comments and many more minor ones. We have attempted to respond to all of them, but few of our reviewers (none of the external reviewers) have had a chance to examine our efforts to incorporate their suggestions. Whatever deficiencies remain are therefore the sole responsibility of the authors.

Not all of our reviewers agreed with each other, and there continue to be several points on which the authors themselves are not in total accord. On all of these issues we tried to reach compromises that we felt were fair to all points of view. The senior author as the final arbitrator, however, is responsible for any failures to achieve a balanced presentation.

We would like to thank all of our reviewers for their thoughtful readings and insightful comments. Because of their valued input, this report is substantially more complete, more technically accurate, and more in tune with real-world bilingual education issues than our earlier draft.

GKT
TCML
NNG

SUMMARY AND RECOMMENDATIONS

What we call bilingual education in the United States is quite different from what the rest of the Western world calls bilingual education. Here the term connotes special programs which are designed for non- and limited-English-proficient, language-minority students and which have two primary objectives: (a) to develop these students' English language skills and (b) to prevent them from falling behind their fully English-proficient peers in other content areas. The students' native language may or may not be taught as an academic subject, but it often serves as the medium of instruction in classes for students whose proficiency in English is too limited for them to benefit from instruction presented in English. When the native language *is* taught as an academic subject, the rationale is usually that developing native language proficiency first will facilitate and enhance the subsequent acquisition of English.

Not all bilingual programs in this country are of the type just described. There are also programs designed to develop, in American school children, the ability to speak two languages. Such programs are often referred to as "additive" bilingual programs. Most often, such programs are not Federally funded under the Bilingual Education Act. They generally exist by virtue of local school district, or possibly, state initiatives.

Formal Federal involvement in bilingual education in this country began with the Civil Rights Act of 1964 and was extended by the Bilingual Education Act of 1968. Neither of those pieces of legislation, however, was prescriptive as to what action needed to be taken to assure language-minority students equal educational opportunities. It was not until the 1974 *Lau v. Nichols* Supreme Court decision that it became clear that something other than regular school services had to be provided. Even that decision left it up to state educational agencies to decide what services were appropriate. Nevertheless, it was the *Lau v. Nichols* decision that provided the impetus for most state and local educational agencies to design and implement bilingual education programs in earnest (see Chapter 1 for additional detail on the legislative history of bilingual education in this country).

The Need for Improved Evaluation Practices

The first requirement to evaluate and report on Federally-funded bilingual programs was laid out in the 1977 bilingual education regulations. Guidance on how such evaluations should be conducted, however, was minimal. This fact, the lack of evaluation expertise at the local level, the low priority and low funding levels provided for evaluation activities, and the technical difficulties inherent in conducting bilingual program evaluations all combined to produce the not surprising outcome of basically useless data. Although several evaluation guidebooks were developed with Federal funds (e.g., Bissell, 1979; Horst et al., 1980; Perez & Horst, 1982), they were unsupported by adequate dissemination and technical assistance systems and had little impact on practices. When systematic reviews of the bilingual education evaluation literature were conducted (e.g., Baker & de Kanter, 1983; Dulay & Burt, 1978; Okada et al., 1982, 1983) only a few evaluations could be identified that met minimal standards of methodological adequacy (see Chapter 2 for more detail on methodological problems and their causes).

The present document represents a renewed attempt on the part of the Federal government to improve the quality of bilingual education program evaluations. It is the first step of a developmental process that will, it is hoped, culminate in a bilingual education evaluation system incorporating methodologically sound designs and procedures validated through field tryout and revision. A major goal for the system is that it be useful at the local level for program improvement purposes. A second objective is that it yield comparable outcome data so that, through appropriate comparisons and aggregations, it will finally be possible to address such questions as what kinds of treatments are most effective for what kinds of students in what kinds of settings and to identify effective instructional practices.

A Validity-Based Framework for Evaluation

We began our efforts to build a comprehensive framework for such a system with an extensive review of the literature. Part of this review focused on the kinds of inferences that might be drawn from program evaluations and the many threats to

the validity of those inferences that have been identified. Four kinds of inferences are discussed in the literature and each is affected by a separate, identifiable type of validity.

Inference	Validity
The students treated, the treatment itself, the setting in which the treatment was administered, and the outcome measures used were all consistent with the research hypothesis being investigated.	Construct
The treatment did indeed have an effect.	Statistical Conclusion
The observed treatment did indeed result from the project.	Internal
The study findings can be generalized to *other treatments, outcomes, students, and/or settings.*	External

To the extent that a particular type of validity is increased, the credibility of its corresponding inference also increases (Lindvall & Nitko, 1981). Similarly, the "amount" of each validity is dependent on how successfully the relevant threats are controlled. A total of 34 threats relevant to the four kinds of validity have been identified and are discussed in Chapter 3.

Ideally, an evaluator will thoughtfully analyze everything that could go wrong in an evaluation, enumerate all the plausible rival hypotheses, and then rule them out one by one during the evaluation's planning, implementation, and analysis stages (Cook & Campbell, 1979). This process is similar to Campbell and Stanley's (1963) "patched-up" design in which specific controls are added, one after the other, to rule out different potential sources of contamination.

> As part of this strategy, the experimenter must be alert to the rival interpretations (other than the effect of X [the program]) which the design leaves open and must look for analyses of the data, or feasible extensions of the data, which will rule these out. (p. 227)

Validity is a matter of degree and by eliminating threats to it, greater confidence is gained about the conclusions drawn regarding treatment effects and their generalizability. If an extraneous influence on outcome measures (threat to validity) cannot be controlled either by the design of the evaluation or by the methods of statistical analysis, its potential biasing effect should be recorded, and the results interpreted accordingly.

The evaluation system we envision should encompass both process (qualitative) and outcome (quantitative) components. The process evaluation is "an analysis of the processes whereby a program produces the results it does" (Patton, 1979, p. 334). It will entail measuring program implementation and the characteristics of students and settings which may interact with outcome measures. Process data can also contribute to the outcome evaluation by providing insights regarding how and why certain results were obtained, and by suggesting variables that need to be controlled. What we are trying to avoid is a simplistic approach to evaluation in which "clients are tested before entering the program and after completing the program, while what happens in between is a black box" (Patton, 1979, p. 324). Implementation information can also be used to monitor the program's progress toward reaching its process objectives.

Planning the evaluation. In planning an evaluation, careful thought should be given to each of the four types of validity discussed in Chapter 3. Strategies for reducing threats to each of them should be examined. In terms of construct validity, the evaluator should exert whatever influence he or she has to see that the treatment is carefully defined, is of the type the project director wishes to implement, and is uncontaminated by other constructs. The evaluator might point out, for example, that if a bilingual immersion project includes a computer-assisted language-development component, it will be difficult to determine whether observed outcomes should be attributed to the immersion strategy or to the computer-assisted in-

struction. A design in which some students received just immersion, some just computer-assisted instruction, and (perhaps) some both, would solve the problem. A second concern related to the construct validity is that, in the presence of high student attrition, the sample of students for whom complete data are available may not be representative of the students served.

When selecting an evaluation design, the evaluators' primary concern should be internal validity. The feasibility of implementing a particular design must also be considered, however. Unfortunately, designs with inherently high internal validities may be impossible to implement in bilingual education settings--or may be implementable only under conditions that pose serious non-design-related threats to their internal validities. These issues are discussed later in this Summary and Recommendations section and in considerable depth in Chapters 5, 6, and 7.

Statistical conclusion validity should be considered in conjunction with the size of the evaluation sample. With projects serving large groups of students, this issue may be trivial. In the case of smaller projects, however, planning should consider the possible need to aggregate data across years or across projects if suitable "matches" can be found. The construct validity of the evaluation sample must, however, always be kept in mind. Other factors related to statistical conclusion validity include the reliability of measures, the extent to which program implementation is standardized, and the extent of quality control over the data collection and analysis processes. All threats to validity can be at least partially avoided through careful planning.

External validity is not something that local-level evaluators need worry much about. Meta-analysts and conductors of national evaluations are the ones for whom external validity becomes a major concern. Their efforts, however, will be greatly aided if local projects carefully document all important treatment, student, and setting variables as discussed below.

Documenting Treatment, Student, and Setting Characteristics

There is great diversity in bilingual education. Students from many different ethnic, linguistic, socioeconomic, and educational backgrounds are served at all grade levels in schools with dissimilar student body compositions in many different types of communities where special programs for non-majority children experience varying degrees of acceptance. To further complicate matters, different instructional strategies are implemented by staff with a wide range of professional and linguistic competencies in programs of varying intensities and durations. All of these various factors are thought to interact in possibly complex ways so that there can be no simple answer to the question, "How well does bilingual education work?" It would be more appropriate to ask, "How effective are different bilingual education treatments for different types of students in different settings"?

If, indeed, the issue of effectiveness is as complex as is suggested by the preceding question (and there is at least some evidence that it is), then all relevant characteristics of students, settings, and treatments must be carefully documented as an integral part of any bilingual education program. Failure to do so would run the risk that educationally significant relationships would be obscured whenever data were pooled across different types of students, treatments, and/or settings.

Chapter 4 is devoted to discussions of treatment, student, and setting variables that have been identified as potentially interactive on the basis of either theoretical formulations or empirical findings. Lists of these variables along with methods for obtaining and documenting relevant information are presented in Tables 2 through 4.

Most of the characteristics that need to be documented are relatively easily determinable matters of fact. Some of the treatment variables, however, can only be determined through classroom observation. It is the treatment *as implemented*, not the treatment *as intended*, that is evaluated. The actual treatment, unfortunately, may bear little resemblance to what was intended and may, consequently, have very low construct validity relative to what the study set out to evaluate.

Treatment characteristics. There are four widely recognized types of bilingual education programs for language-minority, limited-English-proficient students: early-exit transitional bilingual education programs, late-exit transitional bilingual education programs, immersion programs, and English as a second language programs. In addition, the absence of any treatment is often referred to as submersion.

In both immersion and ESL, instruction is conducted in English. In immersion programs, however, the teachers are supposed to be bilingual and able to respond in the students' native language (L1) to student questions posed in L1. In both early- and late-exit programs, instruction is initially presented in L1. It is used less frequently and for a shorter duration in early-exit programs than in late-exit programs. Literacy skills are developed only in English in early-exit programs, whereas L1 and English literacy skills are developed concurrently in late-exit programs. The theory behind late-exit programs is that students will learn English better if they first develop proficiency in their native language.

There is a good deal of theoretical debate over which type of program is most effective. At present, however, the consensus appears to be that some students will do best in one type of treatment while others will do better in a different type. In Canada, immersion has been found to be highly effective for teaching middle-class, language-majority students a second language. There is some research indicating immersion programs in the U.S. are not as effective with low-socioeconomic status, language-minority children. Additional research on these programs is needed; still, it would be a mistake not to document this gross-level treatment characteristic. We recommend, however, that treatments be operationally defined in terms of such variables as percentage of instructional time devoted to L1 language arts, percentage of instructional content areas taught in L1, and the grade levels at which instruction in L1 is provided. There is a great deal of variation on such variables even among programs given the same label. There may even be some overlap between programs given different labels. In any case, the characteristics of instructional treatment, materials, staff, and setting should be documented (as they actually exist, rather than as they were planned). All of these treatment characteristics are at least potentially relevant to program impact.

Student characteristics. In addition to such widely recognized achievement-relevant characteristics as socioeconomic status and parents' educational level, a review of the bilingual education literature reveals a variety of other factors that may affect the outcomes of educational treatments. Not all of the research findings are consistent, but some characteristics are clearly important. Among these are: ethnicity/culture, age, L1 literacy, length of time in the country, and prior educational experiences. A number of research findings run counter to conventional wisdom. It appears to be untrue, for example, that "younger is better" for second language acquisition (except in the case of pronunciation).

Again, the implications are clear. What works for one group of LEP students may not work for another, and it is important to document all student characteristics carefully so that meaningful comparisons of different evaluations can be made.

Setting characteristics. Community and school settings are also believed to be relevant to bilingual education program effectiveness. A good project evaluation will include information such as the poverty level of the community, language usage in the community, and school administrative support for the program.

Measuring Growth

When one considers the gains that LEP students make in English language proficiency and subject matter knowledge over time, it is important to recognize that some of that growth results from the bilingual program in which they are participating and some results from other influences such as television, social interactions, and non-program school experiences. While our primary interest may be in assessing the amount of growth that results from the bilingual program, it is almost always a prerequisite to that objective that we measure total growth. At least we must have the tools and skills required to measure total growth if we intend to identify that portion of it which can be attributed to the treatment. In this document we have decided to treat the measurement and attribution issues separately.

If we had perfect instruments, measuring growth would be no problem. Unfortunately, deficiencies in the available instruments make growth measurements

subject to both random and systematic error. Random measurement error (unreliability) is usually ascribed to test characteristics but, in fact, is as much a function of the test takers and the testing environment as of the test itself. Misinterpretation of test items, luck in guessing, variations in mental alertness, and the number and intensity of distractions during the testing session are just a few of the factors that make test scores imperfect indicators of "true" achievement levels. Lengthy tests and multiple measurements tend to minimize those problems--and when large numbers of students are tested, the means of their scores are very stable indices even when the individual scores are unreliable. Chapter 5 contains a lengthier discussion of test unreliability and other random measurement-related error.

Systematic error is often referred to as bias. Unlike random error, which tends to cancel out when data from a large number of items, situations, and/or individuals are aggregated, systematic error produces scores that are consistently either too high or too low. Aggregating across units does not reduce this bias. And the most difficult thing about systematic error is that its presence, unlike that of random error, is not always easy to detect or quantify.

In Chapter 5, four sources of systematic error associated with simple measurements of growth are discussed: (a) stakeholder bias, (b) statistical regression, (c) cultural and linguistic bias, and (d) curricular irrelevance. Stakeholder bias tends to spuriously depress pretest scores and/or spuriously inflate posttest scores. The effect is thus to inflate growth estimates. Fortunately there are ways (discussed in Chapter 5) for eliminating, or at least minimizing stakeholder bias.

Statistical regression works in the same direction (i.e., so as to inflate gain estimates) but is somewhat more predictable with respect to magnitude. Without going into technical detail, whenever students are selected from a group because of low scores on a test (eligibility for a bilingual program is usually contingent upon scoring below some cutoff on a language-proficiency test), scores on subsequent testings will move toward the mean score of the original group *in the absence of any special treatment.* The amount of movement is predictable from the reliability of the test and the original distance that the mean score of the selected students was below

the mean of the total group that was tested (exactly predictable in theory, less accurately in practice). The predicted movement can be used to adjust statistically for the bias due to statistical regression.

Cultural bias works to depress both pre- and posttest scores. It can usually be assumed, however, that posttest scores are somewhat less depressed than pretest scores because of acculturation occurring between the two testings. This factor, too, works to inflate growth estimates. If care is taken to select tests that have few, if any, biased items, and if students can be taught some test-taking skills before pretesting, however, this source of bias can probably be kept within tolerable limits.

Curricular irrelevance refers both to the testing of material that was not taught and to the non-testing of material that was taught. The effect here is that posttest scores will be lower than they would be with greater curricular relevance. Growth estimates will thus be depressed. The solution to this problem, of course, is to select tests that have high degrees of curricular relevance.

In the second half of Chapter 5 we discuss and make recommendations regarding the types of tests and other measures that should be used for assessing the impact of bilingual education programs.

Bilingual education programs, as discussed here, have two primary objectives: (a) developing English language proficiency in LEP students, and (b) preventing LEP students from falling behind their English-proficient peers in other academic subjects while they are learning English. Individual programs may have additional objectives that local educators regard as equally important. The two cited here, however, are legislatively mandated for all public-school programs serving LEP children--*whether or not they are Federally funded.* For this reason we begin our discussion by considering these objectives.

English language proficiency. A substantial amount of professional literature has been devoted to the topic of what constitutes language proficiency. The current fashion distinguishes between (a) linguistic and (b) sociolinguistic or communicative competence, with linguistic competence typically subdivided into the

four components of listening, speaking, reading, and writing. Sociolinguistic competence refers to a student's ability to recognize the appropriateness of particular communications and to interpret them appropriately in particular contexts.

The four components of linguistic competence are closely interrelated. Some theorists believe, therefore, that they should be measured together as no more than different aspects of a single trait. Others disagree and argue for separate measures. The majority of language proficiency measures currently available measure only oral language proficiency and yield a single index of proficiency level.

Unfortunately, language proficiency tests appear to have serious psychometric inadequacies, especially when used for evaluation purposes (a usage for which they were not designed). Although standardized reading readiness and reading tests may be criticized on the basis that they do not sample all important areas of language proficiency, these instruments appear to offer significant psychometric advantages. We recommend that they be used as soon as program participants are able to respond to them in a non-random manner. Until such time as they can understand the test questions and respond appropriately to them, however, their scores will be meaningless and nothing is to be gained by collecting and analyzing them. Oral language proficiency tests may be the only meaningful alternative, but evaluators should choose among these carefully and be aware of their shortcomings.

Out-of-level testing should enhance test item comprehension and should be a viable strategy to use for English language arts testing, since the content of below-level tests is likely to match the language instruction LEP students are receiving. Below-level testing may be unsuitable in other areas because of content mismatches.

One of the most frequently discussed problems in bilingual education evaluation is the lack of appropriate instruments. This is not so much a problem in the area of English language reading and language arts, where English is the appropriate testing medium and a variety of relevant tests are available. It is in other academic areas, especially when instruction is conducted in L1, that instrumentation issues become especially problematic.

We have recommended that, where instruction is conducted in L1, testing also be undertaken in L1 (unless students are more proficient in English). Under these circumstances, the first choice for an appropriate instrument would be an already existing L1 test that has been professionally developed and standardized, and is psychometrically sound. Such instruments are very rare, however, especially when L1 is a language other than Spanish.

If a professionally developed and standardized English-language test with high construct validity is available, it may be usable with extended time limits or other modifications (see following paragraph). Tied for last place in our hierarchy of choices would be locally developed tests and tests locally translated into L1 (see Chapter 5). Despite their deficiencies, we suggest that even teacher-made, end-of-term tests are likely to yield useful information.

If instruction is in English but students are not fluent in English, the best choice of an outcome measure is a standardized achievement test with high reliability and content validity (if one is available), *despite the fact that scores will be spuriously low because of the language difficulty*. We recommend countering the language difficulty by providing the administrative instructions in L1, extending the time limits, and even translating individual words that the test takers do not understand (although these strategies must be standardized so that they are the same at both pre- and posttest times). These strategies will certainly not remove the effects of language difficulties, but they should minimize them. The important thing is to try to be sure that the test is measuring content knowledge and not English vocabulary. If the pretest measures vocabulary and the posttest measures the intended content area, growth estimates will be meaningless.

Chapter 5 provides more detail on all of these points and also discusses measures of academic aptitude and affective states.

A final point, and one that is discussed in detail in Chapter 3, is the desirability of obtaining multiple measures for each outcome. We are aware that practical constraints and testing burdens limit what can be done along these lines. But even combining teacher judgments and classroom grades with test scores will

enhance the credibility of an evaluation by contributing to the construct validity of the outcome measures.

Establishing Cause-Effect Relationships Between Treatments and Outcomes

In Chapter 6, we discuss eight evalution designs that have been reported in the literature and have been used, if not for bilingual education evaluations, for the evaluation of other educational or social interventions. Six of the eight designs yield no-treatment expectations and thus, when properly implemented in appropriate circumstances, provide a methodologically sound basis for estimating how much of the growth students are observed to make can be attributed to the treatment and how much is non-treatment-related. The other two designs do not yield no-treatment expectations but add information to the simple measurement of growth which contributes to the interpretability of data resulting from their implementation.

There are serious problems associated with the implementation of all six designs that yield no-treatment expectations. One requires random assignment of students to treatment and no-treatment (control) groups. A second requires a highly comparable no-treatment comparison group. Implementation of these two designs is essentially precluded by current civil rights and bilingual education legislation.

Three other designs--the grade-cohort, value-added, and regression-discontinuity designs--all hold some promise for application to bilingual education evaluation--but only in special circumstances that are likely to occur infrequently. In the case of the value-added design, we concluded that applicability was too limited to merit inclusion of the model in any bilingual education evaluation system.

The final design--the norm-referenced design--was judged to be unsuitable for bilingual evaluation applications, although it appears to have merit for impact assessments of educational interventions serving language-majority students. The reason it is unsuitable for use in bilingual settings is that it rests on the fundamen-

tally unsound assumption that, without treatment, LEP students would maintain their status with respect to national norms.

We advocate use of the non-equivalent comparison group design when and if an "only slightly non-equivalent" comparison group can be found. We advocate use of the regression-discontinuity design (with curvilinear regression equations) whenever situations can be found where all students both above and below the cutoff scores are representatives of a single language-minority population. We also advocate use of the grade-cohort design discussed in Chapter 6 whenever pre-treatment scores of new program entrants can provide a baseline for students of the same age who have been program participants for some time. We expect, unfortunately, that the opportunities for such applications will account for substantially less than the entire population of bilingual programs.

Where models yielding valid no-treatment expectations cannot be applied, we believe that growth in areas of intended program impact should still be measured. Such growth assessments can be used for effectiveness comparisons among different treatments serving similar target groups in similar settings--or among similar treatments serving different target groups in similar settings--and so on. Criterion-referenced and gap-reduction[1] interpretations can further enhance the meaningfulness of simple growth estimates, and we particularly recommend the gap-reduction approach which is described in Chapter 6. When coupled with process evaluation data, one can use gap-reduction information to draw inferences about causal linkages on logical grounds.

Aggregating Data and Making Effectiveness Comparisons

The fact that treatment, student, and setting variables all interact with one another and with program outcomes does not mean that no meaningful comparisons

---

1. Gap-reduction designs may employ a variety of gaps. We recommend focusing on the gap between the performance level of the project students and that of either their nonproject grade mates or the 50th percentile of the national norms.

can be made among different programs or that data cannot be aggregated across them. In order to do these things, however, outcomes must be measured on a common scale.

In Chapter 7 we discuss several approaches to the common-metric issue and note that the index typically used in meta-analysis is not ideally suited for comparison and aggregation purposes. The advantages of a "nationally standardized metric" are discussed but the conclusion is reached that its utility for bilingual education evaluations is limited.

Unfortunately, we expect that there will be many situations in which it will not be possible to obtain any estimate of effect size for bilingual education projects. For this reason (and because we recommend that total growth be measured even when it is possible to obtain treatment-related growth estimates), we also need a common metric for quantifying growth. After considering all of the alternatives, our final recommendation was to use a new metric specifically developed for this purpose, the Relative Growth Index. This metric is the standardized raw- or scale-score growth observed in the treatment group minus the standardized growth exhibited by the nonproject comparison group expressed as a percentage of the comparison group's growth. An RGI of 0% suggests that program participants are exactly keeping up with their non-LEP peers (a frequently stated objective for bilingual programs--in non-language content areas, at least). A negative RGI would signify that program students are falling behind their non-LEP peers while an RGI above 0% would signify that they are outgaining them. RGIs do not require the use of standardized achievement tests (unless the evaluator wishes to use normative data in lieu of a "live" comparison group. The metric is independent of group homogeneity and is thus suitable for comparing results between and aggregating them across *similar* projects.

In the final analysis, we believe that reliable, valid, and comparable growth estimates for at least the most salient bilingual education objectives can be obtained through implementation of the practices we recommend. When these measures are interpreted within the validity-based framework we have described and are properly

integrated with process information, we believe that most of the current questions pertaining to bilingual education can be answered.

1. INTRODUCTION

The purpose of this report is to summarize the state of the art in bilingual education evaluation in the United States and to develop recommendations for an evaluation "system" that will be developed, field tested, and disseminated in future phases of this contract effort. The system will provide procedures and materials for evaluating the impact on student achievement of local projects supported by Title VII of the Elementary and Secondary Education Act.

As background, it is important to note that the term bilingual education has a somewhat different connotation in this country from other parts of the Western world--especially Canada. Here we are talking about special instructional programs serving non- and limited-English-proficient, language-minority students--programs that are primarily intended to develop students' English language skills and to prevent them from falling behind their fully English-proficient peers in other academic subjects. In Canada and other Western countries, bilingual education most often refers to programs designed to facilitate the acquisition of a second language by language-majority students. This distinction has important theoretical implications for program and evaluation design that will be discussed later.

While there is no shortage of second-language acquisition programs in this country (including some based on Canadian models), they are not what we generally refer to by the term bilingual education. That term is used almost exclusively to denote the kind of programs described earlier. It is important to note that throughout this report we use the term bilingual education programs to denote special instructional services provided to language-minority, limited-English-proficient students *whether or not* they employ dual-language instruction. Thus, programs that involve no more than English-as-a-second-language instruction are included in our definition.

Legislative History

Bilingual education programs in this country grew out of the constitutionally guaranteed right of all resident children to free and equal educational opportunity.

The Civil Rights Act. Passage of the Civil Rights Act of 1964 was the first step in the movement to provide appropriate instructional services to language-minority, limited-English-proficient (LM-LEP) students. Although the Act did not directly address the language issue, it did stipulate that citizens "regardless of race, color or national origin" should have equal access to federally funded programs and benefits. It was not until six years later that the implications for education were made explicit, however, via a clarifying memorandum issued by the Department of Health, Education and Welfare (DHEW) (see below).

The Bilingual Education Act of 1968. Two years before DHEW's clarifying memo, the Bilingual Education Act (Title VII of the Elementary and Secondary Education Act) was made law. Designed to meet the educational needs of limited-English-proficient students, Title VII provided funds for staff training, purchasing educational materials and equipment, and implementing special programs. The Act supported a transitional bilingual education approach, but gave school districts wide latitude in implementing programs. The definition of what constituted a bilingual education program was vague in the 1968 legislation, and no specific evaluation criteria for determining program effectiveness were provided.

The May 25 memorandum. On May 25, 1970, the Department of Health, Education and Welfare issued a memorandum stating that school districts must rectify the "language deficiency" of "national origin-minority group" children so that they could participate effectively in educational programs.

> Where inability to speak and understand the English language excludes national origin-minority group children from effective participation in the educational program offered by a school district, the district must take affirmative steps to rectify the language deficiency in order to open its instructional program to those students. (Pottinger, 1970, pp. 102)

The memorandum also restricted the use of tracking, and required the removal of students from special (language) ability grouping as soon as their linguistic deficiencies were remedied. No guidelines were provided in the memorandum specifying what "affirmative steps" should be taken to remedy language deficiencies.

It did, however, lay the groundwork for the *Lau v. Nichols* decision (Epstein, 1977; Holt & Arellano, 1980; U.S. Commission on Civil Rights, 1975).

Lau v. Nichols. In 1974, the Supreme Court ruled that equality of educational opportunity was not ensured by the San Francisco School District's policy of "merely providing [Chinese] students with the same facilities, textbooks, teachers, and curriculum...[since] students who did not understand English are effectively foreclosed from any meaningful education [by that policy]" (*Lau v. Nichols*, 483 F. 2d at 566). Significantly, the Supreme Court did not suggest any specific remedies, stating that educational policy was a state function and remedies should be designed by those with educational expertise.

Shortly after the *Lau* decision, the Equal Educational Opportunity Act of 1974 was passed. The 1974 Act required all public school districts to comply with the *Lau* decision, even if they did not receive Federal assistance.

The 1974 amendments. In 1974, the Bilingual Education Act was amended to specify, in greater detail, the policies and procedures local and state educational agencies were expected to follow. The amendments also directed the Commissioner of Education to develop and disseminate bilingual education program models. Finally, they provided funds for research to promote the effectiveness of programs for LEP students (Holt & Arrellano, 1980).

The Lau remedies. In 1975, a year after the *Lau* decision, the Office of Civil Rights provided a set of guidelines for the provision of bilingual educational services. These guidelines came to be know as the "*Lau* remedies." They deviated from the *Lau* decision in several important respects. The *Lau* decision identified target students as those who have "linguistic deficiencies" in English, whereas the remedies identified eligible students as those who have a "primary or home language other than English." The remedies also extended the provision of bilingual education services to students who were equally proficient in English and their native language, but were "underachieving" in school.

The *Lau* remedies stated that districts with 20 or more students from any non-English language group must provide a transitional bilingual-bicultural program for them. The transitional model described in the *Lau* remedies included (a) the development of basic skills in the student's native language (L1) first, and subsequent development of these skills in English; (b) recognition of a student's culture and heritage; (c) bilingual instruction for students in kindergarten through grade 12; and (d) remedial instruction for "underachieving" students who had been exited from the bilingual program.

The *Lau* remedies, although legally only guidelines, acquired the force of regulations as a result of the Office of Civil Rights' statement that districts failing to implement them would be found "out of compliance." This threat led districts to comply with the *Lau* remedies as if they were, in fact, legally binding (Epstein, 1977).

Relevant court decisions. Although the *Lau* decision itself did not mandate implementation of bilingual educational programs, and the *Lau* remedies were "merely guidelines," the situation was markedly altered by three landmark court decisions. The *Serena v. Portales* decision in 1974 required the Portales Municipal schools to provide bilingual instruction on a daily basis for 30 to 60 minutes minimum, depending on grade level. It also required that bilingual instruction be provided to English-dominant Chicano and Anglo students. In the *Aspira v. Board of Education of New York* case, the U.S. Court of Appeals ruled in 1975 that the ESL instruction provided to LEP students in New York City schools did not meet their linguistic needs, and mandated the introduction of a program to develop English language skills. The decision also outlawed the use of pullout and immersion programs and established standards for identifying students entitled to bilingual instruction as well as qualifications for bilingual teachers (Holt & Arellano, 1980).

The **Rios v. Read** decision in 1977 stipulated that the quality of a bilingual program should be assessed to determine compliance with the *Lau* remedies. The court ruled that simply providing a bilingual program was not sufficient to satisfy these guidelines. The program should be designed "to assure as much as is

reasonably possible the language deficient child's growth in the English language" (Holt & Areliano, 1980).

The 1977 bilingual education regulations. Federal regulations governing bilingual programs were published in 1977 and required that programs funded on a multi-year basis submit evaluation reports twice annually. Evaluations were to be based on programs' stated objectives and were to include comparisons of students' English and native language reading skills with estimates of their probable performance in the absence of the bilingual program. Reports were required to include pre- and posttest reading scores (mean scores and standard deviations), and appropriate tests of statistical significance.

The 1978 amendments and 1980 regulations. Additional amendments to the Bilingual Education Act were enacted in 1978. And in 1980, the Federal government published new regulations reflecting the amendments. For the first time, funding was provided for demonstration projects, and the regulations emphasized the need to institutionalize programs. The requirements for student selection and evaluation were altered slightly, requiring programs to adopt measurable criteria for determining when program participants no longer needed special language instruction and to conduct individual evaluations of students enrolled in bilingual programs. Evaluation plans were required to include methods for measuring the acquisition of English language skills and strategies for using evaluation results to guide program improvement. Evaluations were also required to assess attainment of each program objective and utilize comparison procedures to estimate the academic performance of program participants in the absence of any treatment. The results of these annual evaluations were to be used by the Department of Education in making continuation awards (Holt & Arellano, 1980; Liebowitz, 1982).

EDGAR. Federally funded bilingual programs were also required to comply with the Education Department's General Administration Regulations (EDGAR), promulgated in 1980. The primary goal of these regulations was to increase the accountability of Federally funded programs. EDGAR established criteria for judging the evaluation component in funding proposals. These criteria were (a) the appropriateness of evaluation methods to the proposed instructional models and (b)

the extent to which they would produce quantifiable data. Funded programs were required to submit annual evaluations of progress toward achieving their objectives and the impact of the program on participants. In addition to annual evaluations, performance reports had to be submitted which contained comparisons of projected goals with actual accomplishments, explanations for failure to achieve goals, and an analysis of unexpectedly high costs (National Center for Bilingual Research, 1982).

The 1984 Amendments. The Bilingual Education Act was reauthorized in 1984, adding two significant new provisions. First, school districts were required to inform the parents of LEP students, explaining why their children needed special language instruction, describing the different programs that were available, and indicating that they had the right to decline enrollment in any of them. The second significant provision of the new legislation was the authorization of funding for "special alternative" programs that did not require the use of native-language instruction. Programs using an immersion strategy, which were specifically excluded from funding in the past, qualified for Federal assistance under the 1984 Amendments.

The evaluation requirements contained in the 1984 Amendments (P.L. 98-511, section 733) require documentation of (a) the educational background, needs, and competencies of LEP students participating in bilingual programs; (b) the educational activities supported by Federal funds and pedagogical methods, techniques, and materials; (c) the competencies and qualifications of staff implementing the bilingual program; and (d) the degree of educational progress attributed to program participation

> measured, as appropriate, by (a) tests of academic achievement in English language arts, and where appropriate, second language arts; (b) tests of academic achievement in subject matter areas, and (c) changes in the rate of student grade-retention, dropout, absenteeism, referral to or placement in special education classes, placement in programs for the gifted and talented, and enrollment in post-secondary education institutions.

The June 19, 1986 regulations specify that the evaluation design include:

> ...a measure of the educational progress of project participants when measured against an appropriate non-project comparison group. (34 CFR, section 500.50)

The regulations further specify that (a) evaluations be representative of all person, schools, or agencies served by the funded program; (b) instruments and procedures used in evaluations provide reliable and valid measures of the program's progress toward accomplishing its objectives, taking into account the characteristics of the population served; and (c) data collection procedures be employed that minimize error by ensuring proper administration of instruments, accurate scoring and transcription of results, and use of appropriate analysis and reporting procedures. The regulations also specify that evaluations provide objective and valid measures of achievement gains in English language proficiency, native or second language proficiency (for developmental programs), and other academic subjects. Finally, they require documentation of the educational achievement of current program participants (including those who are limited-English-proficient, English dominant, and reclassified LEPs), the amount of time participants receive special instructional services, and their progress toward attaining proficiency in English.

The History of Bilingual Education Evaluations

As Federal funds for bilingual education increased during the early years of the program, concerns about program effectiveness increased correspondingly. These concerns were reflected in the increasingly stringent evaluation requirements spelled out in successive iterations of both the legislation and the regulations. Also indicative of these concerns are the several large-scale program evaluations that have been funded by the Federal government and numerous systematic reviews of the literature that have been undertaken in attempts to determine how effective the program has been. There have been multiple attempts to develop systematic guidelines for evaluating bilingual programs--several of them Federally funded.

Despite these and many varied efforts, it is safe to say that very little is known about the benefits, if any, that have accrued to program participants. Since some 1.7 billion Federal Title VII dollars and certainly several times that amount of state and local dollars have been spent on bilingual projects for which there is so

little evidence of success, it is not surprising that the present Secretary of Education and many others are concerned about the program's cost-effectiveness.

Although yearly evaluations of local projects have been required since 1977, policy makers felt there was a need for additional evidence of bilingual projects' progress, implementation, and effectiveness. Based on this perceived need, four large-scale studies have been undertaken.

The 1972-73 Development Associates study. The first of the large-scale Title VII evaluation studies was conducted by Development Associates in 1972-73. This exploratory study collected descriptive statistics about Title VII programs; assessed the impact of the Office of Education's policy on Title VII program management and operation, and the extent to which programs adhered to OE guidelines; and evaluated the usefulness of products and services provided by special research and development projects. The study found that a high degree of enthusiasm and commitment existed among personnel involved in Title VII programs, and that the programs had fostered institutional recognition of the needs of LEP students. There appeared to be a continued need for technical assistance in management and contracting procedures, language training for teachers, curriculum development, and procurement of classroom materials (Development Associates, 1974). The study did not examine student outcomes.

The 1973-74 General Accounting Office study. During the 1973-74 school year, the General Accounting Office (GAO) examined the Office of Bilingual Education's (OBE) implementation of Title VII legislation. Based on a review of 20 funded projects, GAO concluded that OBE had failed to evaluate and monitor the implementation of programs adequately. As a result of this failure, GAO concluded, little progress had been made in identifying effective bilingual instructional methodologies, training bilingual education teachers adequately, and developing useful instructional materials. GAO's assessment of the evaluation reports submitted by projects was that they "were of little use" (General Accounting Office [GAO], 1976).

The AIR impact study. In 1977, the Office of Education commissioned the American Institutes for Research to conduct the first national impact study of Title VII programs. The results indicated that on the average, Title VII students were performing better in math than their counterparts in mainstream classrooms; however, the latter were performing better in English language arts. The validity of these findings has been criticized on the basis of methodological flaws in the evaluation design, especially the dissimilar initial linguistic competence of the treatment and comparison groups (Cervantes, 1979).

Some of the less *technically* controversial findings of the AIR report included the following facts: (a) only a third of the bilingual program participants were of limited-English-proficiency and (b) 86% of the interviewed program directors reported intentionally keeping children in the program after they believed they could function effectively in mainstream classrooms (Danoff, 1978). These findings are indicative of problems that are endemic to Title VII (as well as other) programs where funding is partially (in the case of Title VII programs) or wholly (in the case of entitlement programs) dependent on the number of target children who can be identified and served.

The Significant Bilingual Instructional Features study. The Significant Bilingual Instructional Features study was a three-year investigation undertaken by a consortium of research organizations headed by the Far West Regional Educational Laboratory and funded by the National Institute of Education. Beginning in 1980, the study was intended to identify, and later cross-validate, the instructional features of successful bilingual educational projects, thereby contributing to the fund of knowledge upon which future programs could be built. The five features identified were:

> (a) congruence of instructional intent, organization and delivery of instruction, and student consequences; (b) use of active teaching behaviors; (c) use of the students' native language (L1) and English (L2) for instruction; (d) integration of English language development with basic skills instruction; and (e) use of information from the LEP students' home culture. (Fisher & Guthrie, 1983, p. 3)

Although this study defined successful bilingual treatments in terms of student outcomes, it used Academic Learning Time (a measure of the amount of time a student is actively and successfully engaged in task-related activities) as a surrogate measure for achievement gains.

Synthesis of local evaluation studies. Several attempts to assess program effectiveness using data from local evaluation reports have also been made (See Chapter 2). Of these, the most widely cited include Zappert and Cruz (1977) who reviewed evaluation reports submitted to government funding agencies and rejected 97% of the studies because they contained serious methodological flaws. Baker and de Kanter (1983) examined some 176 evaluations of bilingual programs and found that only 39 of them were methodologically sound, empirical studies. Okada, Besel, Glass, Montoya-Tannatt, and Bachelor (1982) and Okada, Besel, Bachelor, Glass and Montoya-Tannatt (1983) conducted meta-analyses of Title VII and non-Title VII bilingual programs with the intention of (a) assessing the impact of Title VII capacity building on the ability of schools to meet the needs of LEP students and (b) providing information to improve Title VII program management and operations. More recently, Willig (1985) conducted a meta-analysis of many of the evaluations reviewed by Baker and de Kanter. These syntheses were neither overwhelmingly negative nor overwhelmingly positive (Willig's was the most positive) about the impact of bilingual education programs. The results of the studies did, however, indicate that the quality of bilingual program evaluations was poor.

The Purpose, Objectives, and Scope of This Report

Based on the preceding review of large-scale evaluations and evaluation syntheses, it can be concluded that little is known about the impact of the program on student achievement. Although policy makers and educators all agree that special educational services are needed to help language-minority students obtain an adequate education, there is little consensus as to what instructional approach is most effective for what types of students. Okada et al. noted in 1983 that "researchers and program developers find themselves 14 years after the implementation of Title VII bilingual education, with very little sense of what types of

programs or approaches work for or match the needs of the many diverse linguistic populations" (p. 4).

This study represents another attempt on the part of the Federal government to obtain information about the overall impact of bilingual programs on participating children. Instead of being another national evaluation study, however, this new effort is intended to improve local evaluation practices with the dual goal of enhancing the local utility of evaluation information and providing a data base that will be useful for broader purposes. Although we believe that the question of bilingual education's impact can be only partially addressed by an effort of this type (or by any single national-level study), a methodologically sound, standardized evaluation system should certainly shed new light on the issue.

There is little doubt (as will be shown later in the report) that evaluation practices in bilingual education need substantial improvement (as do the practices employed in evaluating conventional programs). We support the position of the Joint Committee on Standards for Educational Evaluation (1981, p. 5) that "sound evaluation can promote the understanding and improvement of education, while faulty evaluation can impair it." Although the bilingual education evaluation system we envision will generate only rough estimates of the extent to which achievement gains are attributable to bilingual interventions, it should provide teachers, administrators, and parents with useful and accurate information about student performances and program implementation. In this way, the system can meet the local stakeholders' needs for evaluation as well as those of the Federal policy makers.[2]

---

2. While policy makers are generally most concerned about program impact, the needs of local project staff include "obtaining information for modification and improvement of the program, information to support the continuation of the program, and evidence of the effectiveness of the program in comparison to some other locally-feasible alternative" (Gold, 1981). Bissell (1979) provides a more detailed list of the different needs of different evaluation audiences.

Specifically, the two objectives for developing the proposed evaluation system for bilingual education are:

1. To improve the quality of local Title VII project evaluations by providing standardized and methodologically sound evaluation procedures and materials designed to enhance the validity of findings and the utility of evaluation for program improvement purposes.

2. To yield comparable outcome data so that, through appropriate comparisons and aggregations, it will finally be possible to address such questions as what kinds of treatments are most effective for what kinds of students and to identify effective instructional practices.

The prospective system, as we have conceptualized it, will encompass both process and product information, and will be designed for use at the local level. Although primarily designed as a summative evaluation system for determining the merit or worth of a bilingual project, the heavy emphasis on program documentation during the course of the program will provide project staff with adequate information for monitoring and improving program implementation. The system will also reflect a concern for larger issues by addressing topics such as effectiveness comparisons between projects, generalizability, and aggregation.

The evaluation system is designed to minimize threats to the various types of validity that have been identified as important in research and evaluation studies. It will provide a reporting system for local projects specifying what kind of data to collect, and how to collect, analyze, and present them. At the same time, it will allow for variations in local project types, goals, and resources. We believe the explicit Federal evaluation requirements manifested in the evaluation system can increase local evaluation standards. The system builds on existing knowledge and is developed with the realization that local projects will implement only what is easiest and most practical for them.

This document represents a first step in a complete system development, test, and dissemination effort. It attempts to:

- summarize the current state of the art in the evaluation of bilingual education programs (Chapter 2);
- discuss validity issues in evaluation and research and present strategies for maximizing validity in bilingual education evaluations (Chapter 3);
- provide guidelines for the systematic documentation of program, student, and setting characteristics that are important to proper interpretation of evaluation findings in bilingual education (Chapter 4);
- identify measures that are appropriate for quantifying goal-related changes in student achievement and affective status (Chapter 5);
- summarize designs that may be used to relate student outcomes to program inputs (Chapter 6); and
- develop a metric that will enable effectiveness comparisons to be made among programs serving similar target groups in similar settings and the aggregation of data across programs whose impacts are assessed with different instruments (Chapter 7);

In formulating our initial recommendations we have tried to retain as many design and implementation options as we believe might work under *some* circumstances. The nature of bilingual programs is restrictive, however, and several practices that would be useful in other settings (e.g., compensatory education) have been rejected as technically inappropriate or impossible to implement.

2. REVIEW OF CURRENT PRACTICES AND PROBLEMS IN THE EVALUATION OF BILINGUAL PROGRAMS

This chapter examines what may be called the "state of the art" of bilingual education evaluation as determined through an analysis of the pertinent literature. A number of methodological deficiencies common to bilingual evaluation are described. It should be noted at the outset that, while some of these problems are relatively simple to resolve (e.g., through greater methodological rigor), others are not. Except under certain conditions, for example, deriving a valid estimate of how participants would have performed without the program appears to require groups of LEP students who are not participating in bilingual projects, but whose educational needs are similar to those of program participants--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 problem found in the former and not the latter type of evaluation effort. Although the focus of this chapter is on local evaluations, 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.

It has been 19 years since the passage of the Bilingual Education Act in 1968 when direct Federal grants began funding local school districts to develop bilingual programs designed to meet the educational needs of LEP students. The Title VII program is one of several Federally funded programs in education that stress the importance of evaluation. Not only does it demand that every proposal include a detailed plan for demonstrating program effectiveness, 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, evaluations in bilingual education 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 agreed that local evaluation reports are of little value to decision-makers, both at the local and Federal levels (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, et al., 1982; Rodriguez-Brown, 1980; U.S. Department of Education, 1982). Although one recent meta-analysis (Willig, 1985) is more optimistic regarding the efficacy of such interventions, debate continues over the merits of bilingual programs. 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 unfortunate state of affairs is not unique to bilingual programs (Campeau, Roberts, Rowers, Austin, & Roberts, 1975). In examining previous attempts to evaluate the efficacy of special education programs for mildly handicapped children, Tindal (1985) found "serious methodological flaws in these evaluation efforts [which] make our present knowledge in this area very weak" (p. 101). Some of the problems identified include ill-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 recent summative evaluations and concluded:

> ...the metaevaluation 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 poor does not absolve bilingual education evaluations of blame for their own deficiencies. Every effort should be made to improve their quality so that the impact of bilingual education can be more accurately estimated and sound educational practices identified for language-minority students.

In the following pages, we (a) empirically appraise the quality of current practices in bilingual education evaluation, (b) analyze the sources of methodological flaws in bilingual education evaluations, and (c) identify the evaluation needs and the desired characteristics of an evaluation system for bilingual education.

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 reports for further analysis and synthesis. The screening process and its results provide some basis for inferring the state of the art in bilingual program evaluation and some insights into 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 non-qualifying programs were rejected because the evaluation methodology in their program

reports was so flawed that no conclusions could be drawn about 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, non-relevant comparison, (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 performed 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 *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 which 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, or (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 non-relevant (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 and provided 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 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 those submitted during

the 1980-81 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 1980-81 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 were applied in the screening process. The following is a list of these criteria reorganized and simplified by O'Malley (1984):

- General Design Problems
  - no outcome data
  - posttest only, no comparison
  - testing not related to program objectives
  - duration of treatment less than six months
  - no information on duration of treatment
  - pretest data only
- 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%
- Inadequate 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
- Inappropriate Metric
  - 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 average acceptance rate was only 10%

(median = 6%). The acceptance rate of each study was undoubtedly affected by the selection criteria employed and the investigators' 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. 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).

Table 1

Studies Accepted for Review on
Effectiveness of Bilingual Education

	Number Reviewed	Number Accepted	Acceptance Rate
Campeau et al. (1975)	175	8	5%
Zappert and Cruz (1977)	600	18	3%
Dulay and 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 and de Kanter (1983)	176[3]	39	22%
		Mean = 10%	
		Median = 6%	
		SD = 8%	

One other meta-analysis study was described in a doctoral dissertation by Gold (1981). Instead of reviewing evaluation or research reports, the author reviewed 75 proposals of 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

---

3. Source: Baker, 1985 (personal communication).

met by the proposals studied...[and] evaluation designs for Title VII programs showed a consistent lack of conventional evaluation rigor" (p. vii).

Although there are some indications that the quality of evaluations has improved over the years (Baker & de Kanter, 1983; Okada et al., 1982),[4] "program evaluations are still of very poor quality" (Baker & de Kanter, 1983, p. 52). Considering the amounts of time and money that have been spent on bilingual program evaluations, the current state of affairs with respect to impact assessments is discouraging at best. In the next section, we discuss the difficulties in performing evaluations in bilingual education that have led to the inferior quality of evaluation studies.

Sources of Methodological Problems in Bilingual Education Evaluations

The preceding review highlights the fact that there are serious methodological flaws in bilingual education evaluation and research reports. Based on the relevant literature (e.g., Baca, 1983, 1984; Burry, 1979, 1981; Cohen & Laosa, 1976; Evaluation, Dissemination & Assessment Center, 1983a; Gezi, 1981; Hubert, 1982, in press; National Clearinghouse for Bilingual Education [NCBE], 1983; Piper, 1984; Rodriguez-Brown, 1980; "Some Common Pitfalls," 1980; Yap, 1984), these inadequacies can be attributed to four 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 competence. Some of the deficiencies in bilingual evaluations are directly attributable to the lack of knowledge and skills of the individual(s) who conduct the evaluations. Shortcomings such as presentations including insufficient

---

4. In the Okada et al. (1982) review of evaluation studies from 1967 to 1979, the percentage of rejected studies conducted between 1967 and 1976 was in the 90s. It dropped to the 80s from 1977 to 1979.

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, 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), 8 (7%) out of the 123 bilingual project evaluators responding specified evaluation and research as their area of concentration; and 2 (2%) indicated specific preparation in bilingual education. Ninety-one percent of the evaluators surveyed were not trained either in evaluation or in 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 that 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 by a group of distinguished evaluation experts. The review panel *added* 34 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; Worthen & Gagne, 1969). In the most recent formulation (Worthen, 1975) the list comprises 25 tasks requiring some 82 skills and/or areas of knowledge. Worthen describes 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[5] and the use of independent auditors or consultants to add credibility to an evaluation.

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 relationships, 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 which brings together the knowledge, skills, and experience of all its members. To select evaluation team participants, Bissell (1979) offered 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)

---

5. Cook and Shadish (1986) perceived the failure of mandated self-evaluation using internal evaluators 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 tend to be 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."

A related suggestion by Bissell (1979) is to form an evaluation monitoring team including administrators, teaching staff, secondary-level students, school board members, and district testing and evaluation staff, whose responsibilities would be to review, comment on, and facilitate all evaluation activities performed by the evaluator or evaluation team. The formation of such an evaluation monitoring team is probably feasible only in large school districts.

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) suggested several ways in which teachers and project directors can assist evaluators to ensure accurate assessment of their programs. In order to capitalize on these working relationships, it is just as important for the evaluators 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 producing adequate project evaluations. Thus it may be necessary to enhance not only the general level of evaluators' competence, but project directors' knowledge of evaluation as well.

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) pointed out that one of the reasons for low quality evaluation is that:

> 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)

Although some contract researchers would take issue with Bernstein and Freeman, we agree that close monitoring systems should be imposed by the funding agency to

ensure state-of-the-art work (a concern we will address later under state and Federal policy).

Administrative practices. Although evaluators are apparently guilty of misguiding the evaluation process and ultimately producing inadequate reports, local administrators who supervise the evaluators must share the blame. Local administrators and project directors often do not appreciate the importance of evaluation and consider it an extra burden required by their funding agencies. Their cooperation in adapting school routines to accommodate evaluation activities is therefore low. In addition, evaluation reports are often treated as public-relations documents (Rodriguez-Brown, 1980). Consequently, project directors 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 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 are also contributing factors to 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 is almost always beyond the evaluation budget. Presently, 3% of a project's total budget is usually allowed for evaluation (N. C. Gold, personal communication, 1985). For small projects, in particular, this funding level is clearly deficient. To compound the problem, evaluators are often hired 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.

State and Federal policy. State and Federal policy impact on the quality of evaluations in much the same manner as local administrative practices. 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)

As noted by Hubert (in press), "nearly all of the major technical problems in conducting evaluations of bilingual education projects are linked to evaluation practices that are required, encouraged, expected, or tolerated at the Federal level." This observation is substantiated by the lack of specific guidelines provided for local evaluations, and by the quality of evaluation plans in approved Title VII project applications (Gold, 1981).

Although Federal regulations for bilingual education evaluation exist, they provide no specific instructions with respect to the ways in which data should be collected, analyzed, and presented. Even the few guidebooks that have been developed for bilingual program evaluations (e.g., Bissell, 1979; Center for the Study of Evaluation, 1980; DeGeorge, 1980; Perez & Horst, 1982), are generally nonprescriptive regarding procedures for assessing cognitive achievement gains. This lack of technically sound and practical standards for conducting evaluations is undoubtedly a contributing factor to poor practices (Yap, 1984). There thus continues to be a real need to develop an evaluation guide that can prescribe uniform procedures and assure technical excellence. In addition, Title VII project proposals should be routinely reviewed by methodologists.

Reports of both local and large-scale evaluations should also be read by individuals who are knowledgeable about research methodologies and who have the authority to take some action. It is by the active monitoring of the "quality" of evaluation and research in bilingual education by competent specialists that improvement can be assured. To avoid stakeholder bias, it has been recommended that evaluations not be monitored by the same office which funds the program (Cook & Gruder, 1978; Laosa, 1985). Although some have argued that the real goal of bilingual programs is to provide bilingual education per se, and not to study its 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.

Hubert (in press) summarized the situation as follows:

> Major improvement in evaluation data cannot be obtained solely through evaluator training and the use of manuals...it is the policy framework which is most amenable to change, and through which substantial improvements in the quality of evaluation data could most readily be sought.

The new rule allowing a six months start-up period for new projects is a prime example of how policy might significantly improve the quality of bilingual program evaluations. Hubert (in press) offered the following suggestions for additional policy changes: a continuing national study, planned meta-analysis, regionalizing evaluation, mandating longitudinal evaluation, and economizing with a sampling strategy. Other 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 use of evaluations, and (4) using LEA evaluation data at the aggregate level" (p. 6).

Characteristics of bilingual education programs. The three factors discussed above (evaluator competence, administrative practices, and state and Federal policy) 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 significantly restrict what can be done in evaluation.

The most salient and obvious feature of a bilingual program is that *all LEP students served are limited in English proficiency, and their native languages and cultural background are different from those of the mainstream population.* This feature means that available affective and achievement instruments are usually not well suited for use with LEP populations. Very often, pre-treatment 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). 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 taking what tests there are.

Since 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 language proficiency tests are the Bilingual Syntax Measures (BSM), the Basic Inventory of Natural Language (BINL), the Language Assessment Battery (LAB), and the Language Assessment Scales (LAS). Unfortunately, "all of these [instruments], according to the Office of Bilingual Bicultural Education of the California State Department of Education, suffer serious psychometric defects" (Piper, 1984). The major criticism is that what is measured by these language tests does not adequately represent the "English language proficiency" construct (Willig, 1985). This measurement issue is discussed more fully in Chapter 5. 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.

By law, *all LEP students must be served.* This requirement effectively eliminates any possibility of employing a true experimental design with random assignment of students to treatment and control groups. 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. While such a delaying strategy may be attractive from a research perspective, 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 make up for the loss of time.

A variation on the random assignment theme is to conduct 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 *more* needy LEP students 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 (see Chapter 6).

Since random assignment is apparently not feasible, an alternative strategy would be to seek out a pre-existing 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 different from those of the experimental group and to compare the relative effectiveness of the different treatments. Even this possibility is remote, however, since the meaningfulness of the comparison would hinge on the two groups being virtually identical on 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 an 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, for example, McConnell, 1982). Unfortunately, the number of situations in which it will be possible to compile the data needed for this type of assessment may be limited. Still other quasi-experimental designs are at least theoretically possible (see Chapter 6) but the unique characteristics of bilingual education programs typically cause non-trivial implementation problems.

Students served by bilingual programs are mobile. Many LEP sudents are either recent immigrants whose families are still in transition, or migrant students who relocate seasonal 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.

Bilingual education may also require an extensive period of time for its effects to emerge. Ovando and Collier (1985) reviewed several studies including Cummins' (1980) paper and concluded tha. 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. Also, 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 employment figures, statistics on drug addiction and alcoholism, suicide rates, and personality disorders" (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 three-year period. If sample sizes are small to begin with, meaningful data analysis may not be possible.

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 non-mobile 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.

The characteristics of students served by bilingual programs vary. 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 three years of schooling will require a very different instructional strategy from 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 them in order to enhance the interpretability of evaluation findings. This point will be discussed in greater detail later.

Treatment in bilingual education varies. Bilingual education treatments traditionally include instructional, 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 educational needs. If teachers are doing their jobs, they will tailor instruction to meet these individual student needs. Secondly, 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, p. 1-2).

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. Due to the unique 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 which 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, 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 possibly in Spanish) are available on the market. Because of the difficulties in achieving the instructional goals created by these two factors, the implementation of instructional components has been uneven.

The last reason why treatment varies is because bilingual programs are new. Very often a program changes and evolves over time to adapt to local conditions, and for purposes of improvement. Program designs are modified due to practical constraints. Instructional strategies and materials are tried, abandoned, adopted, or adapted to meet the demands and the needs of the students and the school. So 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.

Another characteristic of bilingual programs is the *small number of students served by each project.* 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 (Gold, 1985, personal communication). Not only does this situation result in small sample sizes (which means that small treatment effects may not be detected), it means that unusually effective (or ineffective) teachers, or schools with outstanding (or totally inept) leadership can have a marked influence on the results of the evaluation (Horst, 1982). 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, what is needed to correct current deficiencies in local evaluation is:

(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 internal validity that do not require a randomized control group.

(5) a system of comparing and aggregating data across projects.

(6) a system for utilizing setting, student, and process information in outcome evaluations.

The purpose of this list is to provide the foundation for increasing the *validity* of evaluation findings. In the next chapter, the meaning of research validity is explored in order to shed light on the structure of an integrative evaluation system which will incorporate and expand on the above-listed needs for improvement. Then, in remaining chapters, we attempt to deal with the remaining needs. The need for improved evaluator skills is only indirectly addressed in this document, and can probably be dealt with only through a prescriptive and detailed evaluation system.

Summary

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

Local administrators and project directors can affect evaluation practices by the extent of their cooperation with the evaluator. Their concern for adequate evaluation can result in hiring evaluation staff with the necessary qualifications, which in turn has a direct effect on the quality of evaluations. The unique characteristics of bilingual programs, including the variety of student groups served, necessitates evaluation analysis practices that control for socioeconomic status and initial language proficiency. State and Federal policies restrict local administrative operations through funds allocated for evaluation and through deadlines, regulations, and

late awards. Such policies may also contribute to the hiring of incompetent evaluators, because no standards are set. In addition, the state and Federal regulations can have a direct effect on evaluation practices by failing to provide adequate proposal reviews. Actual evaluation practices, in turn, affect Federal policy as evidenced by Federal initiatives to improve bilingual program evaluations.

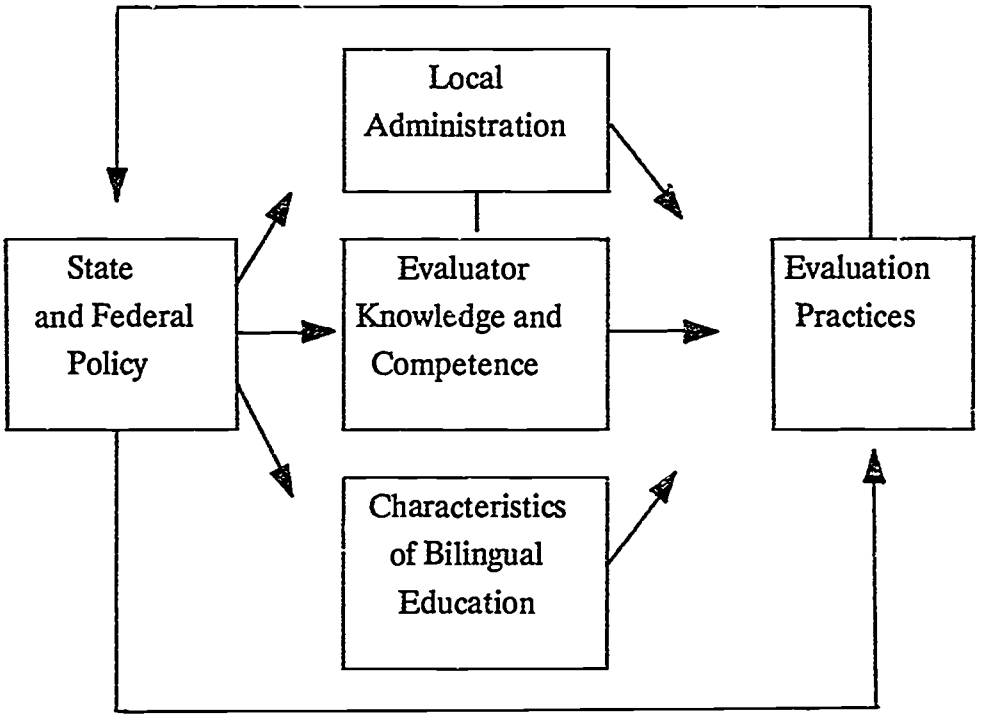


Figure 1. Causal rel -hips among various factors and the technic. quality of evaluation practices.

Discussion and Recommendation

A number of Federal initiatives aimed at improving the quality of bilingual education evaluation have been made in the past but have apparently met with little success (O'Malley, 1984). Before more Federal money is spent on developi. an evaluation system for bilingual education, it seems appropriate to review these efforts briefly in an attempt to establish the direction of this new endeavor.

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 bilingual education training programs in the assessment, evaluation, and dissemination of relevant materials. Although the centers' primary focus was the production and distribution of materials (Rodriguez, Sherman, Pelavin, & Hayward, 1984), numerous workshops on evaluations 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 local projects. More recently, the Evaluation Assistance Centers (EACs), have been assigned all responsibility for the evaluation assistance function. In addition to the supportive services provided by these centers, OBEMLA has periodically sponsored management training institutes for Title VII project directors, designed to familiarize them 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) 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 were developed and field tested. Compared to others of its kind, the project was comprehensive in providing "hands-on" information on how to conduct 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). Although never formally published, the handbook provides numerous forms and instructions for describing and documenting program operations and identifying areas for program improvement.

It also describes procedures for analyzing outcome data to determine student performance levels.

Another attempt by OBEMLA to improve bilingual evaluation practices was its effort to develop validation procedures for demonstration projects (programs of educational excellence). In one project (NCBE, 1983), a panel of bilingual evaluators was formed to work out more relevant (to bilingual education) alternative procedures for validating 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 them. The study's funding period was from January, 1984 to June, 1985.

A concern closely related to the evaluation of bilingual programs is the 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 (DHEW), 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, unfortunately, is intimidating both to practitioners and to 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*, is being conducted by Pelavin Associates, Inc. under contract with the Department of Education's Office of Planning, Budget, and Evaluation. The purpose of the project is to identify procedures and criteria for placing LEP students in and exiting them from bilingual and other special programs. The study has not yet been completed.

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; 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.

> Occasional Papers and some of the papers in the Bilingual Education Paper Series 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)
>
> Guide to Bilingual Program Evaluation (Ulibarri, 1983)

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

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

The Program Impact Evaluations booklet has been well received and widely distributed.

The Midwest BEMSC developed a training module on bilingual education evaluation designs (Secada, 1983) but only in outline form. The BUENO Center BEMSC has "initiated a study of evaluation models and processes...in an effort to facilitate standardization of evaluation practices for Title VII projects" (Georgetown BESC, 1985). The current status of this development effort, however, is not clear.

Most of the guidebooks named above contain a component that deals with language assessment, a key element in bilingual education evaluations. Many ar-

ticles 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, 1978; 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.

It is clear from the preceding review that substantial efforts have been made to improve the quality of evaluation in bilingual education. The seemingly insignificant impact of these efforts can perhaps be attributed to the following problems. First, efforts have focused on disseminating evaluation guides 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 rather limited and unsupported by a a technical assistance system. Third, the documents themselves have tended to be cumbersome, poorly presented, and redundant. Finally, and most compellingly, the materials are generally nonprescriptive. They elaborate on the necessity for certain evaluation practices, assuming that the readers already have or will learn the skills needed to understand and implement the recommendations.

Based on the preceding observations, we believe the following conditions are necessary for success in developing and implementing a new bilingual education evaluation system. First, the evaluation system should be built on the existing knowledge base by incorporating, refining, extending, and elaborating the work that has already been done in bilingual education evaluation methodologies. Linkages with past and current practices should be made explicit. At least the EACs should have some active involvement in the development, tryout, and revision process. Only if they acquire some feeling of ownership for the system will there be effective dissemination of it. Second, the *Users' Guide* should be prescriptive, providing clear how-to-do-it information and real-world examples for readers with various levels of knowledge and skill. Before the system has been completed, the target audiences must be identified. Then they should be made aware of the forthcoming guidelines--preferably through their "friends" in the BEMSCs and EACs (whose support for the guidelines should be earnestly sought). An effective delivery system

should be developed for all system documentation, also involving the BEMSCs and EACs in important roles.

A final precaution is that the *Users' Guide* and its accompanying training materials cannot by themselves make the significant improvement that is needed in bilingual education evaluation. Changes in local, state, and Federal policy will also be required (Hubert, in press). The key is to develop an evaluation system that allows for variations in local conditions and program types, and can be easily adopted without too many extra tasks and complexities for district and program administrators. Most important, the system should have utility for local program improvement.

3. A VALIDITY-BASED FRAMEWORK FOR BILINGUAL EDUCATION EVALUATION

Chapter 2 pointed out that bilingual program evaluation methodology and practices have been so poor that data accumulated over the years provide little convincing evidence about the impact of bilingual programs. Chapter 2 also enumerated the various flaws that rendered evaluations uninterpretable or invalid. In this chapter, we undertake to examine the various aspects of validity that have been discussed in the literature. We hope that this examination will prove useful by providing a systematic framework for readers to use in conceptualizing the material presented in subsequent chapters.

The "validity" concept was borrowed by Campbell and Stanley (1966) from the field of psychological measurement and used by them to describe the quality of various social science research methods and designs. It was later expanded on by Cook and Campbell (1979) and by Judd and Kenny (1981). The validity-based approach to program evaluation was described by Wortman (1983) as "[having] great heuristic value in sorting through the complex issues that inevitably surround any program evaluation" (p. 228). It provides a conceptual framework for understanding the effects of inadequate practices on the quality of evaluations and a guide for developing a comprehensive outcome evaluation system designed to maximize validity within whatever practical constraints may exist.

In this chapter, we explain the meanings of four kinds of validity, describe the conditions under which each of them may be threatened, discuss the relationships and priorities among these kinds of validity, and subsequently explore a general approach for resolving difficulties in bilingual program evaluations. Throughout this discussion we have borrowed heavily from both Cook and Campbell (1979) and Judd and Kenny (1981).

The Meanings of Research Validity

There are four general questions that must always be addressed in educational research and evaluation. These are: (a) Are the constructs involved in the study adequately defined or represented by the treatments, outcomes, samples, and settings studied?, (b) Are the observed outcomes due solely to the treatment and not due to or confounded by other influences?, (c) Is the research design sufficiently precise and powerful to detect the program effects?, and (d) Can an observed causal link between treatment and outcome be generalized to other treatments, outcomes, populations, and settings? These four concerns are the individual aspects of social research validity referred to respectively as *construct, internal, statistical conclusion*, and *external* validity. Their presence, as reflected by affirmative responses to the four questions listed above, are the desirable characteristics of a research investigation or evaluation. Each of these types of validity, however, may be affected by a number of "threats" that could contaminate the results and/or reduce the interpretability of the study. To the extent that these threats are controlled or avoided, the credibility of the research or evaluation findings is enhanced.

Construct Validity

Assume that the Federal government would like an answer to the question, "How effective is bilingual education in helping LEP students attain English language proficiency and other academic goals?" This is, upon close examination, a complex question. To begin with, the terms bilingual education, English language proficiency, other academic goals, and LEP students are all constructs that need to be defined and operationalized before studies can be designed to provide answers to the question. If we wanted our study findings to be generalizable to the population of all bilingual programs, we would have to be sure our sample included all possible program types. We would have to employ appropriate selection and weighting procedures so that we could, with known error probabilities, generalize the findings from our sample to the population of concern--all bilingual programs.

If financial or other constraints prevented us from employing a stratified random sample of all bilingual programs, we might decide to examine only the most

common type of program, transitional programs. Assuming that we studied a representative sample of such programs, we could then generalize our findings to the populations of all such programs with a known probability of error. We would be on shaky ground generalizing to all bilingual programs, however, because our sample lacks *construct validity* for such a generalization--it is an inadequate operationalization of the *construct* "bilingual education"--although it *is* a sound operationalization of the construct "the most common type of bilingual program in the United States."

It should be noted that social science is almost always interested in constructs (e.g., English language proficiency) but research is necessarily conducted with observable operationalizations of those constructs (e.g., scores on a particular language proficiency test). To put it one more way, then, the construct validity of a study is a direct function of the adequacy with which constructs are operationalized.

In educational evaluations, there are always four areas of construct-validity concern: treatment (bilingual education in our example), outcomes (English language proficiency and other academic goals), population (LEP students), and settings (schools). On the following pages, we address each of these areas individually.

Treatment. The construct validity of a treatment is the extent to which the actual program implementation fits the conceptual definition of the program or treatment. According to Sechrest, West, Phillips, Redner, and Yeaton (1979), "it refers to our interpretation of treatments, not to the treatments themselves" (p. 17). In bilingual education, the construct of treatment is difficult to specify and thus to operationalize because there are great variations in program composition (instructional, curriculum development, staff development, and parent/community involvement components) and in instructional models and strategies (see Chapter 4). In addition, as discussed in Chapter 2, programs are implemented with varying degrees of fidelity to their "models." These implementation variations arise through:

> ...normal adaptation to local resources such as skills of personnel, funds, and available facilities, or the natural development of the program over time--the so-called "moving target" problem (Wager, 1979). These changes in what

> Sechrest and his associates (1979) call the "integrity" of the program are viewed by them and others (see Judd & Kenny, 1981) as more appropriately categorized as a threat to construct validity. (Wortman, 1983, p. 227)

Baker and de Kanter's (1981) review of the literature on the effectiveness of transitional bilingual education was criticized for improper definitions of bilingual instructional models such as "transitional bilingual education," "English as a second language," "structured immersion," and "submersion" (Seidner, 1981). If the labels attached to the treatments are incorrect, conclusions based on the study can be in error (Sechrest et al., 1979).

To describe a bilingual program, an analysis must be made of all of the characteristics and activities of all its components. Without these descriptive data, one cannot determine the extent to which the outcomes are attributable to the treatment constructs of interest as opposed to constructs not operationalized by the treatment. This brings us to another validity distinction which is particularly relevant to both treatment and outcome constructs, but which is also applicable to population and setting constructs. In order to have high construct validity, an operationalization must possess the characteristics of both *convergent* and *discriminant* validity. Convergent validity is the extent to which the operationalization of a construct does, indeed, represent the construct of interest. Discriminant validity is the extent to which the operationalization of a construct is uncontaminated by the presence of other, theoretically irrelevant constructs. Taken together, convergent and discriminant validity are the necessary and sufficient conditions for construct validity.

Consider, for purposes of illustration, a hypothetical study of bilingual education that was designed to test the efficacy of a particular instructional strategy. Assume that the teacners in the study were strong advocates of that particular strategy. These teachers might not only take special care to cover all of the curriculum material encompassed by the posttest, but in administering that posttest they might be slightly more helpful in answering student questions than they had been at pretest time. During the interval between pre- and posttests, the teachers might also have devoted some instructional time to test-taking skills.

In this example, the treatment reflects at least three constructs: The bilingual instructional strategy, teaching for the test, and test-taking skill training. The latter two constructs, of course, confound the results of the study and make it impossible to determine just how much of whatever growth was observed could be attributed to the instructional strategy. The extent to which such confounding constructs are not present is reflected by discriminant validity.

To summarize, a treatment's convergent validity is the extent to which it reflects the construct of interest. A treatment's discriminant validity is the extent to which it does not reflect unwanted constructs. Added together, a treatment's convergent and discriminant validities represent that treatment's construct validity.

Outcome measures. The construct validity of an outcome measure is the extent to which it reflects the theoretical construct of interest and does not reflect other, irrelevant constructs. IQ tests, for example, are designed to measure "intelligence." If they are administered to a group of LEP students, however, the obtained scores may reflect not just "intelligence" (convergent validity), but cultural bias, English language proficiency, test wiseness, motivation to perform, and random measurement error as well. The latter sources of variation are not only irrelevant but unwanted. They would systematically bias estimates of the LEP students' intellectual levels (and hence lower discriminant validity).

The construct validity of outcome measures in bilingual education evaluations is undoubtedly affected by students' linguistic, cultural, and educational backgrounds. To the extent that test scores do not reflect subject matter knowledge (low convergent validity) and do reflect irrelevant student characteristics (low discriminant validity), the construct validity of the outcome measure is low.

A major problem identified in Chapter 2 is the lack of valid and reliable assessment instruments for measuring educational achievement and affective growth. As an example, the commonly used English language proficiency tests have been criticized for measuring only some aspects of language proficiency (Fiper, 1984). Such tests can be characterized as having low convergent validity (Gilmore & Dickerson, 1980). Also, it is not clear whether measures of affective outcomes

adequately tap constructs such as "attitude toward school" and "ethnic pride" or are heavily contaminated by students' desires to make socially desirable responses.

Student samples. The construct validity of student samples is the "extent to which the specific students tested in a study represent the theoretical population of interest (convergent validity) and do not represent populations of no theoretical interest (discriminant validity)" (Judd & Kenny, 1981, p. 23). The population of interest, of course, can be defined in any way the investigator wishes. Definitions could range from narrow--"high school Vietnamese refugee students enrolled in the Los Angeles school district who have missed at least two years of schooling and whose English language proficiency is classified as limited by the LAS test"--to broad--"language-minority students in the U.S." What is important, however, is that the sample reflect the definition.

One of the technical standards for bilingual program evaluation design that is specified in the current regulations is "representativeness of evaluation findings [which means that] the evaluation results must be computed so that the conclusions apply to the persons, schools, or agencies served by the projects." Translating this standard into validity terminology, it specifies that evaluations must have high construct validities of student samples (and also of program settings).

To determine the construct validity of a study's sample of students, one should begin by clearly defining, in operational terms, the population to which the results will be generalized. Then biographic and demographic data should be collected on the sample students to determine how representative they are of the defined population. If they are not a good match, it may be necessary to redefine the population to which study findings might reasonably be generalized.

In bilingual education programs, high student mobility often degrades the representativeness of the sample (see Chapter 2). Some of the strategies proposed by Yap (1984) for resolving this problem include using tests with monthly or quarterly norms (or criterion-referenced tests), and using separate comparison standards for subgroups of project students based on length of time spent in the project.

To reduce the burden of testing, sometimes only a subgroup of students is selected for testing through either random, stratified, cluster, systematic, or multiple matrix sampling procedures (e.g., Molina & Shoemaker, 1973). In those situations, it is critical to determine the construct representativeness of the subsample to ensure that the results are generalizable to all students in the project. Without such assurance, it will not be possible to evaluate the extent to which the project sample represents its population.

Other factors that may reduce the sample construct validity are volunteerism and use of available groups (Cook & Campbell, 1979). In cases where students may choose to participate in a project or parents may volunteer to serve on committees and attend workshops, it will probably be inappropriate to generalize findings to non-volunteers. Similarly, threats to construct validity will arise if students in intact groups (e.g., students from one particular school) are selected for the evaluation sample while other "units" of the population served are excluded. Unless the selection process is random, the sample may not represent the target population adequately.

The AIR bilingual education evaluation (Danoff, 1978) provides a good illustration of this general problem. Since 26% of the Title VII group and 83% of the comparison group were monolingual English speakers, the samples were hardly representative of the population in need of bilingual education. Thus treatment-control comparisons did not really answer the research question of interest: Is bilingual education effective for Hispanic LEP children? The samples that were compared had low sample construct validity. (Note: because criteria for entry into bilingual programs sometimes result in the inclusion of monolingual English speakers [K. A. Baker, 1985, personal communication], low sample construct validity may be relatively commonplace.)

Settings. Bilingual programs are implemented in a variety of settings including bilingual centers, small classes, large classes using aides, and others. Evaluation findings may be affected by the setting; hence it is important to ensure that the operationalized setting matches the setting of interest. Suppose we wished to investigate the effectiveness of bilingual tutoring conducted in a bilingual center resource

room. If the tutorial program we studied was actually provided in the rear of the regular classroom instead, the theoretical construct--a typical bilingual center--would be inadequately operationalized. It would be misleading to generalize the results of the evaluation to the bilingual center setting. In other words, the evaluation's construct validity of settings would be low.

Threats to Construct Validity

Construct validity is important in bilingual program evaluation because determining the instructional strategy with the most significant outcomes for different types of students in various settings has been described as a primary goal of bilingual evaluation (Cummins, 1980; Hubert, 1982; Piper, 1984). If the construct validity of treatments, effects, samples, or settings is low, erroneous conclusions and inappropriate generalizations to theoretical constructs are likely to occur. Ensuring high construct validity of treatments, outcome measures, student samples, and settings is thus crucial for local program implementation because it enforces close adherence to program plans which, in turn, are usually based on sound theoretical justifications and empirical evidence.

In Cook and Campbell's (1979) treatment, 10 threats to construct validity were identified.

> They all have to do either with the operations failing to incorporate all the dimensions of the construct, which we might call "construct underrepresentation," or with the operations containing dimensions that are irrelevant to the target constructs, which we might call "surplus construct irrelevancies." (p. 64)

"Construct underrepresentation" and "surplus construct irrelevancies" correspond respectively to convergent and discriminant validities which were discussed previously.

Following are brief discussions of each of the 10 threats to construct validity as they relate to bilingual program evaluation. We have drawn heavily from Cook and Campbell (1979) in these discussions.

The first threat has been given the somewhat intimidating title of *inadequate pre-operational explication of constructs*. What this means is that careful thought should be given to operationalizing the constructs to be investigated. If the constructs are not adequately operationalized, the research will not provide a valid answer to the question the investigator wishes to explore. According to Judd and Kenny (1981), "hypotheses about the validity of an operationalization should be based on experience, convention, common sense, and prior research" (p. 25). A precise explication or redefinition of constructs, and sometimes further research, is necessary. In bilingual education evaluation, the linkages between operations and constructs are, unfortunately, seldom challenged and examined.

The *mono-operation bias* occurs when there is only one example of the treatment construct or a single measure for each of the outcome constructs. Construct validity is threatened in such instances because the single indicator may mis- or underrepresent the theoretical construct of interest. The solution is to employ multiple indicators. In the case of bilingual instruction, for example, educational achievement could be operationalized in a variety of ways: by performance on standardized achievement tests, by course grades, by time-on-task, or by grades on homework or project assignments. Construct validity is enhanced if two or more operations that represent the same construct show the same result. Construct validity would be enhanced, for example, if a student who did well on a math achievement test also received a high grade in math class. As Webb, Campbell, Schwartz, and Sechrest (1966) have argued, "if a proposition can survive the onslaught of a series of imperfect measures, with all their irrelevant error, confidence should be placed in it" (p. 3). In an attempt to reduce the mono-operational bias, Hazen (1980) proposes the use of multi-method research and evaluation in computer-assisted and computer-managed instruction to reduce measurement error and to determine the convergent validity of the effect construct. The five classes of measurement methods he identified are final examinations, attitude questionnaires, naturalistic observations, interviews, and archival data analysis.

Another related threat is the *mono-method bias* which refers to using the same method of administering treatments and the same means of recording responses for all the outcome measures. The method itself becomes an irrelevancy

which may influence the outcome measures. For example, bilingual instruction may be presented only orally without visual aides, while outcome measures may rely solely on multiple-choice, paper-and-pencil tests. If positive findings are observed, they could be attributed partially to the visual mode of presentation and/or to response bias in favor of multiple-choice test items. The obvious solution to this potential threat to construct validity is to vary treatment administration and response recording.

The next three threats all relate to treatment administration. *Hypothesis-guessing within experimental conditions* refers to the staff or students guessing what the evaluator hopes for and trying to please him/her. Classroom observation frequently encounters this threat because students and teachers may deviate from their normal behaviors when they are observed. The classic Hawthorne experiments are sometimes cited as another example of this threat to construct validity. In those experiments, employees reputedly increased their productivity in apparent response to managements' concern for their welfare rather than in response to improved lighting--although this account may be more folklore than fact (Parsons, 1974).

A closely related threat is *evaluation apprehension*. Teachers or students may be nervous when being evaluated by an outsider (supposedly an expert) and their nervousness may affect their behavior either positively or negatively. Another treatment-related threat is *experimenter expectancy*. An evaluator observing an experimental class may unconsciously rate the instruction more favorably than when observing a control class; or more coaching may be given to experimental than control students during testing. The use of non-stakeholders for data collection is an often-recommended approach for controlling this threat.

A threat which relates to program implementation is labeled *confounding constructs and levels of constructs*. When positive effects are not observed in a bilingual program, it could either be that the instructional method was not effective (construct) or that the strength and integrity of the treatment was insufficient to produce any effect (level of construct) (Sechrest et al., 1979). The best approach for determining whether the problem lies with the construct or with the level of construct is to measure the degree of program implementation through classroom ob-

servations, interviews, school records, checklists, staff reports, questionnaires, or attendance records.

Sometimes a school receives multiple funding to provide services to LEP students (e.g., migrant, Chapter 1, Title VII). If a student is served by more than just the bilingual program, outcome measures are confounded with the effects of the other program(s). This is the *interaction of different treatments* threat to construct validity. In addition to documenting the amount and type of service provided by each program, attempts should be made to separate the effects during data analysis and to acknowledge whatever contamination of treatment construct validity may remain.

Another threat to construct validity is the question of generalizing the treatment effects to other testing situations. If a pretest sensitizes the students to the subsequent treatment (Solomon, 1949), the outcomes cannot be generalized to situations where there is no pretest. For example, a bilingual computer-assisted instruction project may administer a computer literacy test and a rating scale measuring attitudes toward use of computers before the instruction is begun. Responding to these instruments may "tune students in" to the subsequent instruction. In an historical-record design, employing multiple measures collected before and after the treatment, there is always some question whether the results can be generalized to another sample not exposed to multiple testing. The concern for the extent to which testing is confounded with outcomes is the *interaction of testing with treatment* threat to construct validity. Careful analysis of the effect of testing in particular situations, and avoidance of potentially sensitizing items are two recommended precautions. If feasible, unobtrusive measures should be employed.

Sometimes a treatment may have positive or negative impacts on dependent variables other than the ones included in the evaluation plan. In bilingual education, for example, such outcomes may include the success of former participants in mainstream classes, their ability to secure jobs after graduation, their social interaction skills, the extent of their involvement in community activities, and so on (see Paulston, 1977, p. 100). If the experimenter wishes to consider these outcomes (and it may be important to do so), appropriate measures must be included in the study.

Without them, he/she will fall victim to the *restricted generalizability across constructs* threat. To minimize the effects of this threat, careful thought must be given to operationalizing the outcome construct.

More detail on all of these threats to construct validity is contained in Cook and Campbell (1979) from which much of the preceding discussion was adapted.

Mortality related to the treatment also constitutes a threat to construct validity because LEP students who drop out of a bilingual program because they (or their parents have) a negative attitude toward L1 instruction or feel that the program is ineffective may be systematically different from those who remain in the program. Evaluation findings based on students who remain in the program cannot, therefore, be generalized to the entire target population. This type of student attrition is considered a threat to construct validity because the apparent treatment effect may be no more that "selecting out those individuals who can potentially be affected" (Judd & Kenny, 1981, p. 37). According to Cook and Campbell (1979), the extent of the problem caused by differential mortality can be estimated by comparing the pretest scores and other background information of the dropouts with those of the students remaining in the program.

Internal Validity

After construct operationalization, the next step in the research process is to determine the causal relationship between the operationalized treatments and the operationalized outcomes or the *internal validity* of the research. A study is internally valid if it can demonstrate in a credible way that the obtained effects are, in fact, due to the treatment. In other words, the outcomes are "caused" by the treatment and not by other irrelevant influences. It should be noted that the causal link is between the operationalizations of the treatment and effect constructs and *not* between the constructs themselves. It is only by inference that we generalize the results to the constructs that the operationalized treatment and outcome represent. Such generalization will be inappropriate in the presence of low construct validity, as discussed above. Our internal validity concern, however, is with the validity of

the cause-effect relationship between the treatment *as it was implemented* and the outcome *as it was measured.*

A total of 13 threats have been identified that can jeopardize the internal validity of research (Cook & Campbell, 1979) by either inflating or deflating treatment-effect estimates. These threats can be grouped in two categories: those that are related to the adequacy of the research design and those that may occur even in the presence of an ideal research design. We will first describe the nine design-related threats.

Design-related threats to internal validity. The first design-related threat is *history* which refers to local events which occur during the treatment period and may affect the outcomes. This threat is most relevant to time series and other designs where there is no contemporaneous control or comparison group. With a contemporaneous comparison group it is possible to determine whether change resulted from the extraneous event or from the treatment. With no such control, it is not possible to make this distinction.

With the passage of time, not only may historical events exert a threat to internal validity; *maturation* may do so as well. Bilingual students grow older and perhaps wiser, and hence perform better or differently on the post-treatment measures than the baseline measures. For example, their attitude toward school and education in general and their cognitive problem-solving skills may change with age. Although both history and maturation cause real change and growth in the individuals, they are nevertheless not treatment-related and therefore are potential sources of bias.

Another source of bias which may be operative during the course of the program is differential *mortality*, or attrition. We discussed mortality earlier when we discussed the construct validity of student samples. Attrition is a threat to construct validity *if* the students' reasons for leaving the program are due to the characteristics of the treatment. There may, however, be non-treatment-related reasons for more or different students dropping out of one group (treatment or control) than another group. This type of attrition constitutes a threat to internal validity when

the treatment effect is estimated from the difference between the posttest scores of the treatment and control groups. Mortality is a major problem in bilingual education evaluations because of high student mobility.

The next two threats are related to measuring outcomes. *Testing* is a threat if the pretest has some carry-over effects on the performance of the subsequent posttest. The presence of this threat is most likely when the same test is given twice over a short period of time and students can recall the correct responses. In bilingual education evaluation, since accountability data are required annually to demonstrate program effects, and because student turnover is often high over the summer, students are likely to be tested on a fall-to-spring schedule rather than on the more preferable spring-to-spring schedule. Use of carefully equated alternate test forms or an annual testing schedule represent effective countermeasures to this threat.

A related threat to internal validity arises when different instruments are used to measure outcomes either across time or across groups. This *instrumentation* threat may arise when different tests, observers, or scorers are used, or when interviewers' proficiencies increase or decrease. Changes in outcome indices are likely to be confounded with changes in instrumentation in such circumstances. In addition, instrumentation bias may occur when a measure has floor or ceiling effects (Judd & Kenny, 1981). This problem is particularly relevant to bilingual education, where floor effects are often observed when pretesting LEP students (Baker & de Kanter, 1983). Use of out-of-level instruments may be effective in countering this threat, but only if the content of the out-of-level test affords a reasonable match to the material taught.

When students are selected for program participation on the basis of either high or low pretest scores, outcome measures will show change unrelated to any treatment. When posttested, such specialized groups will tend to score closer to the mean of the total group from which they were drawn than they did on the pretest, even in the absence of any treatment. This phenomenon is called *regression toward the mean* and is a threat to validity. Even if selected students are administered a separate pretest, there will still be some regression to the mean from pre- to post-

test. In bilingual education where the students served are typically those with the lowest scores on a language proficiency test, this "regression" may have a significant biasing effect unless the treatment and control groups are formed through a random assignment process. This threat to internal validity will be discussed again in Chapter 5.

An even more common threat to the internal validity of bilingual program evaluations is the effect of *selection*. This threat arises when there are systematic differences between treatment and comparison groups and no adequate statistical adjustment for the differences. Since, under current legislation, obtaining a randomized control group appears to be impossible in bilingual program evaluation, this selection bias is present to some extent whenever a non-equivalent comparison group is used as a basis for estimating treatment effect.

All the threats to internal validity discussed thus far can be controlled by random assignment of students to treatment and control groups. The assumption when we have random assignment is that any non-treatment influence that affects the treatment group will affect the control group with the *same* intensity and direction. For example, it is expected that the two groups will have the same amount of outside learning, maturation rates, dropout rates, pretest carry-over effects, instrumentation biases, and regression effects. For this reason, whatever biases may exist will cancel each other out, leaving treatment as the only independent variable acting on one group and not the other.

Of course, the same assumptions cannot be made if there is selection bias, i.e., the groups differ initially. Under these circumstances, selection may interact with the other threats to internal validity to produce differential history, maturation, mortality, testing, instrumentation, and regression between groups. These interactions create a whole new set of threats, known collectively as *interactions with selection*. One that is particularly important is the selection-with-regression interaction. This effect is frequently encountered when evaluators attempt to construct equivalent comparison groups by selecting apparently comparable students from non-equivalent, intact groups through a process of score-matching. If, for example, an evaluator found that the bottom 30% of the third graders at School A had ap-

proximately the same test-score distribution as the bottom 20% of the third graders at School B, the evaluator would be ill-advised to use the School-A subgroup as a control for the School-B subgroup. The school-B subgroup would show greater regression to the mean on retesting because it was further below the mean on the original testing (Thorndike, 1942). Thus the two groups that appeared equivalent would not really be so, and the selection-regression threat to internal validity would bias any evaluation that assumed they were.

Whenever a quasi-experimental design is used in a bilingual program evaluation, special attention should be paid to these interaction-with-selection threats.

The next threat to internal validity is *ambiguity about the direction of causal influence.* Such ambiguity often arises when correlational data are used to infer causality. An example is the correlation between LEP students' attitudes toward school and their academic achievement. It is never clear whether students' academic performance improves because they have more positive attitudes toward school or their attitudes improve because they are having greater academic success. One way to determine the direction of causality between variables is to conduct longitudinal research using structural equation modeling (Sorbom & Joreskog, 1981; Wets, Linn, & Joreskog, 1977) or cross-lagged panel designs (Campbell & Stanley, 1966). The latter designs, however, have been criticized for their conceptual and technical problems (Rogosa, 1980).

All the threats to internal validity described thus far are design-related threats--that is, they can be controlled through appropriate experimental design. The remaining four threats can occur even in randomized control group designs and represent unintended (and undesirable) effects of the evaluation itself. These threats can occur when a comparison group is employed and the treatment is being perceived as desirable.

Suppose a bilingual program is designed to demonstrate the effectiveness of an innovative reading program especially designed for LEP students. The evaluation of this program employs a comparison group that does not receive this treatment. If, however, teachers of the comparison group learn about the program and

feel it would be beneficial for their students, they may adopt at least some of the same methods and materials for their classes. In this way the planned difference between the treatments administered to the two groups is reduced. This threat to internal validity is labeled *diffusion or imitation of treatments.* A variant to the above scenario might find local administrators desiring to minimize the inequity between groups by providing the comparison group with other special services. This *compensatory equalization of treatment* imposes a threat to internal validity if the compensating services reduce the planned difference between groups.

The planned contrast can also be altered if the participants in the comparison group are disturbed by the fact that they are receiving the less desirable treatment. This knowledge of group membership *may* motivate the teacher and/or students in the comparison group to try harder or otherwise compensate for the "unfair" treatment. On the other hand, the comparison group may feel discouraged and resentful and may consequently lower its level of effort. Both the *compensating rivalry* (or John Henry effect, as it is more commonly called--see Saretsky, 1972) and *resentful demoralization* of the comparison group are threats to internal validity.

The four non-design-related threats to internal validity are likely to operate in bilingual program evaluations because the comparison group is usually selected within the same district as the project group, and bilingual teachers and aides from the same district often form a special interest group in which members are aware of each other's activities. In addition, the social-political environment, the lack of adequate teaching and learning materials, and the inexperienced comparison group teachers' need for assistance can all enhance the likelihood that these difficulties will be encountered in bilingual program evaluations. To account for the resulting plausible rival explanation for whatever posttreatment differences between groups are observed, it is necessary to define and monitor comparison group activities during the evaluation period (Chesterfield, Moll, & Perez, 1982; Cook & Campbell, 1979; Kerr, Kent, & Lam, 1985). Simply talking to or interviewing comparison group teachers, students, and/or other school staff can provide insights regarding the extent of the problems. An adequate description of both the experimental and comparison groups can also reveal other group differences or events that may dis-

tort the evaluation findings, e.g., unique local events, data collection procedures, changes in instrumentation, and other potential sources of bias.

In summary, the most effective way to control for design-related threats to internal validity is to employ randomized assignment of students to treatment and no-treatment conditions. As Cook and Campbell (1979) put it:

> When respondents are randomly assigned to treatment groups, each group is similarly constituted on the average (no selection, maturation, or selection-maturation problem). Each experiences the same testing conditions and research instruments (no testing or instrumentation problems). No deliberate selection is made of high and low scorers on any tests except under conditions where respondents are first matched according to, say, pretest scores and are then randomly assigned to treatment conditions (no statistical regression problem). Each group experiences the same global pattern of history (no history problem). And if there are treatment-related differences in who drops out of the experiment, this is interpretable as a consequence of the treatment. Thus, randomization takes care of many threats to internal validity. (p. 56).

Given that random assignment is an impossibility in bilingual program evaluation, the evaluator "has to systematically think through how each of the internal validity threats may have influenced the data. Then, the [evaluator] has to examine the data to test which relevant threats can be ruled out" (Cook & Campbell, 1979, p. 55).

Without randomization, some of the strategies which can be employed to reduce or account for threats to internal validity are:

1. To minimize differential history bias, the comparison group should be selected from classes in the same school or schools in the same neighborhood as the treatment group; and relevant "historical" events that occur during the time the program is being implemented should be recorded (e.g., teacher change).

2. To avoid the confounding effect of maturation, the inter-test interval should be reduced. Such a reduction, however, will increase the likelihood of the testing carry-over effect. A compromise, therefore, has to be made. Spring-to-spring instead of fall-to-spring testing has been recommended for bilingual program evaluation (Horst et al., 1980) and a twelve-month test interval is required by the current regulations for Title VII projects.

3. Some methods of reducing the testing threat are: using parallel forms of the test if available, testing only if it is necessary to address an evaluation question, and using unobtrusive measures if possible.

4. To measure and account for statistical regression artifacts, the sample and population means should be compared, and the test reliability examined.

5. To estimate and control for selection bias, sufficient demographic and biographic data from all project students (experimental and comparison) should be collected to provide a wide data base for determining group comparability, to statistically control for initial group differences, and to assess the plausibility of competing causes.

6. To determine the effects of mortality, attrition rates should be computed, and comparisons should be made between remaining and dropout students on their pretest scores and key background variables (Cook & Campbell, 19779).

7. To minimize instrumentation bias, the same or equivalent tests with high test-retest reliabilities should be used across time or groups. In addition, more than one observer, interviewer, or scorer should be employed to establish inter-rater reliabilities; and the same data collectors should be used throughout the evaluation.

As previously discussed, documentation of comparison group activities is essential to determining the extent of contamination by both the non-design-related

and the interactions-with-selection threats to the validity of the evaluation study. Devising appropriate preventive strategies to minimize the non-design-related threats (imitation of treatments, compensatory equalization, compensatory rivalry, and demoralization in groups receiving less desirable or no treatments) will require consideration of local conditions, human relationships, political factors, and the nature of the program. The methods of counteracting threats are usually project-specific, relying heavily on the evaluator's and project director's ingenuity. Some general strategies include isolating the treatment and comparison groups, educating administrators about research, and misinforming the groups (disguising the treatment).

In 1969, Campbell added "instability" to the list of threats to internal validity. This refers to drawing incorrect conclusions because of unreliable findings. This threat was later expanded to a separate validity category labeled "statistical conclusion validity" (Cook & Campbell, 1979).

Statistical Conclusion Validity

Statistical conclusion validity is defined as "the extent to which the research design is sufficiently precise or powerful to detect effects on the operationalized outcome should they exist" (Judd & Kenny, 1981, p. 20). It relates to the probability of incorrectly concluding that there was no treatment effect when, in fact, there was (Type II error).

The distinction between internal and conclusion validity is that the former is concerned with "sources of systematic bias" while the latter is concerned with "sources of random error and with the appropriate use of statistics and statistical tests" (Cook & Campbell, 1979, p. 80). A source of systematic bias (e.g., learning outside of the bilingual program) can effect the mean of an outcome (e.g., average group score in an oral English proficiency test). Random error does not have that effect. Its effect on a research study is to reduce the chances of obtaining statistically significant results. A parallel type of distinction can also be made between construct and statistical conclusion validities.

Statistical conclusion validity relates to the sensitivity of an evaluation, or its ability to detect true treatment effects of a given size. There are five factors relevant to such sensitivity:

- The power of the statistical test that is selected. Other things (see below) being equal, a more powerful test will detect smaller effects than a less powerful test.

- The probability level at which the evaluator is willing to accept that the observed effect was treatment-related rather than the result of chance. The more cautious the evaluator, the less likely it is that an effect of a given size will be "detected."

- The size of the sample. Small effects will be found statistically significant with larger sample sizes.

- The size of the estimated treatment effect. Larger effects will be more readily detectable than smaller effects.

- The homogeneity of within group performance. Chance differences between treatment and control groups will be high if performance variability within groups is large. It is useful to think of the difference between groups as a proportion of the within group variation. The larger the proportion, the more likely it is that the treatment effect will be detected.

With this background information, we hope that the following discussion of threats to statistical conclusion validity will be useful.

The first threat is *low statistical power.* If a statistical test with low power is used (e.g., nonparametric statistics are less powerful than parametric statistics), it will be necessary to increase the sample size, decrease the acceptable level of statistical significance, or select a more homogeneous group of students in order to detect a treatment effect that could have been detected through the use of a more powerful statistical test without such changes. In bilingual programs, the number of students

served per grade is usually small. Thus, if less powerful tests must be used, one possible solution would be to aggregate data across time or projects (see Chapter 7).

When parametric tests are used to increase statistical power, another potential threat to conclusion validity may arise. Unlike nonparametric statistics, the proper application of parametric tests rests on certain "strong" assumptions. The violation of such assumptions can prompt erroneous interpretations of the evaluation results. For example, in analysis of covariance, if the homogeneity-of-regression assumption is not met, the results of the analysis will be misleading. Another example is the use of students as the unit of analysis when the independence-of-observation assumption is violated, i.e., when students' performances are inter-related because of their sharing of the same teachers. It is fe to say that assumption testing has not been commonly practiced in bilingual program evaluation. This threat is referred to as *violated assumptions of statistical tests.*

Another source of statistical conclusion error is the practice of performing separate univariate statistical tests in evaluations using multiple outcome measures. This practice necessarily lowers the non-chance probability of the statistical indicators--a fact that is often unrecognized. This *fishing and the error rate problem* can inflate Type I error (concluding that treatment effects exist when they do not) and lead to "false positive" findings (obtaining "statistically significant" results by chance). Because of pressures on bilingual project directors to find positive results, it is not unlikely that evaluators will "fish around" the data.

As mentioned earlier, heterogeneity of within-group performance contributes to lowering the sensitivity of an evaluation. High within-group variation produces high standard errors of estimate (error variance) which, in turn, decrease the chance that between-group differences will be statistically significant. The next three factors to be discussed may threaten conclusion validity by increasing the heterogeneity of within-group performance.

If *the reliability of measures* is low, chance factors can contribute to the fluctuation of scores and thus increase the standard error of measurement. If change scores are used as measures of dependent variables, their reliability will be even

lower (Cronbach & Furby, 1970), although the significance of this fact has been challenged in the recent literature (Rogosa & Willett, 1983; Zimmerman & Williams, 1982). In any case, test reliability is a salient problem in bilingual program evaluation because commonly used assessment instruments (particularly language proficiency tests) are notorious for their low reliabilities (see Chapter 5).

In Chapter 2, it was pointed out that bilingual instruction varies from project to project because of the differences in students' educational needs. It may also change from occasion to occasion as a program adapts and improves. This low *reliability of treatment implementation* can increase student performance variability and hence error variance.

The last threat to conclusion validity that can inflate within-group performance variance is the *random heterogeneity of respondents.* A bilingual program often serves LEP students with diverse background characteristics (see Chapter 4). To the extent that some of these student characteristics (e.g., socioeconomic status and L1 language proficiency) correlate with outcomes (e.g., English language proficiency), error variance can be inflated if no control is exercised.

The sensitivity of an experiment is also affected by the magnitude of the treatment effects. Statistical significance can be obtained with a large group difference in means even if the error variance is large and sample size is small. The threat to conclusion validity that can reduce the size of treatment effects is called *random irrelevancies in the experimental setting.* If a bilingual class or tutorial session is being conducted in the library, a resource room, teacher's office, or in the hallway, students are easily distracted. Since different students are affected differently by different program settings, error variance may increase. In addition, error variance may be inflated if students are tested under similarly diverse conditions.

Wortman (1983) added "errors in coding and recording the data" to the list of threats to conclusion validity. Judging from the apparent quality of technical skills of bilingual education evaluators (see Chapter 2), such errors are almost surely present in evaluations. Sometimes such errors tend systematically to favor positive findings (Linn, 1982) and may reflect stakeholder bias (Tallmadge, 1985).

In summary, threats to statistical conclusion validity are probably abundant in bilingual program evaluations. As a step toward accurate assessment of treatment effects, these threats should be minimized. The following are some proposed strategies to deal with each of the eight threats to statistical conclusion validity in bilingual program evaluation.

1. **Low statistical power.** (a) aggregate data across time or projects to increase sample size; (b) use parametric statistics whenever statistical assumptions can be reasonably met; and (c) perform power analyses (Cohen, 1977) in the planning and analysis stages.

2. **Violated assumptions of statistical tests.** (a) be aware of the assumptions underlying each statistical test and, if possible, avoid violating them or minimize the extent of the violation; (b) use nonparametric statistical tests or alternative analysis strategies if the key assumptions are violated, (e.g., see Pedhazur, 1982).

3. **Fishing and the error rate problem.** (a) use procedures which appropriately adjust the significance level when performing multiple significance tests, e.g., adjusted t test, Scheffe's multiple comparison procedure; (b) perform multivariate instead of multiple univariate analyses; and (c) confine data analysis to testing a small number of hypotheses.

4. **The reliability of measures.** (a) add more items to the test; use more aggregated units such as classes; (b) use corrections for unreliability (attenuation); (c) select more reliable tests; use functional instead of grade-level testing (see Chapter 5); (d) write tests or surveys at a reading level appropriate for target LEP students; and (e) train observers or interviewers until they attain higher levels of reliability.

5. **The reliability of treatment implementation.** (a) try to standardize treatment implementation across occasions; (b) allow adequate planning time before program implementation, and (c) measure degrees of program im-

plementation (for both experimental and comparison groups) and use the measures in the data analysis or as additional information to help in the interpretation of results.

6. **Random heterogeneity of respondents.** (a) measure relevant student characteristics and use them as covariates or blocking variables in analysis of variance, or as explanatory variables in multiple regression procedures, or as additional information for explaining findings; (b) use a repeated-measures design if possible.

7. **Random irrelevancies in the experimental setting.** (a) eliminate distracting features in the setting; (b) increase the attractiveness of the treatment, i.e., make it more interesting to the students so as to get their attention; and (c) "measure the anticipated sources of extraneous variance [in the setting] which are common to all the treatment groups in as valid a fashion as possible in order to introduce the measures into the statistical analysis" (Cook & Campbell, 1979, p. 44).

8. **Errors in coding and recording the data.** (a) impose data quality control procedures such as random checking; (b) train observers, interviewers, and scorers to establish high inter-rater reliability; and (c) develop a systematic data management system (see, for example, Consalvo & Orlandi, 1983; Hoover & Kamm, 1981).

The strategies presented above are useful suggestions for minimizing the various threats to statistical conclusion validity. Another useful, but quite different principle for increasing this type of validity is to avoid drawing inferences *solely* from quantitative data (Balasubramonian, 1983). Gathering multiple outcome measures that include data from qualitative, naturalistic evaluations is critical to drawing valid conclusions uncontaminated by measurement errors (Chesterfield et al., 1982; Lee, 1985). If the qualitative data "are contrary to the quantitative results, the quantitative results should be regarded as suspect" (Campbell, 1979, p. 52).

Next we discuss issues concerning the generalizability of a bilingual program treatment effects to other bilingual programs, outcome measures, LEP student populations, and settings.

External Validity

External validity relates to the generalizability of findings to treatments, students, outcomes, and settings other than those specifically studied. An educational evaluation would have high external validity if its conclusions applied to LEP students with diverse ethnic, linguistic, educational, cultural, and socioeconomic backgrounds--or if there were evidence that what was learned about teaching English language skills applied as well to math, science, and social studies--or if what was observed in the classroom was the same as what occurred when the intervention was implemented in the resource room.

Construct and external validities are similar in that they both involve generalization. However, construct validity is concerned with generalizations from observed entities and events to theoretical constructs of treatments, outcomes, persons, and settings. External validity, on the other hand, is concerned with generalizing from specific observed entities and events to other, entitites and events of interest.

The external validity of an evaluation finding may be largely unknown. A particular study, for example, might allow us to conclude that treatment A is effective in reducing the gap on measure B for students C in setting D. Whether it would be effective if other measures were used, or if other students were served, or if it were implemented in other settings can only be determined empirically. That is the reason behind some researchers' (e.g., Shadish, Cook, & Houts, 1986) strong recommendation that evaluations employ multiple operationalizations of treatment constructs, multiple measures, multiple types of students, multiple settings, etc.

Evaluation findings with low or unknown external validity are of limited usefulness. We may speculate that they will hold true for similar measures, similar students and similar settings but, without empirical support for these hypotheses, we

could seriously and expensively over- or underestimate the generalizability of our findings.

The three threats to external validity are conceptualized as interactions of the treatment with student, setting, and "history" (time) variables. Each is illustrated below.

1. *Interaction of the treatment with the students served.* A particular instructional strategy (e.g., immersion in a second language) may be effective for some students (e.g., language-majority students) but ineffective for others (e.g., language-minority students).

2. *Interaction of the treatment with the setting.* A particular instructional strategy may be effective in one setting (e.g., a small group) but ineffective in another (a whole classroom).

3. *Interaction of the treatment with "history".* An unusual event (e.g., a visit by the Secretary of Education) may have occurred and acted as a catalyst to enhance student learning. The effect might not be observed without that specific historical event.

One way to estimate external validity is by using "theory that defines the relationships between constructs, theory validated by prior research, experience, and common sense" (Judd & Kenny, 1981, p. 40). Knowing the similarity in backgrounds between Vietnamese and Cambodian refugees, for example, one may predict that the effect of a bilingual program should be similar for the two groups of students. On the other hand, generalizing from language-majority to language-minority students involves greater hazards.

The best method for assessing external validity is to conduct large-scale evaluations in which all types of students are exposed to all types of treatments in all types of settings. Judd & Kenny (1981) refer to this method as turning external validity concerns into "many simultaneous issues of construct validity" (p. 41). To the extent that the findings are consistent across the entire population (e.g., all

bilingual programs, all academic and affective outcomes, all language minority students in the U.S., and all public schools), external validity is assured. However, this approach is both costly and impractical. In Baker and de Kanter's (1981) review of bilingual education evaluations, only two of the more than 300 studies reviewed attempted national generalizability using sampling procedures.

High external validity can also be acquired by replicating evaluation studies, varying either treatments, outcomes, students, or settings. For example, if an evaluation studying the effects of ESL instruction on the reading comprehension of third-grade Hispanic students were to be replicated with third-grade Vietnamese students, and then with fourth-grade Chinese students, and so on, generalizability of the treatment effects to different groups of LEP students could be examined. In each of these studies, the concern is whether the student sample represents the population of interest (e.g., third-grade Vietnamese students); a concern for construct rather than external validity. If the findings are consistent across different populations of students or treatments, the plausibility of additional, untested generalizations is enhanced and external validity is increased.

A more practical alternative for bilingual education is to conduct syntheses of published studies. Such a synthesis was attempted by the National Center for Bilingual Research using a meta-analytic approach (Okada et al., 1982, 1983). Unfortunately, as was discussed in Chapter 2, it failed because of the poor quality of the research and evaluation studies available for analysis. Nevertheless, this kind of research should be repeated as soon as local evalution and reporting practices have improved. At the same time, research designed to study the differential effects of different bilingual instructional approaches should be encouraged. It is only through this collective effort that the external validity of bilingual education research and evaluation can be increased.

Controlling for threats to external validity has a slightly different meaning than controlling for threats to other validities. While eliminating threats to external validity can enhance generalizability to other treatments, outcomes, students, or settings, knowledge about the existence of these threats is useful in its own right. For example, it is just as useful to know that a treatment which works for one population

does not work for another as it would be to know that it worked for both. That is the reason why we recommend conducting research to determine the match between program types and student types in various settings. The key point is for policy makers to devote more attention to examining external validity, whether the purpose is to increase generalizability or to define its limits.

In 1978, Cooper expressed his pessimism regarding generalizability in bilingual program evaluation:

> ...it is probably not an exaggeration to claim that each of the 400 current local projects of the Bilingual Education program is unique with respect to the sociolinguistic and educational context in which it operates. Thus, we cannot be sure that a program which works well in one context will work well in another. (p. 79)

We believe the picture is not as grim today as Cooper painted it nine years ago. Great diversity, however, will always be a feature of bilingual education, and careful research, in addition to standardized local evaluations, will be required to determine just what treatment-by-ethnicity-by-setting interactions are significant, and where we can generalize across these constructs.

Thus far we have described the four types of validities and how they relate to bilingual program evaluation. Next we discuss their relationships and priorities in the evaluation of bilingual programs.

Relationships Among and Priorities of Validities

The distinction among validity types can be arbitrary at times. For example, the differences between internal and construct validities, and external and construct validities are not always unambiguous. Some threats could be classified under more than one type of validity, depending on interpretation. Two examples are mortality and the treatment-with mortality interaction, both of which can be regarded as threats either to internal or to construct validity. Their biasing effects on the outcome measures can be interpreted as being due to a confounding of the effects of competing causes (a threat to internal validity), or these effects may be due to the

fact that the remaining sample is no longer representative of the population of target students (threats to construct validity). Another example is Wortman's (1983) dissatisfaction with the listing of "reliability of treatment implementation" as a threat to conclusion validity. In his opinion, it is more appropriately a threat to construct validity.

The ideal in any evaluation or research study is to maximize all four kinds of validity. In practice, however, this may not be possible. A procedure used to enhance one type of validity may diminish another type. For example, including different types of bilingual programs in an evaluation will improve generalizability across program types. But the heterogeneity of the resulting sample may at the same time increase unexplained variation (error variance) in the outcome measures, thus reducing conclusion validity. Other relationships between validity types, including the inverse relationship between internal and construct validities, and that between internal and conclusion validities, are discussed by Cook and Campbell (1979, p. 82) and by Judd and Kenny (1981, p. 42). Here it is sufficient to note that, given these tradeoffs between one kind of validity and another, priority among validity types should be established when planning an evaluation or research study. It may also be necessary, however, to modify desired priorities because of the restrictions imposed by practical concerns. If the goal were to maximize internal validity, for example, there would be a conflict between the desire to implement a true experiment and the legislative prohibition of withholding services from needy students. The internal-validity goal would have to be compromised. Given the various compromises that might be required, we next discuss what the priorities may be for the various stakeholders in bilingual program evaluation.

For years, policy makers in bilingual education have been activity seeking an answer to the question, "does bilingual education work?" or more specifically, "how much of the cognitive growth observed in bilingual program participants can be attributed to the bilingual program itself?" The increasingly stringent evaluation requirements spelled out in the bilingual education legislation and regulations and the increasing number of Federally funded evaluation studies are two indicators of this concern. The question, however, is a simplistic one that has internal validity as its major focus. It should be noted that obtaining data with high internal validity is

the ultimate goal of *research*, which is conclusion-oriented, and not necessarily that of *evaluation*, which is decision-oriented and situation-specific (Cronbach & Suppes, 1969).

While research is involved in seeking to confirm the credibility of some hypothesis (e.g., bilingual instruction increases academic achievement of LEP students), evaluation is aimed at gathering information for judging the merits of a project in a particular setting at a specific time (Burry, 1981, 1982) and for making decisions to terminate, modify, or continue the program. Given this distinction, national or large-scale evaluations are, in effect, research efforts (Gold, 1981).

Local evaluations are generally interested in determining how well the project students are performing and how the program can be improved to enhance their achievement. Formative evaluations which provide periodic feedback to project staff about program operations and suggestions for improvement are just as desirable to local project implementors as summative evaluations which indicate to what extent the program has enhanced the cognitive achievement of the students. Their concern for student progress is not coupled with questions about whether it is due exclusively to the program (high internal validity) or to some other factors. In that regard, internal validity is not as crucial to them as construct validity which addresses treatment, student, and setting definitions. Imposing restrictions on program design to assure high internal and conclusion validities may in some way impede services for the target students. Factors other than student achievement, such as program impact on the schools and community, are also being considered in judging the merits and value of the program.

In recent years, a number of bilingual educators and researchers have criticized the utility of attempts by the Federal government to assess the overall impact of a program employing many different strategies, implemented in varied settings, and designed to meet the educational needs of sociolinguistically diverse target populations (e.g., Cummins, 1981; Hubert, 1982; Piper, 1984). A more worthwhile approach, they claim, would be to determine the differential effects of specific programs on different groups of LEP students in various settings. In other

words, construct and external validities should be stressed, instead of internal validity.

The approach these authors have proposed is in agreement with the cautions urged by Cronbach and his associates (1980) who wrote: "external validity--that is, validity of inferences that go beyond the data--is the crux; increasing internal validity by elegant design often reduces relevance" (p. 7). Judd & Kenny (1981), on the other hand, emphasize the importance of the construct validities of samples and settings in field research because, "the purpose of such research is to gain knowledge about an effect in a specific setting for a given population rather than to gain more basic theoretical knowledge of causal relationships in the abstract" (p. 44). The Significant Bilingual Instructional Features study (Fisher, 1983; Tikunoff, 1984), which had the goal of identifying significant attributes of successful bilingual classrooms using ethnographic research techniques, is an example of research concerned with construct validity.

Conclusion validity in applied settings should also be emphasized, according to Judd & Kenny (1981) "because of the number of studies that have found little or no effects for large social programs" (p. 44). This recommendation is clearly applicable to bilingual education where small treatment effects are commonly observed and expected. However, the enhancement of conclusion validity should not be at the expense of services for project students. For example, heterogeneity of treatment, although imposing a threat to construct validity, should nevertheless be allowed because of its beneficial effects on learning.

When these various recommendations are combined with the well documented difficulties in gaining high internal validity, it seems clear that local bilingual program evaluations should seek to achieve respectable levels of conclusion and construct validities, taking into consideration the conflicts between them. This is not to say we should abandon internal validity in local evaluations. On the contrary, efforts should be made to control for all design-unrelated threats to internal validity, and to examine ways to rule out or at least to document the effects of design-related threats. Although external validity is also desirable, it is solely the concern of research efforts that try to generalize conclusions beyond the specific entitites and

events that were studied. It is by the accumulation of findings from adequate local evaluations and well planned research studies that we can begin to address external validity. It should not be a concern for local evaluations (Popham, 1975; Rose & Nyre, 1977; Weiss, 1972).

Given these various considerations, the four types of validity should be prioritized as follows at the local level: (a) construct, (b) conclusion, (c) internal, and (d) external. For large-scale evaluations and research studies, a more appropriate ordering would be: (a) internal, (b) construct, (c) external, and (d) conclusion. While these orderings may represent slight departures from tradition, we believe that in bilingual education particularly both local and national evaluation efforts should expend relatively more energy than they usually do attending to construct and conclusion validities since both are critical if we are to learn about effective ways to help language-minority students attain an adequate education in the American school system.

4. TREATMENT, STUDENT, AND SETTING VARIABLES IN BILINGUAL EDUCATION EVALUATION

There is great diversity in bilingual education programs, their settings, and the students they serve. Students with diverse ethnic, linguistic, socioeconomic, and educational backgrounds are served at all grade levels, in schools with dissimilar student-body compositions, in many different types of communities. To complicate matters further, different instructional strategies are implemented by staff with a wide range of professional and linguistic competencies in programs of varying intensities and durations. All of these factors are thought to interact in complex ways so that there can be no simple answer to the question, "How well does bilingual education work?" It would be more appropriate to ask, "How effective are different bilingual education treatments for different types of students in different settings?"

If indeed the issue of effectiveness is as complex as is suggested by the preceding question (and there is some evidence that it is), then ideally all relevant characteristics of students, settings, and treatments would be carefully documented as an integral part of any bilingual education program. To fail to do so would run the risk of obscuring educationally significant relationships whenever comparisons are made between programs or when data are pooled across different types of students, treatments, and/or settings. In the real world, however, it is rarely possible to predict and document every relevant variable, and there is no research that conclusively demonstrates interactions between student, setting, and instructional variables.

In this chapter, we discuss treatment, student, and setting variables that have been identified as potentially interactive on the basis of either theoretical formulations or empirical findings. Treatment variables are discussed under the four headings: instruction, materials, staff, and parent/community involvement. Family background, prior educational experiences, attitudes, and initial skills are among the student variables discussed. The setting variables include school and community characteristics. By briefly summarizing the relevant literature, we hope to make clear the importance of documenting as many of these program characteristics as

possible, both to facilitate meaningful comparisons among (and aggregations across) comparable programs, and to discourage inappropriate comparisons and aggregations. We begin with a discussion of treatment characteristics.

Treatment Characteristics

Effectiveness. Although there have been numerous attempts to investigate and compare the effectiveness of different bilingual education instructional approaches, the findings have been inconclusive. Tikunoff (1985) reports that effective bilingual teachers use English about two-thirds of the time for basic skills instruction, while Wong Fillmore (1983) and Legarreta-Marcaida suggest an even balance between L1 and L2 is more effective. Of the 35 studies they reviewed, Cohen and Laosa (1976) report that some found that the exclusive use of L1 for instruction produced the best results, others indicated that the sole use of L2 produced the best results, and still others concluded that L1 and L2 could be used simultaneously with good results. Cohen and Laosa attribute these apparently contradictory findings to (a) differences in the educational treatments investigated; (b) characteristics of students in the samples; (c) contextual characteristics; (d) the research design, methodology, and instrumentation of the studies; and (e) the interactions among these various factors. Tikunoff (1985) attributes differences to (a) the L1 or L2 proficiency of the LEP student population, (b) the percentage of the class that is LEP, (c) the number of languages represented by the LEP students in a class, (d) the time of year, (e) instructional objectives, and (f) content areas.

A different explanation was proposed by Lambert (1975). He notes that numerous studies of immigrant and language-minority students who were learning a second language showed that these students exhibited poor academic performance (Darcy, 1953; Diebold, 1968; Jensen, 1962a, 1962b; Lambert & Tucker, 1972; Macnamara, 1966; Vildomec, 1963) while other studies consistently found cognitive advantages to be associated with second-language acquisition (Albert & Obler, 1979; Bain, 1975; Balkan, 1970; Cummins & Gulutsan, 1974; Cummins & Mulcahy, 1978; Duncan & De Avila, 1979; Genesee, Tucker & Lambert, 1975; Hakuta & Diaz, 1983; Kessler & Quinn, 1980; Liedke & Nelson, 1968; Mohanty, 1982; Peal & Lambert, 1962; Scott, 1973). To explain these conflicting findings, Lambert suggests that

there are two types of bilingual programs: "additive" and "subtractive." Subtractive bilingualism occurs when L1 is replaced by a dominant and higher status L2; additive bilingualism occurs when L1 is maintained while L2 is learned. A student who has learned another language under the latter conditions is less likely to attain native-like proficiency in either L1 or L2. On the other hand, a majority of the studies that found bilingualism to be associated with cognitive advantages studied children who acquired L2 through the additive process. Lambert suggests that the subtractive process is the cause of the negative effects observed in the earlier studies.

An alternative explanation for the conflicting findings is that students must acquire a certain level of proficiency in both languages to avoid negative affects and a still higher level before a beneficial effect appears. Cummins (1983) has also theorized that while minority-language students may within a year or two acquire English proficiency in context-embedded, face-to-face interactions (basic interpersonal communicative skills or BICS), several more years of bilingual education will be required before those same students acquire the level of English proficiency necessary for complex, context-free academic tasks (cognitive/academic language proficiency or CALP). Therefore, different definitions or assessments of English proficiency may also contribute to different research results.

Krashen (1981) suggests that growth in language is stimulated by linguistic input that is just beyond the learner's understanding, but which the learner can make comprehensible by using non-linguistic clues. If the input is not geared to a level that the student can make sense of, or if the input is at a level already achieved by the student, no language growth will occur.

Bilingual education treatments can be described as including four main components: (a) instruction, (b) materials development, (c) staffing, and (d) community involvement (Alkin et al., 1974). On the following pages we discuss each of these four components. Other components that have been considered integral to some bilingual programs include management improvement and evaluation improvement. For reasons of parsimony, however, we have decided to exclude these components from separate consideration.

Instruction. It has become common practice to categorize programs serving LEP students into four instructional types: (a) early-exit transitional, (b) late-exit transitional, (c) immersion, and (d) English-as-a-second-language.[6] In addition, the term, "submersion" is frequently used to denote the absence of any special treatment. Although the distinctions among these instructional approaches are not always clear-cut, we shall begin this discussion of treatment characteristics by describing each of the four major bilingual program types.

Early-exit transitional bilingual education programs are the most frequently implemented in the United States (Gonzalez, 1979). Native language instruction is used, but only until students are proficient enough in English to benefit from all-English instruction. The main goal of the early-exit model is to "transition" LEP students into an all-English curriculum as quickly as possible. Federal guidelines and some state guidelines regulate the length of time students can remain in Federally or state-funded transitional programs.

Although LEP students are initially taught in their primary language, L1 instruction is used only to facilitate the acquisition of English language skills and to prevent students from falling behind in other content areas while they learn English. The curriculum in early-exit programs is not designed to develop or maintain students' primary language. Early-exit programs reduce the amount of L1 instruction and increase the amount of L2 instruction over time until the entire curriculum is taught in English.

Early-exit transitional bilingual programs vary in the degree to which L1 and L2 are developed. At one extreme are programs that develop comprehension and verbal skills in both the primary language and English, but develop literacy skills only in English. At the other end of the continuum are programs which try to develop comprehension, verbal, and literacy skills in both L1 and L2 concurrently or consecutively.

---

6. Although two of these program types may involve no instruction in L1, we refer to all four as bilingual programs throughout this report.

Late-exit transitional programs (also referred to as developmental programs in Federal legislation) provide instruction in both the students' native language and in English, and continue to use both languages for the duration of the program. The goal of late-exit programs is to enable students to develop equal proficiency and competence in their primary language and in English. Unlike other bilingual models, late-exit programs try to sustain L1 and develop literacy skills in both L1 and L2. Skills in understanding, speaking, reading, and writing in L1 and L2 are developed concurrently or consecutively.

At the elementary level, most instruction initially occurs in L1, and literacy skills are usually developed in L1 before English literacy is taught. Instruction in the primary language decreases over time as instruction in English increases, until the two languages are used equally. Both L1 and L2 are used for the duration of the program in some or all subject areas, i.e., math, science, and social studies (Dominguez, Tunmer, & Jackson, 1980).

The *immersion* model has been widely used in Canada for many years (Genesee, 1978, 1984; Genesee, Polich, & Stanley, 1977; Lambert & Tucker, 1972; Swain, 1980). In an immersion program, the instructor, although bilingual, usually speaks in L2. Students, however, are permitted to speak to the teacher in their native language if necessary. Subject matter instruction is conducted in L2 from the beginning, and the curriculum is st ured so that it does not assume prior knowledge of L2 (i.e., L2 is "sheltered" with vocabulary developed simultaneously with subject matter content). Although there is variation among programs, all immersion programs have one essential characteristic: L2 is used both as the target language and as the medium of instruction in other academic subjects.

Immersion programs are not always total; partial immersion programs also exist, although primarily in Canada (California State Department of Education, 1974; Genesee, 1984). In the partial program model, L2 is used for instruction from the beginning but L1 instruction is introduced after students have been in the program several years (Genesee, 1978; Genesee & Lambert, 1983; Lambert & Tucker, 1972). The amount of time L1 is used for instruction may vary from 20% of the time to as much as 60% (Genesee, 1978; Morrison et al., 1979).

English as a second language (ESL) is usually a component of an early-exit (transitional) or late-exit (developmental) bilingual program; however, it may also be provided by itself as a "pullout" program (Ovando & Collier, 1985). ESL instruction may itself use an immersion or sheltered English strategy. The primary objective of ESL instruction is to provide students with the English language skills they need to communicate with teachers and other students, and to enable them to benefit from instruction in English. In a typical ESL program, students receive subject-matter instruction in regular, English-only (mainstream) classrooms, but are "pulled-out" for special instruction in English (usually at times when non-academic subjects are taught). ESL instruction varies in duration from 20 minutes to an hour per day, depending upon school resources. Students usually remain in the program for one to three years depending upon how quickly they achieve proficiency in English (Schinke-Llano, 1984).

While the labels given to the four program types just described may provide a convenient shorthand terminology, they are insufficient to characterize the instructional strategy actually employed. Other, more explicit classification schemes have been proposed which can offer more consistent and informative descriptions of bilingual programs. Dominguez, et al. (1980) say that bilingual education has three components: (a) the percentage of instructional time devoted to L1 language arts, (b) the percentage of instructional content areas taught in L1, and (c) the grade levels at which instruction in L1 is provided. The U.C.L.A. Center for the study of Evaluation (undated) suggests that a description of a bilingual program should include: (a) distribution of instructional time between L1 and L2, (b) kinds of instructional activities conducted in each language, (c) length of time students remain in the program, and (d) assessment categories of linguistic competence. Descriptions and/or classifications such as these can provide a much clearer picture of bilingual programs than more general categories such as "early-exit transitional."

The instructional component of bilingual programs can be described in even further detail. Ovando and Collier (1985), for example, discuss patterns of language use and classroom organization. In *concurrent teaching*, the teacher may use L1 and L2 interchangeably for content area instruction, or two teachers may team teach one lesson and each use a different language. The *preview-review* design is used

primarily in team-teaching situations. One teacher introduces a lesson in one language while the lesson itself is presented by the second teacher in the other language. Both languages are used concurrently for the review and reinforcement of the lesson. The *alternate-language* design separates the two languages completely. Most bilingual classrooms employing this design will use one language for instruction in the morning, and the other language for afternoon instruction. Some classrooms may alternate the language of instruction by subject area (some subjects in L1 and others taught in L2) while others may employ the two languages on alternate days.

Alex Law, as quoted by the Center for the Study of Evaluation (undated) has added to these categories *translation* (lessons are presented in English, then translated to a second language), *language-other-than-English immersion* (English oral language skills are developed, but a language other than English is used for academic instruction), and *eclectic* (combining two or more of the other approaches).

In addition, instruction in bilingual classrooms may be provided by one teacher, a team of teachers, one teacher and one aide, or one teacher and several aides (Ovando & Collier, 1985). The length of time that aides are assigned to a classroom varies, as do the duties assigned to them. Some aides may provide instruction, particularly if they are proficient in L1 and the teacher is not, while others perform only clerical tasks. In some classrooms, aides may work primarily with one group of students (e.g., non-English-speakers); in other classrooms, the teacher and aide(s) work alternately with small groups of students. A resource teacher may also be available to provide additional instructional support.

The grouping of students also varies depending upon the instructional approach. In some programs, students receive part of their instruction each day in mainstream classrooms and part in bilingual classrooms. Other programs provide a comprehensive, full-day program with a bilingual teacher or a monolingual teacher with a bilingual aide in a self-contained classroom. Programs using an ESL approach usually pull students out of regular classes for one or two periods of ESL instruction with a specially trained teacher.

Materials. The effectiveness of bilingual programs is significantly affected by the presence of materials appropriate for second-language learners. In particular, primary-language reading materials are needed by transitional programs to conduct subject-matter instruction and to promote reading at all grade levels, which many theorists believe is important to LEP students' academic achievement (Rosier & Holm, 1980; Santiago & de Guzman, 1977; Thonis, 1976, 1980, 1981). Since reading materials in L1 are seldom present in low-income homes or in community libraries, it becomes the responsibility of the school to have such materials "to extend opportunities for growth in reading and thinking skills" (Cummins, 1981, p. 176).

When Title VII projects were first implemented, there were few appropriate instructional materials available for use in bilingual classrooms. Since then monies have been made available to regional educational laboratories and other agencies with expertise in materials development, and bilingual projects have reduced their involvement in development activities. Earlier, however, materials development was an important component of most bilingual projects. Even today, appropriate materials may be difficult or impossible to find for some linguistic groups. In such situations, materials development continues to be an important program activity.

Even when suitable materials for bilingual instruction can be acquired, programs may not have enough of them to meet the needs of participating students (due to lack of funds, reluctance of administrators to purchase materials for bilingual classrooms, or lack of information about their availability). Several studies have found that a shortage of adequate materials hampers program implementation (Berman & Pauly, 1975; Charters & Pellegrin, 1973; Crowther, 1972; L. Downey Research Associates, 1975; Gross, Giacquinta, & Bernstein, 1971). Whether or not a project has materials development as one of its components, the availability and appropriateness of materials used in bilingual program classrooms should be reported since these materials can influence the success of the program. If materials development *is* a program component, the quality and appropriateness of materials should be assessed as one of the program outcomes, and also as a moderating variable which may limit the effectiveness of the instructional treatment.

Staff. Studies indicate that the ability of a teacher to speak the primary language of LEP students with native or near-native proficiency has a positive impact on both primary language development and on second language acquisition (Carrasco, 1981; Cazden, 1985; Merino, Politzer, & Ramirez, 1979; Penaloza-Stromquist, 1980; Ramirez, 1978). Students' language learning also appears to be affected by the acceptance and sensitivity of teachers to the varieties of L1 the students speak (Adams & Frith, 1979; Legarreta-Marcaida, 1981; Merino, et al., 1979; Penaloza-Stromquist, 1980; Rosier & Holm, 1980). Some research has found that a teacher's knowledge of second language acquisition and primary language development processes has a beneficial impact on English acquisition and primary-language development by linguistic minority students (Penaloza-Stromquist, 1980; Ramirez & Stromquist, 1979; Rodriguez, 1980; Thonis, 1976, 1981). Other studies indicate that teachers mediate effective instruction for LEP students by using L1 and L2 for instruction (alternating languages whenever necessary to ensure comprehension), and integrating English language development with academic skills development (Tikunoff, 1982, 1983; Tikunoff et al., 1981). Thus, the ethnic characteristics, language abilities, academic qualifications, and previous experience of staff are an important part of any bilingual program description.

The hiring of teachers who are qualified and trained to teach in bilingual classrooms has been a continual problem for school administrators since Title VII projects were first funded (Berman, McLaughlin, Bass, Pauly, & Zellman, 1977; Kaskowitz, Binkley, & Johnson, 1981; Oxford et al., 1981). For this reason, staff development continues to play an important role in the provision of effective bilingual instruction. In addition to offering teachers and aides the knowledge and skills they need to work with LEP students, staff development is frequently used to orient staff to the components of a specific bilingual program design. Even staff with experience in bilingual classrooms will need pre-service and in-service training when a new bilingual program is implemented.

Because numerous demands are made on teachers' time and energy, it is important to the effectiveness of staff development activities that administrators encourage teachers' participation and make needed resources available to them (Cole, 1971; Hamingson, 1973; Miller & Dhand, 1973; Shipman, 1974). Research indicates

that teachers who both participate in pre-service and in-service training *and* receive instructional materials implement programs more effectively than teachers who only receive instructional materials (Hess & Buckholdt, 1974; Solomon, Ferritor, Hearn, & Myers, undated). Opportunities for the instructional staff to discuss implementation problems and obtain feedback from others also improves implementation (Berman & Pauly, 1975; Center for Educational Field Studies, 1970; Charters & Pellegrin, 1973; L. Downey Research Associates, 1975; Gross et al., 1971; House, 1975). The combination of staff training and frequent meetings has a beneficial impact on success and fidelity of implementation and student learning (Berman & Pauly, 1975). Pre-service training and the provision of model units and demonstration lessons appears to be particularly useful to teachers (Cole, 1971; Crowther, 1972; Hestand, 1973).

A review of related research clearly indicates the importance of teacher training to program implementation. Generally, evaluators have been satisfied with documenting activities and attendance and have failed to examine (a) whether the sponsored activities have met the needs of teachers and their students, (b) whether they have helped staff teach more effectively, (c) whether they have helped staff resolve implementation problems, and (d) whether continual, internal evaluation of activities has been conducted and follow-up assistance has been provided to staff needing or requesting additional support.

Hall & Louchs (1978) have developed a "Stages of Concern" model which can be used to diagnose group and individual needs of teachers who are attempting to implement new teaching practices or a new instructional program. The model can be used to plan appropriate staff development activities or to evaluate whether staff development activities met the concerns and needs of most instructional staff. According to the model, staff implementing a new program or method will progress through the following stages. *Awareness*--staff indica[illegible] little concern or involvement. *Informational*--staff are generally aware and interested in learning more. *Personal*--staff are concerned about their roles and adequacy in the new program. *Management*--staff are concerned about efficiency, organization, and scheduling demands. *Consequence*--staff are concerned about the impact of the program on students. *Collaboration*--staff focus on working with others involved in the program.

Refocusing--staff focus on maximizing benefits, including changing or replacing the program. According to the authors' research, staff development activities at an inappropriate level will be perceived as useless and will not have an impact on staff knowledge or behavior.

Parent/community involvement. Since passage of the 1978 amendments to the Bilingual Education Act, bilingual projects have been required to involve parents and community members in the development of funding applications and in the implementation process. The impetus for involving parents comes from two sources: the advocacy of special interest groups, and research documenting the impact of parent involvement on program success. During recent years, there has been increased pressure from various ethnic and community groups to increase parental involvement in the schools. Legislators have responded by establishing a formal role for parents in the planning and implementation of Federally funded programs. Schools are required to establish parent advisory groups that meet to discuss educational issues and make recommendations to school and program administrators. Parents can participate in training activities sponsored by bilingual projects. The most common form of participation is for parents to volunteer services to help with extracurricular, social, or fund-raising activities. Some parents also serve as classroom aides or participate in evaluation activities.

Research indicates that parents can play an important role in the academic survival and success of their children. A 1984 study by Crespo and Louque indicates that parent involvement in school matters plays a crucial role in preventing Hispanic students from dropping out of school. Fantini (1970), Gordon (1978), Levin (1970), Schimmel and Fisher (1977), and Stearns, Peterson, Robinson & Rosenfeld (1973) report that school programs with involved parents and community members reflect community interests and, consequently, are more likely to achieve program goals. Parent and community involvement have a positive effect on a child's learning and school socialization according to Henderson (1981) who also reports that parent involvement in almost any form has a beneficial impact on students' achievement. The amount of impact varies in direct proportion to the extent of parent involvement in decision-making, tutoring, observing, and/or classroom management.

The critical factor affecting the impact of parent involvement is that it be well-planned, comprehensive, and long-term. Parental involvement is an indicator of parental interest in their children's education and is mediated by the development of attitudes conducive to achievement (Henderson, 1981). Since students whose parents are involved in school matters tend to make the greatest academic gains, community involvement activities sponsored by a bilingual program could significantly affect not only parental behavior, attitudes, and (for immigrant parents) familiarity with the U.S. educational system, but also learning outcomes.

Table 2 lists the treatment characteristics variables that significant figures in the field consider important to document in a bilingual education program.

Student Characteristics

When evaluating bilingual programs, the evaluator should take into account all student characteristics that may interact with one or more treatment characteristics in such a way as to affect the outcome variables being assessed (achievement, language proficiency, student attitudes) and thus confound the evaluation results. Balasubramonian (1979) warns that evaluations will be useful for program improvement only if all variables related to impact are included in the evaluation design. Increasing the number of variables complicates the evaluation process and increases the probability of obtaining less reliable data; however if interacting variables are ignored, treatment effects may appear weaker than they really are or even be totally obscured. In the following section, we will identify those student characteristics that researchers believe should be taken into account when evaluating instructional programs for LEP students.

Socioeconomic status and minority culture. Numerous studies have shown that the background characteristic which most directly affects school achievement is socioeconomic status (SES). Coleman et al. (1966), Jencks et al. (1972), Moore and Parr (1979), Baral (1979), Veltman (1980), De Avila (1981), Izzo (1981), and Rosenthal, Milne, Ellman, Ginsberg, and Baker (1982) all report that SES, determined by parental income and/or education, has a significant effect on academic achievement. Sociological studies suggest that low SES children are deprived of

TABLE 2

List of Characteristics That Should Be
Documented in Evaluation of Bilingual Education Programs

Instructional Variables

1. Language(s) in which literacy skills are developed.
2. Language(s) in which subject matter content is taught.
3. Proportions of instructional time in L1 and L2.
4. Point of introduction of instruction in English literacy.
5. Pattern of language usage.
6. Classroom staffing pattern and staff member duties.
7. Student-teacher (aide) ratio.
8. Student grouping pattern.
9. Duration of treatment.
10. Treatment hours per subject per week.

Materials

1. Availability of L1 and L2 materials.
2. Appropriateness of L1 and L2 materials.
3. Adequacy of resources and time available for materials.

Staff

1. Staff characteristics and qualifications.
2. Adequacy and appropriateness of staff development opportunities.
3. Rates of teacher attendance at voluntary training.
4. Extent to which "on waiver" teachers become credentialed.
5. Extent to which teacher-shortage problem is being ameliorated.

Parent/Community Involvement

1. Adequacy of outreach activities to obtain parent/community involvement.
2. Adequacy of parent training.
3. Extent of available involvement opportunities.
4. Responsiveness of program to parent/community inputs.

material advantages which promote better performance, such as books, calculators, and a quiet place to study (Mercer, 1977, So & Chan, 1984). The socioeconomic status of the students has been offered as one reason Canadian French immersion programs are more successful than similar American programs for language-minority students (Lambert, 1977; Cohen, 1976).

Cultural differences also affect the educational attainment of minority students (Deutsch, 1973); Hess, 1970; Shipman & Bussis, 1968). Out of the body of research on the effects of poverty and minority status, the concept of the "hidden curriculum" has been developed. The hidden curriculum refers to the rudimentary orientations, motivations, and prerequisite skills that prepare a child to benefit from schooling (Chan & Rueda, 1979). These attitudes and skills are generally developed in early childhood through socialization experiences and exposure to learning tasks in the home. Deutsch (1973) found that children from low-income, minority families were deficient in rudimentary cognitive skills required in formal learning settings, and in their ability to speak standard English. Katz (1967) found that these children also lacked the motivation to attend and perform well in school. Low-income and minority children reportedly did not behave in the ways that were expected or tolerated in the classroom (Rosenfeld, 1971). Cummins (1979) has suggested that low SES minority-language children are dependent on the school to provide the prerequisites for the acquisition of literacy skills, while high SES children may receive these prerequisites at home.

Low-income, minority families have limited resources they can allocate for training their children. Their financial situation also restricts their access to information about good child-rearing practices and support from social agencies, and they are often misinformed (Hurwitz, 1975). Low-income adults do not make extensive use of printed media (40% read less than one hour per week) (Hurwitz, 1975), apparently relying on electronic media as their main source of information (watching six hours of TV per day on the average) (Dervin & Greenberg, 1972). Generally, language-minority LEP students come from cultural groups that are called "caste minorities," meaning they may be viewed as innately inferior by the dominant group (Ogbu, 1978).

Studies indicate that socialization practices of minority cultures can result in the acquisition of adaptive behaviors that conflict with the development of factors related to academic success (Gallimore, Boggs, & Jordan, 1974; Hirata, 1975). Conflict between cultural patterns learned at home and student behaviors sanctioned by the school can create problems for minority students.

Gallimore et al. (1974) report that Hawaiian children, who are accustomed to being cared for by older siblings and are peer-oriented, may be accused of cheating when they consult older siblings and peers, or monitor the behavior of other students without their teacher's permission. Mexican-American children, who are also peer-oriented, work effectively in small, cooperative groups and are most diligent when they understand and accept the purpose of school tasks (Wong-Fillmore, 1983). Students from cultures which foster the development of a cooperative style and promote the welfare of the community sometimes find it difficult to function in the individualistic, competitive orientation of the American classroom (Klienfeld, 1979; Wong-Fillmore, 1983). Eskimo and Native American students have been viewed negatively by teachers because they are reluctant to bring attention to themselves and tend to withdraw in class when called upon (Cazden, John, & Hymes, 1972; Klienfeld, 1979).

Cummins (1984) suggests that the perception of powerlessness in minority communities may influence patterns of parent-child interaction and linguistic and motivational styles transmitted to children. Parents may not communicate a positive feeling toward school nor provide successful early learning experiences for their children, particularly if they have no formal education or have had negative experiences in school. Years of discrimination and cultural isolation can result in ambivalence toward the majority culture and insecurity and shame about the home culture and language (Heyman, 1973; Mougeon & Canale, 1978-79); Skutnabb-Kangas & Tokomaa, 1976; Troike, 1978). When low self-esteem is reinforced by negative attitudes of school staff toward minority languages and cultures, students "mentally withdraw" from academic tasks (Carter, 1970).

Low-income, minority communities are often unstable due to high unemployment and mobility. Unemployment can result in depression, apathy, dis

orientation, and withdrawal according to Levin (1975). Prolonged unemployment creates an unstable environment which affects childrens' early socialization and has a negative impact on their educability (Chan & Rueda, 1979).

Low-income, minority groups are often highly mobile. Some families are employed as migrant farmworkers and must travel to different work sites. Some families move in search of new jobs. Some Mexican and Puerto Rican families periodically move back to their original homes for extended periods of time. Children may also be shifted between parents or relatives in the event of divorce or separation. It is common in bilingual programs for students to miss several months of school, enter school late, or leave before the end of the school year. Interruptions in schooling and/or transfer to different schools can delay or impair the acquisition of academic skills, and will certainly reduce the effectiveness of any instructional program.

Minority-language students' school readiness is not only affected by the development of pre-literacy skills, it is also influenced by the amount and quality of primary language use in the home (Cholewinski & Holliday, 1979; Cooley, 1979; Cummins, 1979; Laosa, 1975; Shafer, 1978; Wells, 1979). Studies conducted by Carey and Cummins (1983), Ramirez and Politzer (1975), and Yee and La Forge (1974) indicate that the use of L1 in the home does not hamper the acquisition of L2 academic skills in school. Research by Chesarek (1981) and Bhatnager (1980) suggests that switching to the use of L2 in the home is correlated with poor academic progress. The crucial factor in terms of academic success is not which language is used in the home, but rather the quality of interaction between children and adults. If parents use English and they are limited in their proficiency, parent-child interactions will be restricted and children's language development will be hampered.

Age. Research indicates that linguistic outcomes are affected by a child's age. Contrary to popular belief, adults acquire a second language more quickly than children (Asher & Price, 1969; Oyama, 1976; Snow & Hofnagel-Hohle, 1978). Those who are exposed to L2 in childhood (in a natural setting) achieve a higher level of L2 proficiency than those who acquire L2 as adults, however, (Krashen, Long, & Scarcella, 1979). On the other hand, older children, aged 12 to 15, learn

morphology and syntax faster than adults (Snow & Hoefnagel-Hohle, 1978), and more quickly than younger children in either a natural or formal environment *if* the exposure is equivalent.

Asher and Price (1969), Olson and Samuels (1973), and Fathman (1975) report that students 11 to 15 years old are superior to students less than 10 years old in their acquisition of morphology and syntax. Seven- to nine-year-olds are superior to four- to six-year-olds in their morphology, syntax, and pronunciation (Ervin-Tripp, 1974). Furthermore, children who begin formal L2 instruction at a later time (senior or junior high school) catch up to those who begin earlier (in elementary school) (Bland & Keisler, 1966; Burstall, 1975; Oler & Nagato, 1974; Ramirez & Politzer, 1975; Vocolo, 1967). After reviewing the research on age of acquisition, Eckstrand (1979) concludes that "general cognitive development, native language learning, second language learning, learning ability and memory, perception initiation, and social learning will all improve with age and are positively interrelated" (in elementary and secondary grades).

Length of U.S. residence. Research indicates that the length of time LEP students have lived in the U.S. affects their achievement. The relationship between achievement and length of stay in the U.S. is not a straightforward one, however. Christian (1976) reports that recent immigrants rarely experience the educational problems faced by native-born Mexican-Americans. According to Troike (1978):

> it is a common experience that...children who immigrate to the United States after grade six...rather quickly acquire English and soon outperform Chicano students who have been in United States schools since grade one. (p.21)

Observational studies indicate that students born in Mexico achieve at a level equal to or better than second and third generation Mexican-American students (Anderson & Johnson, 1971; Kimbal, 1968). Carter (1970) reports that many teachers and administrators surveyed in four southwestern states believe that children who recently immigrated to the U.S. perform better academically than native Mexican-American students, and also acquire English rapidly.

On the other hand, Baral's 1979 study found that students who transfer to U.S. schools in the primary grades do not perform as well academically in junior high as native-born students. A study conducted by Skutnabb-Kangas and Toukomaa (1976) concludes that the length of schooling in the primary language may be critical to second language learning, although Baker and de Kanter (1981) question the validity of this inference. Baral (1979) proposes several explanations for these contradictory findings: (a) the impact of immigration status is confounded by SES factors; (b) the expectancies of teachers affect their treatment of students and consequently the students' performance; or (c) the benefit of native language instruction may be attained only after prolonged instruction in L1 (more than several years of L1 instruction). Clearly, additional research is needed to clarify these findings.

Prior education and experience. The educational experiences of LEP students prior to their entrance into bilingual programs can modify the impact of instructional interventions. Students do not always enter programs at the kindergarten level, and they may not have received bilingual instruction or any prior instruction at all (Hubert, 1982). Incomplete exposure to a particular program may reduce the effectiveness of the intervention. Some research indicates that the educational prognosis of late arrivals may differ from that of students entering school at an earlier time (Skutnabb-Kangass & Toukomaa, 1976). Differences in prior educational treatment--participation in preschool; exposure to monolingual instruction in either L1 or English; exposure to bilingual instruction; participation in special education or gifted programs--may modify the impact of the program being evaluated (Hubert, 1981).

Home language environment. The language used in the home appears to have some impact on the academic progress of students. The National Assessment of Educational Progress (NAEP) (1983) study of the impact of minority home language found that:

> ...some students from homes where English is not spoken often are much better readers than others. And some, in fact, read better than many students from English-dominant homes... Consequences of coming from an other-language-

> dominant home are not the same for students of different racial and ethnic backgrounds...
>
> White youngsters from other-language-dominant homes have a strike against them when it comes to reading skills. At age 17, these pupils are about 5 percentage points below whites from English-speaking homes in reading performance.
>
> For Hispanos, however, language spoken in the home doesn't appear to make much difference in reading abilities. For 17-year-olds, students from both other-language-dominant and English-speaking homes lagged about 9 percentage points behind the nation in reading skills. (p.3)

Several other studies provide further evidence of the complex interactions between home language use and intellectual performance. Bhatnagar (1980) reports that students who speak only L1 at home perform significantly worse than those who used both L1 and L2 at home; however, these results are confounded by length of U.S. residence.

Studies involving different linguistic groups provide evidence that supporting the home language does not interfere with acquisition of L2. Chinese students, whose L1 development is supported by exposure to a Chinese-speaking community and attendance at a Chinese school, perform better on the WISC than their peers who are not exposed to L1 outside the home (Yee & LaForge, 1974). Hispanic students who maintain L1 as their dominant home language perform better academically than those who switch to English as their dominant home language (California State Department of Education, 1981). Clearly, the interaction between home language and achievement indicates language exposure in the home should be taken into account in analyses of program outcomes.

Attitudes toward primary culture and language. Research on the often observed underachievement of language-minority students indicates that the attitudes of these students toward their culture may act as an intervening variable between educational treatment and achievement. Negative attitudes toward the majority culture by language-minority groups have been documented in different countries (Cummins, 1981). Heyman (1973) describes the attitudes of Finnish immigrants toward their primary and second languages:

> Many Finns in Sweden feel an aversion, an sometimes even hostility, towards the Swedish language and learn it...under protest. There is repeated evidence of this, as there is, on the other hand, of Finnish people--children and adults--who are ashamed of their Finnish language and do not allow it to live and develop. (p.131)

Cummins (1981) suggests that the reason students who immigrated after beginning school do better than U.S.-born minority students is that they did not experience ambivalence toward their culture in their early schooling and developed a secure identity and positive academic self-concept.

Cummins' interpretation is supported by acculturation studies. Chesarek (1981) and Bhatnagar (1980) report that "acculturated" students (those who adopt the culture of the majority and switched entirely to L2) demonstrate lower levels of academic achievement than students who maintain their allegiance to their native culture and the use of L1 at home.

Parents' ambivalence about the value of their native culture and language may also result in their children developing a negative self-image and a negative attitude toward L1. In contrast, parents who are proud of their culture are more likely to transmit their heritage to their children and "negotiate meaning" in L1 with them. Studies indicate that the process of "negotiating meaning" is a strong predictor of future academic success, and that children who are encouraged to develop L1 skills at home are better prepared to handle the communicative demands of school than those who are not (Chesarek, 1981; Wells, 1979).

Table 3 lists all of the student characteristics that research suggests may interact with treatment variables and thus affect achievement outcomes.

Setting Characteristics

The impact of bilingual education programs is also influenced by the community and school. The community in which a minority culture student lives will often reflect the characteristics of his home environment. As discussed earlier, such factors as minority status and poverty are thought by many researchers to have

TABLE 3

List of Student Characteristics That May Interact with Treatments in Bilingual Education Programs

---

Student Characteristics

1. Socioeconomic status
2. Age
3. Ethnicity
4. Sex
5. Length of Residence in the U.S.
6. Immigrant vs. native resident status
7. Prior educational history
   - (a) preschool (yes/no)
   - (b) years of schooling outside U.S.
   - (c) years of schooling in U.S.
   - (d) years of schooling in L1
   - (e) years of schooling in L2
   - (f) years of bilingual schooling
8. Age entered program
9. Early or late entry
10. Language proficiency at time of entry
    - (a) in L1
    - (b) in L2
11. Home language environment
    - (a) percent of time L1 spoken
    - (b) percent of time L2 spoken
12. Attitudes toward native language and culture
13. Academic aptitude

an effect on the achievement of students. For evaluation purposes, a description of the community from which a program's LEP population is drawn may be used to supplement or replace a description of the characteristics of individual LEP students.

School setting. School administrators and administrative procedures have an important impact on the implementation of instructional programs. The support of school administrators increases the probability that bilingual programs will be implemented as planned, while lack of administrative support often results in inadequate or inconsistent implementation (Ortiz, 1979; Teitelbaum, Hiller, Gray, & Bergin, 1982), particularly if the complexity of instructional tasks increases (Cohen, Deal, Meyer, & Scott, 1979). Without support, teachers and specialists, those directly involved in implementation, are insulated from administrative direction (Gross et al., 1971). Furthermore, teachers are typically oriented toward means and administrators are typically oriented toward ends (Wolcott, 1977), and this conflict eventually aggravates the separation of process from outcomes ("loose coupling") (March & Simon, 1958; Weick, 1976). Conflict regarding goals and means and "loose coupling" hampers successful implementation of school programs (Berman, 1978).

If school administrators are not actively supportive and involved in bilingual programs, coordination between the mainstream curriculum and the bilingual curriculum will be impaired, and bilingual teachers will be isolated from other teachers (Piper, 1984). This isolation has two unfortunate effects: teachers who teach the same children may not meet to plan a coherent comprehensive curriculum; and it is less likely that children in the bilingual program will be taught the same curriculum as children in the mainstream program (Cazden, 1985).

Administrative support can also be translated into resource support--the provision of time, materials, equipment, and other facilities. Lack of time and adequate materials are significant barriers to successful implementation (Charters & Pellegrin, 1973); Crowther, 1972; L. Downey Research Associates, 1975; Gross et al., 1971). Inadequate materials, space, and equipment create problems in program implementation (Berman & Fauly, 1975). Providing sufficient time for teachers to

familiarize themselves with new materials and methods and to work on problems individually and collectively contributes to the success of bilingual programs (Hamingson, 1973).

An important type of administrative support is providing teachers with feedback, particularly during early stages of implementation (Charters & Pellegrin, 1973; Gross et al., 1971; Center for Educational Field Studies, 1970). Feedback from consultants (Cole, 1971; Crowther, 1972) and other teachers (L. Downey Research Associates, 1975) was also found to support successful implementation. Regular and frequent staff meetings, which provide feedback, enhance implementation outcomes (Berman & Pauly, 1975; House, 1975).

Classroom climate. Many observers believe that certain classrooms, including some bilingual classrooms, have a positive atmosphere or learning environment which contributes to successful student outcomes (Wong-Fillmore, 1983). While classroom climate is difficult to measure, researchers have identified teacher behaviors and characteristics which have positive effects on student performance and which seem to be tied to the atmosphere of the teacher's classroom. Brookover, et al. (1977), Rutter, et al. (1979), and Weber (1971) discuss in this regard teachers' beliefs that they can make a difference and that all their students have the ability to succeed. Communication by teacher of high expectations of their students and a sense of their own ability to teach all students has also been named as significant specifically in teaching LEP students (Tikunoff, 1985). There is a large body of research indicating that structured classrooms are more beneficial than unstructured ones. Structure includes clear academic and social behavior goals (Santiago, 1975; Stallings, 1876; Stoll, 1979), supervision of students' work (Good & Grouws, 1978; Good & Grouws, 1979; Rosenshire, 1976; Rutter, et al., 1979; Tikunoff, 1985; Weber, 1978; Wright, 1975) and the use of lesson previews and reviews (Alexander, et al., 1979; Anderson, et al., 1979; Good & Grouw, 1979; Lawton & Fowell, 1979; Levin, 1973). In a structured classroom, students understand their tasks and a minimum of time is spent on non-learning activities such as behavior management or preparation of learning materials. This permits students more time to be engaged in assigned academic tasks, which has been correlated with higher student achievement (Fisher, et al., 1978).

Researchers have also found that warm, supportive teachers have a positive effect on their students (Brophy, 1976; Cantrell, et al., 1977). Appropriate, discriminating praise and encouragement by the teacher also seem to be associated with student achievement (Cantrell, et al., 1977; Frederick, et al., 1979; Brown & Epstein, 1978; Crawford, et al., 1977; Weber, 1978, Good & Grouws, 1977; Brookover, 1976). The use of cooperative goal structures, in which students can work together in groups to accomplish tasks, has been found to be important (Johnson & Johnson, 1974; Johnson, et al., 1978; Lucker, et al., 1976; Slavin, 1978). Competition among groups (as opposed to among individuals) has also been found to be effective (Brookover, et al., 1976; Clifford, 1971).

For LEP students, an atmosphere in which the student's home culture is recognized and respected in the classroom has also been identified as an important part of classroom climate that is related to student achievement (Tikunoff, 1985). Students' home cultures can be recognized by the teacher in such ways as using cultural referents during instruction, and observing the values and norms of the home cultures even while teaching the norms of the majority culture. Krashen (1982) has hypothesized an affective filter which is lowered in a culturally positive atmosphere. When learners feel that their languages and customs are understood and respected, their second language acquisition is enhanced because their resistance is lowered.

While the classroom climate will be affected by the school environment and by student characteristics, the literature on the characteristics of successful classrooms indicates that, to a large degree, it is the teacher who controls the classroom climate. Thus, teaching behaviors identified as contributing to an effective learning environment in the classroom can be measured as an index of the extent to which the teacher has created the desired classroom climate.

Table 4 lists the community, and school (setting) characteristics that have been found to impact on bilingual education program effectiveness.

TABLE 4

List of Setting Characteristics That Should Be Documented in Evaluations of Bilingual Education Programs

Community Characteristics

1. Poverty level of community
2. Degree of parent acculturation/literacy
3. Family stability/mobility
4. Language usage (percent L1 and L2)

School Characteristics

1. Administrative support for bilingual program
   (a) integration of program with other school programs
   (b) time and resource support
   (c) administrative attitudes toward program
2. Classroom climate

Measuring and/or Documenting Treatment, Student, and Setting Characteristics

It should be apparent from the discussions presented in this chapter that there are a large number of variables that may affect the outcomes of a bilingual program. It should be equally apparent that the task of measuring and/or documenting these variables will be substantial even if an extremely austere approach is adopted. Nevertheless, without adequate documentation, program evaluation will fail to serve many of its intended purposes.

In the most general terms, an evaluation should play two roles:

> One role of evaluation is formative; it serves to help and advise program planners and developers to describe and monitor program activities, assess the progress achieved, pick up potential problems, and find out what areas need improvement. Another major role of evaluation is summative; it is designed to provide a summary statement about the general effectiveness of the program; to describe it, judge achievement of its intended goals, pick up unanticipated outcomes, and possibly compare the program with similar ones. (Burry, 1982, p. 2)

Before conclusions are drawn in a summative evaluation, results of the formative evaluation should be known.

> It is clear that effects and potential effects of bilingual education cannot be evaluated adequately until a reliable process is found for determining the level of use that bilingual education has reached in the innovation-adoption process within the classroom, the school, and the district. (Dominguez, et al., 1980)

As research by Hall and Louchs (1977) has shown, a bilingual program may not be fully implemented until it has been in existence for several years. Levels of implementation will differ among teachers, classrooms, and schools. This has serious implications for summative evaluations, since, as Cordray (1986) points out, "strong causes produce strong effects and weak causes produce weak effects." If a program has been only partially implemented, a summative evaluation will show that it had minimal impact on students. If, however, the evaluation groups separate students receiving a fully implemented treatment and students receiving less than a full treatment, the effects of a thoroughly implemented program will become evident.

The formative role of evaluation involves comparing actual program events and activities with the intentions of the program designer or director. If the intentions have been well defined, the formative evaluation process will often entail little more than the identification of discrepancies between the program model and the program as implemented (Proves, 1971). If there is no detailed program model, one will have to be developed.

It should be noted that not all discrepancies will be "bad." Sometimes changes to a program model may be required to adapt it to a particular setting or to make it "work" with a target group different from the one it was originally designed to serve. In any case, it should be clear that detailed information about what a program is must be obtained before any conclusions about the program can be drawn.

On the other hand, good documentation is also essential if local evaluations are to be used to address the question, "What works for whom in what settings?"

Burry (1982) says, "Implicit in the concepts of documentation and evaluation is the desire to discover those effective practices maintained in the parent site which may then be adopted at other sites." Meta-analyses of sound, well documented local evaluations may afford an even better opportunity to address issues of effectiveness than large-scale studies--but only if the local evaluations are indeed sound and well documented.

On a smaller scale, good documentation is essential if meaningful comparisons are to be made between programs, or if data are to be aggregated across programs. Without such documentation, the kinds of interactions discussed throughout this chapter would only serve to obscure the benefits that may accrue from bilingual education. To draw an analogy, "medicine" has important health benefits--but only if the appropriate treatments are prescribed for specific diseases. Because some treatments will have negative effects on certain health conditions, "medicine" might be found ineffective if treatments were indiscriminately assigned to diseases.

The question of how to measure and/or document program characteristics is one that deserves attention. Unfortunately, there is an inevitable tradeoff between quality and cost. A variable such as the percent of time that instruction is conducted in L1, for example, is most effectively determined through classroom observation. As already mentioned, however, the simple fact that they are being observed may cause teachers to behave differently than they would if not observed. A classroom observer should thus be present long, or often enough so that the reactive effect of his/her presence will wear off before data collection begins. Such desensitization, of course, adds to the cost.

Estimates of L1 teaching time could be obtained for less cost by interviewing teachers and/or students, but one would have less confidence that the obtained data would be valid. A still cheaper and possibly still less valid approach would be to use questionnaires. Burry (1982) provides an excellent discussion of the various options available to the evaluator. Hall and Louchs (1977) have developed a level of use questionnaire which has been used to determine which components of a bilingual program were actually implemented in the classroom (Dominguez, et al., 1980).

Unfortunately, the cost-validity tradeoff for any particular bit of information will usually have to be governed by cost considerations. And, at this point, there is simply not enough known about bilingual education programs so that guidelines can be provided as to what proportion of the available resources should be expended on documenting each program characteristic. It is clear, however, that without knowing (a) whether the program exists, (b) what the program looks like, and (c) whether the program was implemented as planned (Center for the Study of Evaluation (undated)), it will not be possible to draw conclusions about program effectiveness.

5. MEASURING ACHIEVEMENT AND/OR AFFECTIVE GROWTH

An essential ingredient of any evaluation is a reliable measure of growth. (For our purposes, growth is broadly defined as improvement or even simply change--usually from pretest to posttest). Some growth may be due to special educational interventions. The remainder results from maturation, and from learning experiences other than those provided by the "treatment." The distinction between treatment-related and non-treatment-related growth is the subject of Chapter 6. Here we are concerned with total growth--the sum of treatment-related and non-treatment-related growth.

What we measure, with the instruments we use, we shall call *observed growth.* This observed growth reflects both *true growth*--the growth that the students actually experience--and whatever *error* is associated with the measurement process. Thus:

OBSERVED GROWTH = TRUE GROWTH + MEASUREMENT-RELATED ERROR

As can be seen from the above equation, if measurement-related error is small, observed growth will reflect true growth fairly accurately. As measurement-related error gets larger, however, observed growth provides an increasingly inaccurate estimate of true growth, and the statistical conclusion validity of any evaluation that includes large error components will be correspondingly low. For this reason, it is always an important goal of any evaluation to minimize measurement-related error.

For the purposes of this discussion, it is useful to consider two types of measurement-related error: systematic error or bias, and random error. Our equation for observed growth thus becomes:

OBSERVED GROWTH	=	TRUE GROWTH	+	SYSTEMATIC ERROR	+	RANDOM ERROR

Systematic error results when test scores are consistently either raised (for example by test wiseness) or lowered (for example by cultural bias) by factors other than the ability or trait of interest (irrelevant constructs). Random error is the result of unsystematic (chance) factors that affect test scores.

Components of Systematic and Random Error

From a program evaluator's viewpoint, systematic error may result from several causes among which are (a) not measuring things that were taught (low convergent validity) and (b) measuring things that were not taught (low discriminant validity). These two aspects of *curricular irrelevance* both reflect a mismatch between the content of the test and the content of the curriculum.

When one is dealing with cultural- or linguistic-minority students, another important source of systematic error is *cultural* and/or *linguistic bias.* A third source of systematic error arises when individuals who have a stake in the findings of an evaluation also participate in some aspect of data collection or analyses. Under such circumstances, it is not uncommon to see pretest scores somewhat depressed and/or posttest scores somewhat inflated compared to what they would have been had all operations been conducted by persons with no stake in the findings. Whether the influences that stakeholders exert are conscious or unintentional, their net result is that growth is overestimated. This source of systematic error is often referred to as *stakeholder bias.*

Finally, when either low- or high-scoring individuals are selected from a larger group to participate in some type of educational intervention, their scores, on successive subsequent testings, will move closer to the mean of the original group than they were on the selection test. Although this *regression-toward-the-mean* phenomenon was discussed briefly as one of the threats to the internal validity of evaluations, it deserves additional attention here as it is both poorly understood and frequently encountered.

If our concern is limited to the specific students for whom we have pre- and posttest scores, then the random component of measurement-related error is con-

fined to *measurement error* or what we shall call *test unreliability.*[7] If, on the other hand, we wish to generalize from the sample tested to the target population (assuming that the students tested are an unbiased sample from that population), then we must also consider random error due to sampling. *Sampling error* arises whenever we evaluate less than the entire population of interest and wish to generalize from the evaluation sample to that population. When dealing with groups of students, both test unreliability and sampling error are reflected in a statistic called the standard error of the mean. The standard error of the mean quantifies the amount of random error present in the means of a group's pre- and posttest scores. Since growth is defined as the mean posttest score minus the mean pretest score, a related statistic, the standard error of the difference (between means) is actually of more direct interest.

Assuming that no systematic error is present, the standard error of the difference can be used to establish "confidence limits" around the amount of observed growth. These confidence limits, in turn, provide an estimate of the amount by which observed growth is likely to be larger or smaller than true growth. Before proceeding, it is important to note that the random error reflected in the kind of confidence limits we just described can be reduced (and the confidence interval correspondingly narrowed) either by increasing the reliability of the test or by increasing the number of students in the evaluation sample.

As mentioned above, both random and systematic errors may be either positive or negative--that is, they may act so as to spuriously increase or decrease whatever quantity the evaluator is attempting to estimate. The most significant difference between the two types of error is that the direction in which random errors operate in any specific instance cannot be known in advance (and may not be known

---

7. Measurement error is one component of random error which, in turn, is one component of measurement-related error. To avoid possible confusion between measurement error and measurement-related error, subsequent discussions of measurement error substitute the term, test unreliability, for the term, measurement error.

even after the fact) whereas the direction of systematic errors is generally predictable. In flipping any given number of coins, for example, our best guess is that we will get a 50-50 split between heads and tails. Because of the random nature of coin flipping, however, we will often obtain different splits (sampling error)--and we have no way of predicting in advance whether we will get more heads or more tails than we expected.

This difference between random and systematic error has important implications. Consider our coin tossing example again. If we flipped just one coin, we would always get either a head or a tail. On a single flip, then, we would get either 100% heads or 100% tails and the deviation from our 50% expectation would be very large. As we increased the number of coins per flip, the tendency would be for our obtained results to come closer and closer to the expected 50-50 split of heads and tails. Sampling error thus tends to approach zero as the number of observations (individual heads or tails) comprising the unit of analysis (coins per flip) increases. A similar example could be worked out for test unreliability. Its effect on mean scores also approaches zero as the number of observations per unit of analysis increases. Unfortunately, systematic error (e.g., regression to the mean) does not cancel out in a similar fashion but remains a constant bias that is independent of the number of observations.

Because systematic error is unaffected by the number of observations, evaluators working with large samples should make it the focus of their effort to minimize measurement-related error. In small-sample studies, however, evaluators may have a choice between two methodologies, one that involves both systematic and random error and another that has a larger random error component but no systematic error. Despite its bias, the former method may be preferable if it yields a measure of observed growth that is closer to true growth than is provided by the latter method. It may even be possible to correct for the systematic errors if other studies have provided a means for estimating their magnitude. The point here is that bias is not necessarily worse than random error. This point should be kept in mind when reading the following discussion of the components of random and systematic error.

Test unreliability. The sources of test unreliability can be grouped under three general headings: task, student, and environment. Task variables include the nature and quality of the instrument itself. If a test is poorly constructed with ambiguous items and instructions, it tends to encourage irrelevant responses and thus introduce random error. It should be noted, however, that the ambiguity of both items and instructions will vary as a function of the students tested. What is perfectly clear for one group may be ambiguous for another.

Tests are appropriately regarded as samples of behavior. Most often they are focused on just one aspect of behavior (e.g., reading), but even when they are restrictively focused, tests sample behavior rather than examine it exhaustively. A vocabulary test, for example, may contain only 35 words--but those words may have been drawn from a list of the 2,000 most commonly used English words. Ideally, we would like the test score to tell us something about the students' understanding of the 2,000 most commonly used words. But a student who knows, say, 75% of the 2,000 words may know a substantially higher or a substantially lower percentage of the particular 35-word sample included in the test. He or she would most probably get a different score on a different 35-word sample drawn from the same 2,000-word population. Such differences between scores on alternate forms of a test reflect one type of random error that contributes to test unreliability.

The student is a second source of random error. At the time of testing the student may be particularly well rested, attentive, and motivated. Or he or she may be tired, excessively worried about the outcome of the test, and unable to concentrate. These time-to-time variations in "mood" will cause students to perform differently on the same test at different times. Variations in "luck" will also occur. Students may guess on items they do not know. They may not make the same guesses on successive administrations of the same test, and they may make more lucky guesses on one test than another.

The third category of source of error is environment, which includes both the testing and the scoring environments. Examples of testing conditions that can introduce error are physical arrangements such as temperature, lighting, and noise level; rapport between examiner and examinee; and variations in administrative practices.

Test scoring practices can produce error if there are clerical errors in scoring the tests, converting scores, and compiling summary statistics.

In addition, the three sources of error described above can interact in different ways. For example, some students are not as easily distracted by noise in the testing environment or as easily frustrated by difficult tests, clerical errors in scoring may be less likely with some tests than with others, some tests may hold students' attention better than others and thus be less sensitive to potential distractions, and so on. By delineating the different variables that can threaten test reliability, an evaluator may be able to devise strategies for minimizing their impact.

Most of the sources of unreliability discussed above tend to decrease as the size of the sample of behavior increases. Particularly relevant to this discussion is the fact that the reliability of any test will increase as its length increases. The relationship between test length and reliability is expressed by the Spearman-Brown formula:

$$\hat{r}_{tt} = \frac{nr_{tt}}{1+(n-1)r_{tt}}$$

where

$\hat{r}_{tt}$ = the estimated reliability of the lengthened test.

r_{tt} = the measured reliability of the original test.

n = the number of times by which the original test is lengthened (e.g., if the test has been lengthened by 50%, n = 1.5).

In bilingual education in particular, it is important to note that the length of a test may not correspond to the number of items printed on its pages. The *effective* length of a test is the number of items that test takers respond to. If those test takers understand and respond to only 20% of the items on a 50-item test, then the effective length of that test is 10 items. If test takers are able to comprehend only one or two of the items (or none, for that matter), their test scores will be virtually without meaning.

The manuals of some tests provide percentiles and even grade-equivalents corresponding to raw scores of one or two. This practice appears to lend "respectability" to very low test scores, but very low *raw* scores will typically have correspondingly low reliabilities. Low *percentile* scores, on the other hand, may have adequate reliabilities if they are derived from raw scores on tests that are written at appropriate difficulty levels (as could be the case when below-level tests are used). Low-achieving students will be able to respond to more items on the easier test level and their raw scores will thus be more reliable.

The low-score/low-reliability issue is particularly relevant to the testing of LEP students. If they do not have enough English-language proficiency to comprehend the questions on tests written in English, then there is no point in administering such tests to them. This disclaimer applies to tests of English vocabulary, reading, and language arts as well as to tests in other subject matter areas.

Low scores may not be the only cause of test unreliability when LEP children are tested with instruments designed for non-LEPs in mainstream classrooms. In this regard, it is important to point out that reliability is not inherent in an instrument but is a characteristic of a particular set of scores obtained by a particular set of students who took the test. The reliability figures presented in test manuals should thus be regarded with a good deal of skepticism. When culturally or linguistically different students are tested, reliabilities will almost certainly be lower--perhaps substantially lower. It thus becomes one of the evaluator's important responsibilities to make sure that whatever instruments are used have adequate reliabilities for the target group.

LEP children may or may not have better skills in their native language than in English. If they have, then testing them in their native language may be a viable strategy for obtaining adequately reliable test scores. Where suitable instruments are not commercially available, teacher-made translations of standardized tests *may* prove quite serviceable (a possibility which is discussed in some depth below). Another option, as was mentioned earlier, is to use below-level tests. Below-level testing, however, is only appropriate where the content of the test matches the con-

tent of the instruction the students receive. It is more likely that the content of a below-level test in English language skills will match the instruction provided to LEP students than in other subject areas where instruction is likely to be at grade level but in the students' native language (L1) In the latter situation, below-level testing, even in L1, is unlikely to yield any information useful for evaluation purposes.

One additional strategy for dealing with the low-score/low-reliability problem applies to groups where only some (necessarily fewer than half) of the students lack sufficient English language proficiency to obtain meaningful test scores. For such groups, the median score will be a more viable statistic to use for impact assessments than the mean. Although the standard error of a median is 25.3% larger than that of a mean when distributions are normal, medians will be substantially more accurate in situations where test ceilings or floors are encountered, or where there are significant numbers of "outliers." Use of the median under such circumstances would serve to reduce the instrumentation threat to internal validity.

The possibility remains, of course, that no adequate solution can be found to the low-reliability problem. In such cases the only alternative is to wait until the students attain language proficiency levels that enable them to understand the kinds of tests that are appropriate for assessing their academic progress.

Curricular irrelevance. A substantial amount of research attention has been focused in recent years on the content overlap between tests and the curricula of programs they are used to evaluate. As pointed out by Leinhardt and Seewald (1981):

> When a set of test scores are used to help evaluate the impact of instructional programs, knowledge about the extent of overlap is critical to interpretation of the results. If different instructional programs have varying degrees of overlap with the criterion measured, then results can be biased in favor of the program with the greater overlap. (p. 85)

Precisely this kind of situation occurred in the setting of a bilingual education program and was described by Cabello (1983):

> The CTBS and its Spanish version are, for the most part, equivalent in terms of vocabulary, content, and format. The Spanish language test is relatively free of language which might favor one ethnic group over another. The translation is generally accurate and the format is identical across tests. However, examination of curricular match in terms of vocabulary and general topics suggests that the English language version has a stronger match to English basal readers [than the Spanish language version has to Spanish basal readers]. (p. 48)

In this particular case, it is not clear whether the CTBS Espanol should be considered inappropriate for use in evaluating the effectiveness of the L1 instruction. It is a fact, however, that instruments with high curricular relevance will necessarily result in larger growth estimates than instruments with lower curricular relevance, all other things being equal.

The relationship between curricular relevance and effect size is one that makes a great deal of sense--it is clearly appropriate to test students on what they were taught and equally inappropriate to look for significant achievement gains on subjects that were not taught. Unfortunately, it is a relationship that program administrators and/or evaluators *could* manipulate to make their programs appear more effective than they really are. Narrow, highly focused curricula and tests that cover exactly what was taught (and no more) will show much larger effect sizes than broader curriculum- and domain-referenced tests. It would be possible to produce a very dramatic effect--one in which the lowest posttest score exceeded the highest pretest score, for example--by spending an entire year teaching a group of language-minority children 10 rarely encountered English vocabulary words. Clearly these students would be better off if an equal amount of time were spent teaching the alphabet and letter-sound relationships, developing decoding skills, and working on 500 frequently used vocabulary words. Unfortunately, the latter approach would appear to be less effective than the former.

The first point that needs to be made with regard to this apparent paradox is that we should not allow programs to have very narrow objectives. The objectives of any program should be appropriate to the educational needs of those it serves. Those needs will certainly not be confined to the kind of highly focused objectives referred to above. They will be the kinds of broader-based proficiencies reflected by standardized achievement tests. As Mehrens (1984) put it:

> The whole basis behind giving...various standardized achievement test batteries is that tests covering fairly general domains provide valuable information. People ordinarily wish to infer to the general domain. If one only wants to know about achievement on a particular and unique set of instructional objectives one should construct his/her own test. But let us not confuse such an *audit* with an *evaluation* of the program. (p. 11)

What Mehrens is saying is that an evaluation must consider the adequacy of the objectives as well as the extent to which they were achieved.

The second point is that testing need not be confined specifically to what was taught, particularly if we wish to infer to the general domain as Mehrens suggests. Green (1983) makes this point very nicely:

> If the students have learned fundamental skills and knowledge and understand it, they will be able to answer many questions dealing with material not directly taught...generalized skills and understandings do develop...since all the specifics can never be taught...this development is highly desirable and tests...should try to assess it. This can only be done by having items that ask about content *not* directly taught. (p. 6)

Of course the material tested but not taught should fall within the realm of what might conceivably be generalized or understood from what *was* taught.

The ideas discussed here all relate, albeit somewhat obliquely, to construct validity. If a treatment has the objective of developing language proficiency, the outcome measure should have high construct validity for language proficiency as operationally defined. Such validity may or may not imply a high degree of content overlap--depending on whether the treatment is well or poorly designed for produc-

ing the intended outcome. A high degree of content overlap could occur in the absence of any construct validity whatsoever. It is the possibility that such an absurd state of affairs could actually occur that has prompted some educational researchers to disparage the use of criterion-referenced tests (see below).

In the final analysis, it is simply not possible to specify the exact amount of overlap that should exist between test and curriculum. Mehrens appears to believe it is just as possible to have too much overlap as too little. He also seems to feel, however, that the overlap between standardized achievement tests and most curricula is about right. He does allow that different standardized achievement tests will have different amounts of overlap with different curricula. It should then follow that, if two equally effective curricula are evaluated with a single instrument, the one with the greater overlap will *appear* to be the more effective. Given this relationship, and the opinion that all standardized achievement tests have *about* the right amount of curriculum overlap, it appears that a well informed evaluator would wish to examine all candidate tests on an item-by-item basis and select the one that has the greatest overlap with the curriculum being evaluated.

In the realm of bilingual education, of course, the problem becomes more complicated. Instead of needing to choose among several tests that have appropriate levels of curricular relevance, it may be impossible to find *any* well suited instrument--especially if testing is to be done in L1. It is also clearly beyond the financial reach of local school districts to construct and standardize instruments that have the same level of psychometric sophistication as commercially published tests. This shortage of suitable instruments is one of the more difficult obstacles confronting well intentioned bilingual education evaluators.

On the other hand, the severity of the problem may have been overstated. We believe that less-than-ideal instruments can prove serviceable. Even instruments that are poorly matched to curriculum content will be able to detect educationally significant treatment effects if sample sizes are large (or can be made so by aggregation across time or across comparable treatment groups). Instruments that are psychometrically unsophisticated and whose reliabilities are substantially lower than those of standardized achievement tests will also prove useful under the same

circumstances. Before dismissing the possibility of doing any impact evaluation whatsoever, one should, therefore, examine all potentially useful instruments written in L1 *and* consider the possibility of developing others--either from scratch or through translation.

The question arises as to what kind of instrument development/modification activities do fall within the realm of economic feasibility. Unfortunately no clear-cut answer can be given. Even teacher-developed, classroom-type tests are likely to yield some usable information, however. Local translations of professionally developed English-language tests would seem to represent the next step up and should be considered if adequate time and expertise are available. Neither of these approaches appears *economically* out of reach, except, perhaps, for very small districts. Choosing a less-than-ideal but already available L1 test is an even less costly alternative. Further discussion of these various options is presented later in this chapter.

Cultural and linguistic bias. Concern with biases in tests is not new. Eleven papers summarizing the problem and attempts to deal with it are contained in Wargo and Green (1978). The literature citations in those papers most often came from the late 1960s and early 1970s--and work in the area has continued into the mid-1980s. Test debiasing methods have been developed and assessed (Ironson & Subkoviak, 1979; Marascuilo & Slaughter, 1981; Plake, 1980; Rudner, Getson, & Knight, 1980; Scheuneman, 1979), and there have been at least two major symposia on the topic (one sponsored by Johns Hopkins University in 1980 and an earlier one sponsored by the National Institute of Education in 1975).

While most of the attention that has been paid to test bias issues grew out of concern other than bilingual education, the topic has been correctly recognized as relevant by professionals in that field. Like those concerned about fairness to other minority groups, bilingual educators point out that whatever bias exists in tests used for assessing language-minority students works to depress the scores of those students. In other words, such children achieve lower scores than they would if the tests were truly unbiased.

Achievement or aptitude test scores that are spuriously low because of cultural or linguistic bias can have extremely unfortunate consequences. They can (and sometimes do) result in the misclassification of students as mentally retarded or learning disabled when their abilities really fall within the range served by regular school programs. These students may then be mistakenly placed in special education programs. In a similar fashion, spuriously low scores might cause bright students to be assigned to slow tracks--or cause their teachers to formulate slow-learning expectations for them. Clearly, any of these outcomes constitutes a valid reason for concern regarding the use of achievement tests.

While culturally or linguistically biased tests necessarily yield spuriously low scores, they do not have the same effect on assessments of *growth* or *change* (often simply the treatment group's mean posttest score ($\bar{Y}_T$) minus the same group's mean pretest score ($\bar{X}_T$). Measures of growth, in fact, will reflect zero bias *if* pre- and posttest scores contain equal amounts of bias. If, as is more likely, posttest scores reflect less bias than pretest scores, growth estimates will be *positively* biased, thus making the bilingual program appear more effective than it really is. The following paragraphs illustrate these two sets of circumstances.

If we assume that whatever bias exists in the pretest is present to an equal extent in the posttest, we can see that no bias remains in computations of growth:

$$\begin{aligned} \text{Growth (biased test)} &= (\bar{Y}_T - \text{bias}) - (\bar{X}_T - \text{bias}) \\ &= \bar{Y}_T - \cancel{\text{bias}} - \bar{X}_T + \cancel{\text{bias}} \\ &= \bar{Y}_T - \bar{X}_T \end{aligned}$$

Where:

$\bar{X}_T$ = the mean pretest score the treatment group would have had on an unbiased test.

$\bar{Y}_T$ = the mean posttest score the treatment group would have had on an unbiased test.

Under the assumptions of equal pre- and posttest bias, the growth estimate is thus unbiased.

If we assume that the posttest is less biased than the pretest--and it is reasonable to expect that acculturation occurring between pre- and posttest would cause it to be so--we can see that the amount of bias residing in the pretest which is not matched by bias in the posttest would actually serve to inflate the growth estimate.

$$\begin{aligned}\text{Growth (biased test)} &= (\bar{Y}_T - B_y) - (\bar{X}_T - B_x) \\ &= (\bar{Y}_T - B_y - \bar{X}_T + B_x \\ &= \bar{Y}_T - \bar{X}_T + (B_x - B_y)\end{aligned}$$

Where:

B_x = bias in the pretest
B_y = bias in the posttest

Unfortunately, we will not generally know how much acculturation occurred between pre- and posttests; thus we will not know by how much our growth estimates are inflated.[8] With carefully developed tests that have been submitted to one or more debiasing procedures, however, the absolute amount of bias in both pre- and posttests should be relatively small and the differences between these amounts should be smaller still.

---

8. It should be noted that, when we refer to acculturation, we are talking about the gradual learning of societal conventions that may facilitate the understanding of culturally biased test questions. We are *not* talking about the crossing of "linguistic thresholds" that may dramatically change what skill or knowledge is being measured by a single instrument from pre- to posttest administrations. Throughout this discussion, we are assuming that pre- and posttests measure the same content. If this is not the case, we would say that both pretest and growth indicators are uninterpretable.

One factor which is not considered when tests are debiased, however, is that scores can be affected by cultural differences in attitudes toward testing situations, strategies for coping with them, and the test wiseness that results from being tested frequently. The importance of these biasing factors has been well documented by Laosa (1982). Evaluators should certainly be aware of this source of systematic error and would be well advised to attempt to develop the students' test-taking skills prior to pretesting them. Other strategies for reducing this form of cultural bias would be to extend time limits and to clarify the directions given to the students.

Although we feel that *growth* estimates derived from instruments containing culturally biased items will not be seriously biased, any significant lessening of the effective length of the tests will increase the standard error associated with such growth estimates. This increase in the standard error will make it less likely that observed growth will be statistically significant. Effective programs might then be dismissed as ineffective. This possibility underscores the importance of minimizing cultural bias through use of any or all of the strategies referred to above.

In view of all that has been said thus far, evaluators should assume that some positive bias exists in all growth estimates derived from tests not developed specifically for the ethnic group tested. On the other hand, we believe that the bias will be small enough so that it will not render the growth estimates derived from such tests useless.

Stakeholder bias. It has already been mentioned that when evaluation data are collected and/or analyzed by persons who have a stake in the evaluation showing positive treatment effects, pretest scores appear to be somewhat depressed and/or posttest scores somewhat inflated compared to what they would have been had the evaluation been conducted by non-stakeholders. This stakeholder bias has been discussed by Keesling (1984), Linn (1982), and Tallmadge (1985) in conjunction with ESEA Title I evaluations where stakeholder involvement is the rule rather than the exception.

Although much of the evidence supporting the existence of stakeholder bias is indirect, it is compelling. One bit of direct evidence comes from a study by Elman

(1981) which showed that errors in test scoring and score conversions made by stakeholders produced positive-growth biases compared to machine processing. Other factors that have been suspected of contributing to stakeholder bias include: (a) minor differences in administrative procedures between pre- and posttesting, (b) instructions on test-taking skills between pre- and posttesting, and (c) "teaching to" the test.

An obvious approach to the prevention of stakeholder bias is to use only non-stakeholders for all facets of evaluation data collection and analysis. This approach also requires that the content of the test be kept secure from program teachers. The only threat that would remain uncontrolled if these practices were followed is that of providing instruction in test-taking skills.

For evaluations that track participating students for multiple years, an alternative strategy is to employ annual testing cycles where one year's posttest also serves as the following year's pretest. This practice, which has been advocated by all three of the investigators cited above, effectively defeats the behaviors that produce stakeholder bias. Such behaviors might produce inflated growth estimates for one year, but they would simultaneously have the opposite effect on the next year's findings.

Regression-to-the-mean. Regression biases affect many quasi-experimental evaluations and can work so as to either depress or inflate gain estimates. Bilingual education evaluations are particularly susceptible to inflated growth estimates because program participants are typically selected by virtue of their obtaining low scores on a language proficiency test. There will be quite large amounts of apparent growth from that selection test to all subsequent assessments of language proficiency. Such apparent growth, however, is purely artifactual and has nothing to do with real growth. Even if students are pretested after they have been selected, there will be small amounts of spurious apparent growth from pre- to posttest. The size of these various regression-effect biases will depend on both the reliabilities of the tests used and the correlations between the selection test scores and all subsequent test scores.

Regression artifacts do not affect treatment-effect estimates derived from randomized experiments. They are also at least theoretically controllable in non-equivalent comparison group designs (see Chapter 6). On the other hand, regression effects may introduce significant biases in other evaluation designs, *particularly* if the evaluator is unaware of the hazards associated with certain practices that are likely to appear sound to the uninitiated. Post-hoc score matching for the purpose of creating (seemingly) equivalent treatment and comparison groups is a classic example of an apparently sound practice that can produce highly misleading results (Thorndike, 1942).

Being aware of the dangers associated with regression effects is the first step toward controlling the biases they may introduce. Such knowledge will prevent evaluators from engaging in fundamentally unsound practices. Beyond that, there are certain statistical (converting raw scores to so-called true scores) and procedural (administering separate selection and pretests) controls that can be employed. While these controls may fail to eradicate regression biases from evaluations, they can reduce them to a level where they can be tolerated.

Instrument Selection/Development

The theoretical discussions presented above are all relevant to instrument selection/development decisions and are frequently referred to in the material that follows. It is not easy, however, to bridge the gap between theoretical considerations and the real-world instrumentation decisions that must be made by the local evaluator. To facilitate that decision-making process, the following presentation is organized by type of instrument.

Standardized achievement tests. Several authors have pointed out that standardized achievement tests were not developed for program evaluation purposes and have asserted that they are not well suited for such use (e.g., Carver, 1975; Hanson, Schutz, & Bailey, undated). This point of view, however, has not garnered much support among professional educational evaluators. Major nationwide evaluations of compensatory and bilingual education programs continue to rely

heavily on such instruments (e.g., Carter, 1984; Development Associates, 1983; Ramirez, Wolfson, Tallmadge, & Merino, 1984).

Standardized achievement tests, when used for program evaluation purposes, have most often been criticized for lacking curriculum specificity. In other words, the content of the test does not exactly match the content of the curriculum. Some experts, however, feel that this characteristic has significant benefits for program evaluation. Mehrens (1984) feels that standardized achievement tests are well suited for program evaluation purposes and quotes extensively from Cronbach (1963, 1971) to support his position. In taking this stance, he describes as assets for program evaluation precisely those characteristics of standardized achievement tests that Carver (1975) and Hanson et al. (undated) feel are liabilities (lack of a pretest match with instructional objectives, coverage of generalized rather than specific skills).

Without taking sides on the curriculum-specific/broad coverage debate, one thing *can* be said. Curriculum-specific tests will almost certainly be more sensitive to treatment effects than tests with broader content coverage. This characteristic is highly desirable *if* the goal of the evaluation is simply to detect treatment effects. If one wishes to compare the effect sizes of several different treatments that may have somewhat different instructional content, however, curriculum-specific tests are nearly impossible to deal with. This is a very important issue and one to which we have devoted nearly an entire chapter (see Chapter 7) of this report. We hope that, after having read Chapter 7, the reader will have a better appreciation for one of the characteristics of standardized achievement tests that we judged to be of considerable significance.

Standardized achievement tests *do* have several advantages with respect to other types of instruments. They are generally well constructed both editorially and in terms of their content. They encompass a range of item difficulties that is appropriate for the intended target group. They have high internal-consistency reliabilities. And items that are sexually or culturally biased have (usually) been systematically identified and removed. Such tests are generally easy to administer and

score (scoring services are often available), and they frequently provide normative data and/or other aids to score interpretation.

Standardized achievement tests seem nearly ideally suited for assessing the progress of LEP students in their acquisition of English language skills. It is important, of course, that the students whose progress is being assessed be able to comprehend the questions they are asked. If they are unable to do so, even on below-level tests, their scores wil be meaningless and should not even be collected.

Achievement in subject matter areas other than English is probably best assessed using tests written in the language of instruction. Where instruction is in L1, it is almost certainly because the students are more proficient in L1 than in English. Under these circumstances, testing them in English will result in scores that are spuriously low since language difficulties will prevent the students from revealing the full extent of their subject-matter knowledge. Unfortunately, there are few standardized achievement tests in languages other than English.

The InterAmerica Series: Tests of General Ability are the only instruments developed specifically with parallel English and Spanish forms. Although *user* norms are provided, they are not nationally representative and thus have somewhat limited utility. The California Test Bureau has developed a translation of the Comprehensive Tests of Basic Skills, Form S (CTBS-S) which is called the CTBS Espanol. The publisher undertook an equipercentile equating of the English and Spanish versions using a sample of test takers judged to be "balanced bilinguals." Through the equated scores, the CTBS-S national norms can be accessed by users of the CTBS Espanol. To the authors' knowledge, these are the only Spanish-language standardized achievement tests available to local evaluators. We are not aware of *any* standardized achievement tests in other non-English languages although "unofficial" translations have almost certainly been made (see *translated tests* below).

While we would certainly like to see standardized achievement tests developed in other languages and feel that such instruments would provide the best possible means for assessing growth in subject matter areas taught in L1, we are not

optimistic that such developments will occur. Laosa (1985) feels that developing appropriate instruments should be given high priority and that the potential market is sufficient to repay developmental costs. Although he may be correct, we believe that the market has been researched by the major test publishers and that they have reached different conclusions. Government subsidies of test development activities, however, might afford a reasonable solution. In the interim, other approaches need to be considered.

When no suitable standardized achievement test written in L1 can be found, the evaluator *could* elect to administer a test written in English with full knowledge that the scores of LEP students will be depressed by the language barrier. To compensate for the student's language difficulties, instructions could be given in L1, time limits could be extended (standardized achievement tests are designed to be "power" rather than "speed" tests anyway), and bilingual proctors could even assist the test takers with unfamiliar English words. While such procedures are certainly less than ideal, they may be preferable to the other available options.

Language proficiency tests. The literature on language proficiency is voluminous, complex, and largely theoretical (see Ramirez et al., 1984, for a brief summary). Perhaps for this reason, many language-proficiency tests have been developed, often reflecting diverse theoretical perspectives. Generally, the instruments have been developed by linguists with limited psychometric expertise. Even tests that have been standardized have been the object of strong criticism on psychometric grounds. According to Willig (1985), who cites seven references to support her position:

> It is a known fact...that language tests in general, and the language tests in particular that are used to determine entry and exit into bilingual programs, have low reliability and low convergent validity...In fact, some of the tests actually correlate negatively with each other. (p. 301)

Language proficiency tests are most often used for bilingual program entry-exit decision-making purposes. They are occasionally also used for evaluation purposes, however. Although their psychometric properties suggest that they are less

than ideally suited for either application, it is only the latter that is of concern here. Would-be users of these instruments for evaluation purposes are strongly advised to examine the literature carefully to verify that the test they select will, indeed, be able to provide the needed measurements. Unreliable tests will lower the statistical conclusion validity of any evaluation, while instruments with low convergent validity can only raise doubts as to whether the construct of interest is being measured at all.

Numerous technical and practical reviews have been prepared describing the major language proficiency instruments used in grades K-6 (Bye, 1977; Evaluation, Dissemination and Assessment Center, 1976; Horst et al., 1980; Law, 1978; Pletcher, Locks, Reynolds, & Sission, 1978; Ramirez, Merino, Bye, & Gold, 1982; Rivera & Simich, 1981; Silverman, Noa, & Russell, 1977; Texas Education Agency, 1977; Troike, 1981; Ulibarri, Spencer, & Rivas, 1981). The California State Department of Education established a Language Proficiency Instrument Review Committee to evaluate and designate instruments to be used in the Annual Language Census. This committee produced a set of thorough and accurate critiques of the major instruments (1982). Although the critiques omit a few considerations, such as the amount of time needed to score tests, they represent one of the most thorough and up-to-date descriptions of the major tests. Because of the many negative conclusions of the Committee, we hesitate to recommend *any* of the reviewed instruments for use in bilingual education program evaluations. New instruments are being developed and standardized, however, that hold promise for resolving some of the problems of their predecessors (Abbot, 1985; De Mauro, 1985; O'Brien, 1985). Standardized reading readiness tests also appear to be viable alternatives to language proficiency tests.

Criterion-referenced tests. Criterion-referenced tests were described briefly above in conjunction with the discussion of curricular relevance. Basically they are instruments composed of items derived directly from the objectives of the instruction. The items may be samples from a clearly defined domain the students were intended to master (see Shaycoft, 1979), in which case test scores may reflect the proportion of the domain actually mastered. Alternatively, the items may reflect specific instructional objectives the students were expected to achieve. In this latter case, test scores reflect the number of objectives "mastered." In both cases, each test

item is directly related to the content of instruction. There is no material tested but not taught, or taught but not tested (although the test need not include all possible items--it need not, for example, include all possible items involving the addition of two-digit numbers without "carrying").

As has already been mentioned, criterion-referenced tests are almost certain to be more sensitive than norm-referenced tests to instructional effects. When constructing such tests, in fact, items may be selected based on their ability to discriminate between a group that has received the instruction in question and a group that has not. Another asset of criterion-referenced tests is that they are particularly useful for identifying which program objectives are being achieved and where the curriculum needs strengthening. They cannot, however, provide local program staff with a good perspective on how well students are doing with respect to the more general domains sampled by standardized achievement tests.

The major weakness of criterion-referenced tests is that they are curriculum-specific--a feature which precludes (or at least makes difficult) comparisons of impact between programs or the aggregation of evaluation findings across programs. These are important drawbacks even for local-level evaluations. Another *potential* weakness is low construct validity if the curriculum to which the test is matched is not well designed for producing the *desired* outcome.

Tests may be both criterion- and norm-referenced and such instruments may represent the best of both worlds. A few criterion-referenced tests with national norms are available commercially (e.g., California Test Bureau, 1982). Techniques have also been developed to "customize" norm-referenced tests so that they will yield information more directly related to local learning objectives (Jolly & Gramenz, 1984; Wilson & Hiscox, 1984). Still another option exists--that of building locally relevant criterion-referenced tests from commercially developed item banks such as the one offered by Science Research Associates.

Perhaps the majority of criterion-referenced tests that are used for program evaluation are locally developed. As such they are subject to all of the psychometric shortcomings that typify locally constructed instruments--items that have inap-

propriate difficulty levels, negatively discriminating items, and generally low reliabilities. We do not wish to imply that high quality instruments cannot be developed locally--only that instrument development is best left to qualified professionals who have adequate skill, time, and resources to do the job properly. Few school districts have either the personnel or the time and money needed to produce high quality instruments. Unfortunately, low quality instruments are likely to have such high measurement error components that they are incapable of detecting treatment-related change. This problem is particularly acute with criterion-referenced tests where only a few items measure each instructional objective.

Teacher-made tests. Taken individually, teacher-made tests are probably best classified as less-than-ideally-constructed, criterion-referenced tests. One would hesitate to suggest that such tests be used by themselves for pre-to-posttest growth assessments. On the other hand, cumulative class records compiled over the course of a semester or a whole school year appear to have substantial validity. As such they may constitute an inexpensive and useful source of evaluation information. Their usefulness, however, may be greatest in "mixed" classrooms where bilingual program participants or former participants receive English-medium instruction along with mainstream children. In such settings, the mainstream children can be regarded as a sort of norm group. If the LEP or reclassified LEP children maintain or improve their relative achievement status with respect to their mainstream peers, that evidence could be taken as indicative of program success. Losing ground, conversely, could only be taken as evidence of program failure.

One of the goals of bilingual education programs is to enable LEP students to progress effectively through school. A logical inference from this objective is that reclassified LEP students ought to be able to keep up with their monolingual English peers in mainstream classrooms. Cumulative classroom grades derived from teacher-made tests would appear to offer a valid basis for assessing ability to keep up. One potential problem here, however, is that keeping up in a slow-track classroom is very different from keeping up in a fast-track classroom. We would not wish to consider a program successful if it achieved that "success" by placing all reclassified LEP students in slow-track classrooms.

Translated tests. The literature on test translations contains numerous articles claiming that translated tests are useful, valid instruments (e.g. Hansen & Fouad, 1984; Lega, 1981; Mercer, Gomez-Palacio, & Padilla, in press). An equal number of articles can be cited on the other side of the issue, however (e.g., Chavez, 1982; Merino & Spencer, 1983; Rosenbluth, 1976). Of considerable interest is the fact that both the proponents and opponents of test translations cite highly comparable evidence to support their positions. Opponents are likely to say that tests lose too much reliability in the translation process. Proponents, using comparable statistics, say that only a little reliability is lost in translation. One is faced with the need to decide whether a given amount of reliability loss is too much or only a little. We shall suggest that the choice is best made after careful consideration is given to the alternatives to translated tests.

Most of the literature on test translations comes from the field of cross-cultural research where, as McCauley and Colberg (1983) point out, tests must be translated so precisely that "semantic and syntactic variables...[are]...absolutely non-culturally dependent (e.g., free of colloquialisms, idiomatic expressions, semantic localisms, and particular language-bound syntactic usage)" (p. 81). The authors go on to describe a procedure for rendering translated tests of reasoning ability "transportable" across cultures. As evidence of the success of their approach, they offer comparable reliability figures, high correlations of relative item difficulties across several language groups, and a small proportion of the total test score variance "accounted for by disordinal country x item interactions" (p. 90).

In a comment on the McCauley and Colberg (1983) paper, Van de Vijver and Poortinga (1985) point out that total-score differences between language groups could not unequivocally be attributed to differences in reasoning ability--the possibility of cultural or linguistic bias could not be dismissed. This particular problem, they suggest, has "no generally accepted solution" (p. 157).

What is of interest in this exchange is that the reasoning test, translated into the various languages, appears to be a valid, reliable instrument *within* each language group. It is only when between-group comparisons are made that the issue of bias comes to the fore. Cast in French, for example, the test may simply be some-

what more difficult than when cast in Castilian--but it discriminates between good, average, and poor reasoners in both languages.

In our earlier discussion of bias we attempted to show that the presence of bias is not nearly as significant when tests are used to measure change over time as when they are used to assess status at some particular point in time. Applying that logic to McCauley and Colberg's reasoning test, we would expect their instrument to yield valid measures of gain following a course of instruction in reasoning in all language groups. If different treatments were given to different groups, the instrument could also be used to quantify treatment impact (in a non-equivalent comparison group design) even though *ability* comparisons between groups at pre- or posttest time might be invalidated by cultural or linguistic bias.

When tests are translated, it is not always the case that reliabilities will remain high or that item difficulties will retain the same rank orders. The conditions under which these desired outcomes are likely and unlikely to occur have received some consideration in the literature, however. There is evidence, for example, that translations to similar languages (e.g., English to Spanish) are more likely to be successful than translations to dissimilar languages (e.g., English to Navajo). An example of the former class of translation is provided by Mercer et al. (in press) who concluded that:

> the internal consistency among the WISC-R and the [Mexican translation] measures of academic intelligence are comparable across three cultural groups...[and] the internal consistency among subtests of the ABIC, a measure social-behavior intelligence, is [also] comparable ; oss the three cultural groups. (p. 20)

With regard to English-Navajo translations, however, Rosenbluth (1976) reported that

> The Navajo version of the Boehm Test of Basic Concepts is a harder test than the English version. At least 30% of its items within acceptable ranges of difficulty and discrimination appear to be measuring a different meaning than that intended by the English. Only about 20% of the items measure in the same way in both groups. (p. 42)

Another factor that is, not surprisingly, relevant to the success of translated tests is the quality of the translation. Translators often, unwittingly, change meanings, and having translations "back translated" into original language frequently reveals rather dramatic differences. Chapman and Carter (1979) provide some interesting examples from an earlier study in which the Classroom Behavior Survey was translated into Iranian and then back to English:

Item 16.	Original:	This teacher never knows when to stop answering a question.
	Back translated:	The teacher of this class does not know how to stop lengthy answers given by students.
Item 33.	Original:	The teacher doesn't involve the students in discussions.
	Back translated:	The teacher of this class does not allow the students to participate in class discussions.

When back translations are done, it is a relatively easy matter to identify problem areas. Items can be retranslated until a version is found that back-translates unambiguously. The advantages of this approach are obvious, and it has been widely recommended (e.g., Brislin, Lonner, & Thorndike, 1973; Werner & Campbell, 1970). Unambiguous back translations are not always obtainable, however, due to unclear phrasing of the original or because the concept contained in the original does not have a counterpart in the second language. Even when an unambiguous translation can be achieved, the difficulty of the vocabulary may not match. Thus even when a back-translation procedure is employed, the psychometric characteristics of the translated instrument may be different from the original.

A related point is that there may be important differences between dialects within a particular language. Differences are often cited, for example, between Mexican and Cuban Spanish. Such differences suggest that translations should be done by local people (e.g., classroom teachers) who are thoroughly familiar with the vocabulary and linguistic conventions of the group to be tested. If existing translations are to be used, their adequacy should be checked by means of *local* back translations.

Various authors have presented suggestions for improving the adequacy of test translations. One of the earlier ideas is that of "decentering" (Brislin, 1976) in which both the original and the translated items are altered until the translation can be made unambiguously and both versions are clear and unstilted. Unfortunately, it is not always possible to change the original item. Other authors have suggested other approaches including a micro-propositional analysis (Valdes, Barrera, & Cardenas, 1984), a neo-Piagetian approach (DeAvila & Havassy, 1974), and cross-cultural transportability theory (McCauley & Colberg, 1983).

It is not clear how relevant much of the discussion of translation issues is to bilingual education--especially to bilingual programs for young children. Testing of young children generally involves short questions in the active voice involving specific rather than general terms. Metaphors, colloquialisms, and the subjunctive mood are rarely encountered. And vague words such as probably, frequently, unlikely, and sometimes are uncommon. These are precisely the characteristics that Brislin et al. (1973) list as the characteristics of *translatable* English. As long as we are dealing with English-language instruments written in that manner, translating them into other languages should provide a good solution to the problem of tests not being available in languages other than English.

Even where the instrument's language is substantially more complex, it is clear that tests *can* be translated successfully without substantial loss of reliability or discrimination power. Mercer et al. (in press) describe a translation of the Revised Wechsler Intelligence Scale for Children (WISC-R) that was developed by local researchers in Mexico City. Although these researchers did more than simply translate the WISC-R (they omitted some items that they considered biased and substituted Mexican "equivalents" for others), the resulting test had subtest reliabilities that were only slightly lower for Mexican children than for Mexican-American and Anglo children tested with the English version of the test. Subtest intercorrelations for all three groups were "of about the same magnitude" as those reported by Wechsler for the standardization sample (p. 20). Certainly, a test of comparable psychometric quality could not have been developed for the same cost or within the same time frame.

To summarize, there are certainly hazards to be confronted when dealing with test translations. Even careful translations of highly translatable material are likely to introduce some cultural bias and erode the psychometric quality of the instruments somewhat. On the other hand, modest amounts of cultural bias in instruments used to quantify growth are of little, if any, consequence. And the somewhat eroded reliabilities will almost certainly compare favorably with those of tests constructed locally "from scratch" even if those instruments are developed by trained psychometricians. In other words, we believe that the psychometric quality of carefully translated instruments will exceed that of available alternatives and will certainly meet minimum standards for the intended usage of such instruments.

Measures of academic aptitude. In theory, achievement and aptitude tests should be distinctly different. The latter are intended to predict future achievement while the former assess the extent to which learning objectives have been achieved. In practice, however, the two types of tests often bear more than a superficial resemblance to each other--particularly when the aptitude tests are of the group-administered, paper-and-pencil variety. Nonverbal aptitude measures such as the Raven Progressive Matrices test (Raven, 1940) and the performance subtests of individually administered intelligence tests are less similar to achievement tests--but then, they also tend to be less efficient predictors of future academic performance (Cronbach, 1970).

Occasionally, aptitude tests have been used as outcome measures for educational interventions, although this practice has usually been confined to early childhood programs. More often such measures have been used as predictors of performance and as covariates to adjust for preexisting differences between treatment and comparison groups in educational investigations where random assignment was not feasible. Another possible application is to assist in the interpretation of growth estimates resulting from bilingual education and other special instructional programs. All of these applications are discussed in Chapter 6. Our intention here is simply to discuss the strengths and weaknesses of all types of aptitude measures.

When scores on aptitude tests are used as outcome measures, our concern is with posttest-minus-pretest difference scores. As pointed out above, such difference scores are much less subject to cultural and linguistic bias than either the individual pre- or posttest status indicators from which they are derived. In the more common usage of aptitude tests, however, we have only a single, one-point-in-time status indicator. This indicator is likely to be significantly depressed by the cultural and/or linguistic biases inherent in whatever instrument was used to obtain it. There is a real danger then that students will be misassigned to slow academic tracks and/or that teachers will formulate low expectations for them. It is this possibility that lies at the heart of the anti-testing movement.

Of course, spuriously low scores on achievement tests can be misinterpreted and misused in exactly the same manner. There is a difference, however, insofar as achievement deficits are commonly regarded as "fixable" while aptitude test scores have a higher potential for depriving students of appropriate educational opportunities than achievement test scores.

If there were some way of reliably measuring true aptitude, there would, of course, be no problem. Individually administered intelligence tests in the children's native language probably come closest to this ideal. An additional increment of validity *may* be obtained, however, by administering such tests using both English and L1, since, as McConnell (1985) points out, bilingual children often have "a split language capability with some words and concepts in one language, and some in the other."

Nonverbal aptitude tests, not surprisingly, are less subject to cultural bias than verbal tests. And aptitude tests in L1 provide a hedge against linguistic bias. All of these measures, however, are likely to underestimate the true aptitudes of minority students. Their use in evaluations should be limited to applications where scores will not be available to teachers or administrators who might misuse them for other purposes.

We believe that aptitude measures can be useful adjuncts to evaluations of bilingual education programs. Even if they are culturally biased and therefore in-

valid for any across-ethnic-group comparisons, they can be useful indicators of *relative* academic potential within ethnic groups.

All evaluators should be aware of the pitfalls associated with using academic aptitude measures. They should be aware that individual indicators may be quite misleading and should seek multiple indicators wherever possible. At the same time they should know how to use such measures to advantage in improving the internal validity of their evaluations and interpreting their findings.

Other types of measures. All of the comments presented above apply to any paper-and-pencil measures that evaluators may use in attempting to assess the impact of educational intervention--including questionnaires, interest inventories, personality scales, attitude surveys, etc. The more subjective instruments tend, however, to be less reliable and more subject to situational influences than academic measures. They are probably also more subject to cultural biases and to translational difficulties. While they may provide some useful information, we hesitate to recommend them for anything other than supporting roles.

On the other hand, there are indicators that have substantially smaller error components than even the most objective achievement tests. Statistical data on attendance, tardiness, dropping out, grade retentions, referrals to special education and gifted programs, enrollment in secondary and/or postsecondary education, and even numbers of books checked out of the library fall into this category. They may also be highly sensitive indices of program impact.

Since the collection of such data neither burdens the students nor detracts from the amount of instruction they receive, we strongly encourage the use of this resource. Even with these seemingly objective measures, however, it is important to make note of relevant administrative policies and criteria to assure that comparisons can be made across administrative units. This caution applies especially to such statistics as grade retentions and referrals to special programs where local policy can have a far greater impact than treatment differences. Evaluators must be especially alert to any policy changes that occur during the course of an evaluation.

6. EVALUATION DESIGNS

The purpose served by evaluation designs is *not* to quantify growth. As discussed in the preceding chapter, growth can be measured via pre- and posttesting with the same (or equated) instrument(s). What evaluation designs are intended to do is determine how much of the observed growth can be attributed to the treatment. This is the essence of internal validity as applied to educational evaluations.

In Chapter 5 we introduced the following model for achievement or affective growth:

$$\text{OBSERVED GROWTH (OG)} = \text{TRUE GROWTH} + \text{MEASUREMENT-RELATED ERROR (MRE)}$$

True growth, however, has two components: treatment-related growth (TRG) and non-treatment-related growth (NTRG). Our model thus becomes:

$$OB = TRG + \widehat{NTRG} + MRE$$

The majority of the evaluation designs we discuss in this chapter provide *estimates* of non-treatment-related growth (NTRG). In doing so, however, they introduce a new source of error--the amount by which the estimated non-treatment-related growth exceeds or falls short of the actual non-treatment-related growth (lacks internal validity). We will refer to this source of error as design-related error (DRE).

$$OG = TRG + \widehat{NTRG} + DRE + MRE$$

What we are really interested in, of course, is treatment-related growth, and we can *estimate* this quantity by solving the above equation for TRG. We have:

$$\widehat{TRG} = OG - \widehat{NTRG} + DRE + MRE$$

The accuracy of our estimate of treatment-related growth is thus a direct function of the accuracy with which we measure the observed growth (reflected by the measurement-related error term, MRE, in the preceding equation) and the accuracy of our non-treament-related growth estimate (reflected by the design-related error term, DRE, in the preceding equation).

Measurement-Related error was the principal focus of the preceding chapter. This chapter is similarly concerned with design-related error and the threats to (primarily) internal and statistical conclusion validity that were described in Chapter 3.

The first six of the evaluation designs discussed below provide some form of empirically derived estimate of non-treatment-related growth. This growth shall, for convenience, be called the no-treatment expectation. The models differ from one another both in the method they employ to generate this no-treatment expectation and, more importantly, in the amount of design-related error they introduce in particular circumstances. This latter factor, together with considerations relating to the feasibility of model implementation, should constitute the primary basis for whatever decisions are made regarding inclusion of the model in the Title VII evaluation system.

Because of anticipated technical or implementation difficulties with all of the six designs that generate no-treatment expectations, we have elected to describe two evaluation designs that neither generate no-treatment expectations nor enable observed growth to be divided into treatment-related and non-treatment related components. While this deficiency is indeed a major one, the designs are capable of fulfilling other evaluation functions.

True Experiments

True experiments can take several forms. In all of them, however, treatment and control groups are created through a process of random assignment of students

drawn from a single population.[9] After the groups are formed, the treatment is administered to the treatment group and withheld from the control group and both groups are posttested.

If a pretest is also administered, we have what Campbell and Stanley (1966) refer to as the Pretest-Posttest Control Group Design. If no pretest is administered, the label Postest-Only Control Group Design applies. These two designs may be combined to produce the Solomon Four-Group Design.

In posttest-only designs, the treatment-related growth estimate is unbiased--that is, the designs are free of any systematic influences that would tend to favor one group over the other at posttest time. Pretest-posttest designs are also unbiased, but the use of covariance analysis can increase their precision by adjusting for whatever pre-treatment differences between groups resulted from the random assignment process. Covariance analysis also affords a more powerful test of statistical significance for between-group differences.

In all of these designs, the posttest performance of the control group (adjusted or unadjusted for pretreatment differences) is the no-treatment expectation, and the difference between the treatment and control groups' posttest scores (adjusted or unadjusted) is the estimate of treatment-related growth. The credibility of this estimate rests on the assumption that the control group's posttest performance is exactly what would have been shown by the treatment group had that group not received the treatment--an assumption whose credibility hinges on four sub-assumptions, all of which were discussed as threats to construct or internal validity in Chapter 3.

---

9. Some authors (e.g., Lord, 1967) have suggested that the designs may be used with pre-existing, intact groups if the assignment of students to groups was "random-in-effect"--that is, if the treatment and control groups are as much alike as they would have been if formed through random assignment.

- The pretesting experience (if there was a pretest) did not serve to sensitize the treatment group in such a way that it benefited more from the treatment than it would have in the absence of pretesting (selection and testing interaction threat),

- Awareness of group membership did not result in Hawthorne (hypothesis guessing threat) or John Henry (compensatory rivalry) effects or in resentful demoralization,

- The experiences of treatment and control group members during the course of the experiment were equivalent in all respects save that the presence or absence of the treatment (history and maturation threats), and

- The control group did not receive a partial (diffusion or imitation threats) or alternative treatment (compensatory equalization threat).

The first of the above-listed sub-assumptions is effectively dealt with in the Solomon Four-Group Design. It can also be avoided through use of the Posttest-Only Control Group Design, but, in that design, one loses the statistical advantages afforded by covariance analysis (which almost always employs pretest scores).

The three remaining sub-assumptions are not design issues. Under some circumstances, actions can be taken to increase the probability that they are met. In field settings, however, the evaluator may not be able to exert sufficient influence over events, and the validity of the designs may be seriously threatened.

Despite such threats, most evaluation methodologists consider true experiments to be so far superior to any other designs that they believe should be employed whenever there is any possibility of doing so. Articles by Boruch (1978), Boruch and Cordray (1980), and Campbell and Boruch (1975) all contain rather elegant pleas for the use of true experiments. Boruch and Cordray go so far as to recommend that Congress...

> ...authorize the Secretary [of Education] explicitly in each evaluation statute to use high quality designs, especially randomized field experiments, for planning and evaluating new program components, program variations, and new programs. (p. 7-2)

Although this advice may only make sense for special, Federally funded studies, it has not been heeded, even for such restricted application. Existing laws and deregulations governing Federal education programs typically require that services be provided to the students with the greatest need. Such provisions preclude random assignment. They also make the possibility of finding groups that could be considered equivalent on the basis of random-in-effect assignment extremely remote. This single impediment is sufficient to prompt the judgment that true experiments cannot be implemented in Title VII settings unless current legislation is changed.

Non-Equivalent Comparison Group Designs

The most common form of the non-equivalent comparison group design (and the only way that will be discussed here) is the Pretest-Posttest Two-Group Design. Both treatment and comparison groups are pre-existing intact groups, and the most important consideration when implementing the design is the similarity of the groups. Either "regular" or some modified form of covariance analysis incorporating pretest scores as a covariate is usually employed to adjust for whatever between-group differences existed when the evaluation began. It should be noted, however, that analysis of covariance (ANCOVA) is theoretically "correct" only when assignment to groups is random and within-group regressions are homogeneous. Neither of these assumptions is likely to be met in nonequivalent group designs. On the other hand, there is at least some evidence that ANCOVA is robust to violations of these assumptions (Overall & Woodward, 1977). Alternative analysis strategies such as Kenny's (1975) standardized gain approach are also available and have less restrictive assumptions.

Probably more has been written about the non-equivalent comparison group design than all other quasi-experimental designs combined (see, for example; Bryk & Weisberg, 1977; Campbell, 1963; Campbell & Erlebacher, 1970; Campbell &

Stanley, 1966; Cook & Campbell, 1979; Judd & Kenny, 1981; Reichardt, 1979a, 1979b; and Wortman, 1983). In addition to sharing all of the threats to internal and external validity associated with true experiments, non-equivalent group designs are plagued by the fact that, in order to adjust correctly for pre-existing differences between groups, one must have a covariate that reflects all of the differences between groups that cause differences in their test-score performance. With covariates that fail to meet this requirement, attempts to adjust for pre-existing differences between groups will almost always introduce systematic over- or under-correction biases.

For statistically unsophisticated readers, Reichardt (1979b) provides probably the clearest explanations of the various biases (threats to internal and statistical conclusion validity) that can be introduced when attempting to adjust for pretreatment differences between groups. As he points out, "regular" covariance analysis that uses less than perfectly reliable pretest scores as the single covariate will always systematically underadjust posttest scores for initial differences between groups. This underadjustment will work so as to favor the group with the higher posttest scores. Thus, if the control group scored higher on the posttest than the treatment group, the estimated treatment effect would be smaller than the real treatment effect. Conversely, if the treatment group outscored the control group on the posttest, the bias inherent in covariance analysis would make the estimated treatment effect larger than the real treatment effect. Multiple covariates add further complications.

One commonly used approach for dealing with the unreliable (single) covariate problem is reliability-corrected covariance analysis. In its simplest form (Porter, 1968) the pretest covariate is "corrected" for its lack of (preferable) alternate-form reliability, thus removing the undercorrection bias described above. Porter's correction, however, rests on the assumption that the measurement error in the pretest score is uncorrelated either with the true pretest scores or with the measurement error in the posttest scores--an assumption others have questioned.

Other correction strategies have been proposed by other investigators for both single- (e.g., De Gracie & Fuller, 1972) and multliple-covariate (e.g., Sorbom, 1978) analyses. Even more complex covariance-related models are available but need not be discussed here. All rest on assumptions about the unknown and un-

measured differences that exist between the treatment and comparison groups. To deal with these unknownables, a bracketing strategy is often recommended where treatment-effect estimates are generated using both a procedure thought to underadjust for pretreatment differences between groups and one thought to overadjust for such differences. The evaluator can then conclude that the true effect size probably falls somewhere between the limits established in this manner. Even when this practice is followed, however, findings should be described as tentative. Reichardt concludes:

> Typically, a large amount of uncertainly will remain regardless of how much data sifting, careful reasoning, and creativity goes into the analysis. The size and direction of some biases will probably still be largely unknown, and one or more of them may provide a reasonable alternative explanation for any alleged treatment effect. (p. 201)

One point usually overlooked in discussions of non-equivalent group designs is the fact that the severity of the analytic problems to be dealt with is a direct function of the pretreatment differences between groups. With large differences, any statistical adjustment is extremely hazardous. On the other hand, if there are no educationally relevant differences, posttest scores need no adjustment. Even if pretest scores are found to be equal, however, other important but probably unmeasured differences are likely to remain. Age, grade level, socioeconomic status, academic aptitude, motivation, and attitude toward school are a few examples. In bilingual education, home language, prior exposure to English, family mobility, and prior schooling are certainly variables that should be taken into consideration.

If comparison groups can be found that are highly similar to treatment groups in all of these respects, a non-equivalent group design would be a viable model for evaluating bilingual education programs, assuming that the small differences that *do* exist are measured and that appropriate adjustments are made for them. Unfortunately, it is extremely unlikely that such groups can be found. More probably, available comparison groups will differ markedly from groups of bilingual program participants; thus whatever flaws there are in the adjustment procedure may be magnified beyond tolerable levels. Such designs cannot be recommended

for Title VII applications except when highly similar comparison groups can be identified. (This caveat also applies to secondary analyses that involve comparison between, or aggregations across groups--see Chapter 7.)

Regression-Discontinuity Designs

Regression-discontinuity designs represent a special case of non-equivalent comparison group designs. Usually, the appropriate implementation of non-equivalent comparison group designs requires finding comparison groups that are as similar to the corresponding treatment groups as possible in all educationally relevant ways. In the regression-discontinuity design, the strategy is very different. A group of students is subdivided into treatment and comparison subgroups so that there is no overlap whatsoever between them in terms of the measured pretreatment status indicator. A cutoff score is established, and all students on one side of it are assigned to one subgroup while all students on the other side are assigned to the other subgroup. One subgroup receives the treatment while the other does not. Then both subgroups are posttested. Finally, within-subgroup regression lines are calculated, and the distance between their intercepts with the cutoff score represents the treatment effect. Figure 2 illustrates the regression discontinuity design in a situation where the treatment has had a substantial impact.

The regression-discontinuity design was "invented" by Thistlewaite and Campbell (1960) some 25 years ago. It has always presented serious implementation and analysis problems, however, and has not received as much attention in the professional literature as might otherwise have been the case. Over the years it has been periodically resurrected by Campbell and his students at Northwestern University. Most recently Trochim (1984) has demonstrated that sophisticated analytic routines can overcome many of the problems that have been associated with the model.

A variant of the regression-discontinuity model, the Special Regression Model, was described by Horst, Tallmadge, and Wood in 1975 and was subsequently incorporated in the Title I Evaluation and Reporting System (Tallmadge, Wood, & Gamel, 1981). Subsequent investigations by the same research group, however,

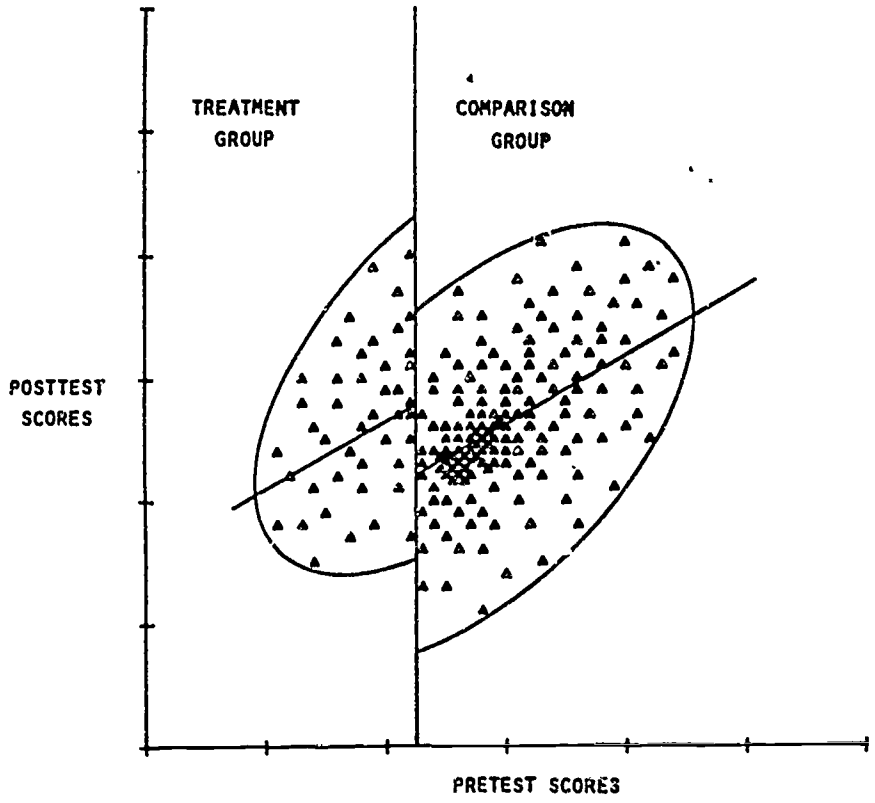


Figure 2. The regression-discontinuity design showing a substantial treatment effect.

identified serious problems with the model when implemented with linear regression equations (Stewart, 1980). In simulations performed on student groups to which no treatment was provided, it was common for regressions to be curvilinear (perhaps because of test ceiling or floor effects). In the presence of such curvilinear regressions, linear modeling produced different size "pseudo-effects" with different placements of the cutoff score (see Figure 3). It now appears that models using higher-order regression equations would have minimized--perhaps eliminated--such pseudo-effects.

It was Joyce Sween who first investigated higher order regression-discontinuity models in her 1971 doctoral dissertation at Northwestern. Boruch

(1974), Boruch and De Gracie (1977) and Trochim (1980) continued these developments, and the current state of the art is summarized in Trochim (1984).

The analytic approach currently suggested is to fit successively higher order regression equations to the data and to chart the resulting treatment-effect size estimates. The task is to determine the point at which the model becomes slightly overspecified and to stop there. In practice this may mean going several steps too far, examining the outcomes (including plots of the regression lines) and making a parsimonious and intuitively sensible choice of the "best fitting" model.

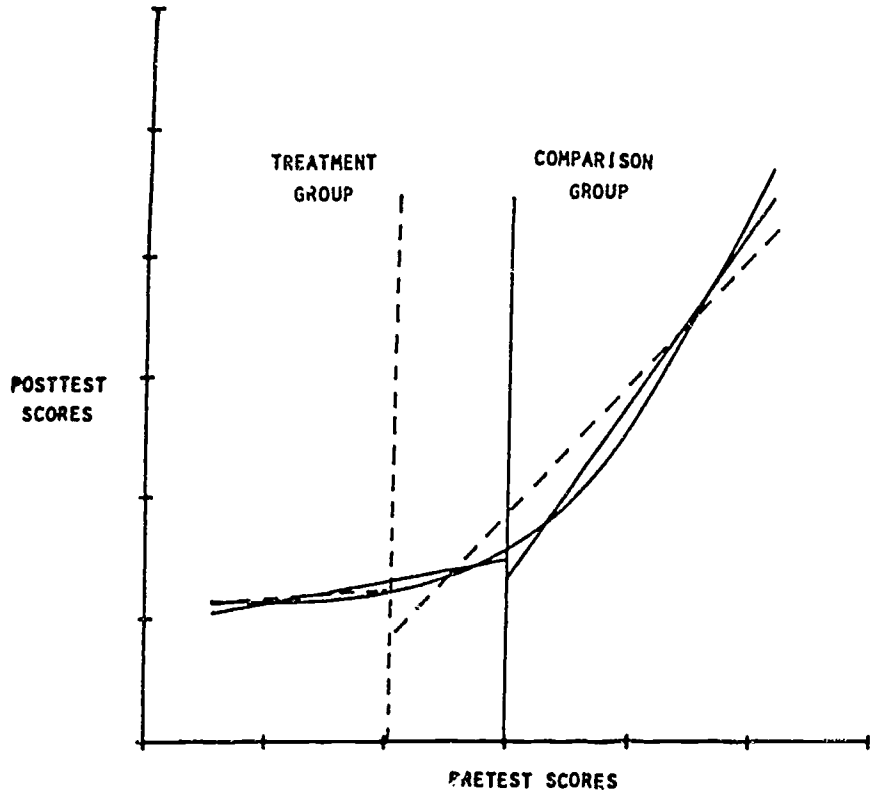


Figure 3. Different size psuedo-effects resulting from different placements of the cutoff score when linear models are fitted to a curvilinear regression function.

This approach produces a separate treatment-effect estimate for each order of regression equation that is investigated. It is up to the evaluator to pick the right one. Unfortunately, the choice is not always clear-cut, but fortunately, successive

estimates tend not to differ radically. An incorrect but "close" choice would thus not be too misleading. Indeed, it may be advisable to select two estimates to bracket the range within which the true effect is thought to lie.

There are three problems that come immediately to mind when one considers use of the regression-discontinuity model at the local level. First, large sample sizes are required if regression lines are to be stable. Second, it is computationally complex, and the required analyses can only be carried out by computer. Although Trochim (1984) provides computer programs, it is likely that many LEAs will not have convenient access to the required hardware or data processing personnel.

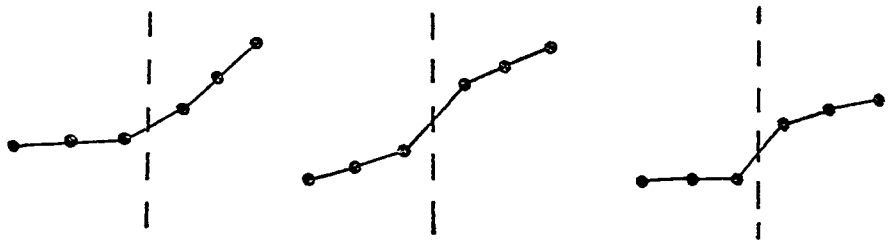
The third problem is that, even after all the analyses are done, selection of the "right answer" depends heavily on the expert judgment of the evaluator. The level of technical expertise that is required to make the right selection probably exists in very few LEAs nationwide.

A fourth problem is specific to bilingual education programs. The model assumes that the students above *and* below the cutoff score are representatives of a single population. Where the selection/pretest is a language proficiency test, the preponderance of students below the cutoff will be LEPs, while the preponderance of students above the cutoff may be native English speakers. Two distinct populations could thus be compared in much the same manner as they are in the norm-referenced model (see below). It would almost certainly be inappropriate to use the regression line of native English speakers to provide a no-treatment expectation for LEPs. In situations where there were enough reclassified LEPs above the cutoff to enable a stable regression line to be drawn, however, the model might be quite useful.

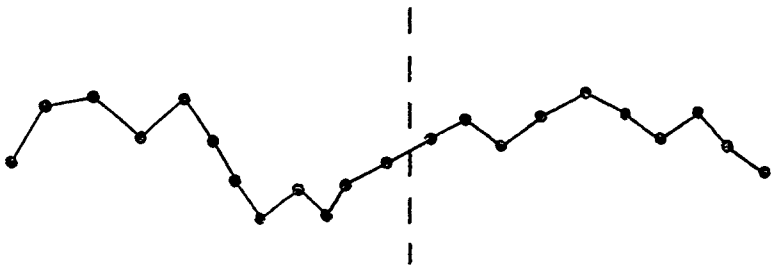
Time Series and Quasi Time Series Designs

In time series designs, a series of observations are made over some time period prior to an intervention, and another series of observations are made after the intervention. "Trend" lines can then be plotted through the "before" and

"after" data points. A treatment effect is inferred if the before and after trend lines have different slopes or if there is a discontinuity between the trend lines (with or without a change in slope). Three forms of positive evidence for a treatment effect are illustrated below.



The three illustrations above all provide relatively clear-cut and convincing evidence of project impact. Unfortunately, data points rarely fall on straight lines, and the effects of (particularly) social interventions are often difficult to detect in the presence of measurement error and other forms of "noise." A more realistic set of before and after data points is illustrated below.



Here one cannot be sure whether the treatment has had any impact without the aid of statistical data analysis.

At this juncture, it is important to point out that the label, time series design, has been applied to several, quite different modes. Textbook treatments of the topic generally discuss applications where there are large numbers of observations both before and after an intervention. Weekly counts of automobile accidents, for example, could be examined over periods of several years before and after nation-

wide adoption of the 55 miles-per-hour speed limit. With a data set of this nature, it is possible to pull out such influences as seasonal variations in accident rates that might contaminate the data if only brief pre- and post-intervention periods were studied. Suppose, for example, that the speed limit became effective just at the end of a particularly severe winter. The snow-free highways and improved visibility accompanying spring might themselves reduce accident rates compared to the preceding winter months. This effect could mistakenly be attributed to the lowering of the speed limit if seasonal influences could not be identified in the data and statistically controlled through the analytic process.

As Cook and Campbell (1979) point out, the common rule of thumb is that about 50 observations are required to perform a "competent" time series analysis. With fewer data points it is simply not possible to determine, and thus control for, the structure of the correlated error in the series.

The statistical analysis of time series data further complicated by the fact that adjacent (in time) data points tend to be closer in (dependent variable) value to one another than points that are separated by longer time intervals. This serial dependence (or autocorrelation as it is usually called) introduces bias into tests of statistical significance that are based on "ordinary least squares" regression. To eliminate this bias, experts on the topic today (e.g., Glass, Willson, & Gottman, 1975; Judd & Kenny, 1981; McCain & McCleary, 1979) recommend using the autoregressive integrated moving average (ARIMA) models described by Box and Jenkins (1970). A discussion of these models is beyond the scope of the pretest paper. It is relevant to note, however, that the statistical complexities of ARIMA models are non-trivial.

The need for a large number of data points is, in itself, sufficient to rule out this type of analysis for bilingual education program evaluations. If time series analyses are to be considered at all, they must be some sort of abbreviated version. Such designs, are, of course, possible, but they suffer from an inability to identify and control for sources of correlated error such as seasonal variation.

Glass et al. (1975) draw a distinction between repetitive and replicative time series designs. Repetitive designs are those that track the same entities over time--

as would be the case in a longitudinal educational evaluation. Replicative designs involve different entities at each data point. In a replicative design, one might, for example, track the end-of-year achievement test performance of second graders for several years before and several years after the introduction of a new curriculum.

McConnell (1982) described a quasi time series design which is replicative before the intervention and repetitive after the intervention. It appears to be particularly useful for the type of bilingual education setting in which she works. Although other settings may differ in ways that preclude application of this model, it is described here since it may be applicable in many sites other than its original home. We shall refer to this design as the *grade-cohort design*.

The bilingual program in question serves primarily migrant children who travel between Texas and Washington. It has an instructional component at both sending and receiving sites as well as one that travels with the students when they migrate. Significant numbers of new students enter the program each year at all age levels the program serves (age three through third grade). It is this last feature which enables the grade-cohort design to work.

Pretests administered at the times children entered the program (say at ages three, four, five, and six) provide cross-sectional, pre-intervention data points, while scores of tests administered after varying lengths of time in the program (say at ages seven, after one year in the program; eight, after two years of program participation; and nine, after three years) provide longitudinal post-intervention data points. With such data, it is possible to construct trend lines through the two sets of points and to look for discontinuities and/or differences in slopes as is typically done in time series analyses. Another possibility for data analysis (and, if fact, the one that McConnell employed) is simply to compare the scores of students who had participated in the program for some time with the pre-entry scores of students at the same age/grade levels. To control for the mortality threat to internal validity, of course, the comparison should include only the pre-entry scores of students who remained in the program as long as students in the treatment group. Failure to exercise this control could result in a substantial self-selection bias.

To summarize, full-blown ARIMA-type time series analyses are almost certainly not feasible in bilingual education settings. Abbreviated, quasi time series designs, like the grade-cohort design just described, appear to hold greater promise. Situations where it is possible to obtain pre-intervention test scores on sufficient numbers of children at all ages served may not be common, however. Thus the model, while having substantial merit, may have somewhat limited applicability.

Value-Added Designs

Bryk and Weisberg (1976) bear primary responsibility for the development of value-added designs. These designs have much in common with both time series and norm-referenced designs in that they generate a no-treatment expectation without requiring a control or comparison group. This is typically done (in value-added designs) by regressing the pretest scores of students on their ages, determining the number of "points" gained per month under no-treatment conditions, and multiplying the treatment duration in months by this value. When the result of this multiplication is added to the pretest score it becomes the no-treatment posttest expectation. Actual posttest scores minus this no-treatment expectation represents the *value added* by the treatment. Other factors of known relevance, such as socioeconomic status, may be included in the regression equation or controlled for using some sort of blocking strategy. The resulting growth curves are then used to predict achievement levels at posttest time.

Although design applications were developed for Title I early childhood programs (Bryk & Woods, 1980), they seem not to have found widespread adoption. The designs have also received little attention in the literature save the few papers by their developers. Reichardt (1979b) simply mentions that they do not provide "easily calculable significance tests" (p. 196) while Judd & Kenny (1981) feel there are "serious unit-of-measure problems." The latter authors also criticize the designs as deterministic and not adequately reflective of environmental influences on social and intellectual growth.

The limitations of value-added designs are clearly spelled out by Bryk and Woods (1980) who note that (a) they should be used only when the duration of the

intervention is considerably shorter than the age range of the pretest sample, and (b) children in the pretest sample must not have been exposed to any formal educational treatment prior to the pretesting. the design also rests on the assumption of linear pretest-on-age regression--an assumption that, according to Bryk and Woods, is unlikely to be met for treatment periods exceeding six to eight months.

If all of these conditions are met, the usefulness of the value-added design is largely dependent on the strength of the age-pretest correlation. A fairly high correlation--perhaps as high as .90--would be required to reliably detect effects of a likely size in typical treatment groups. Such correlations are unlikely to be observed. While adding predictors to the regression equation would increase the predictability of posttest scores, the sample size should be approximately doubled with each predictor added. Groups large enough to produce high enough (reliable) multiple correlations are unlikely to exist in any educational setting. They are even less likely to be found in bilingual education settings.

To summarize, the value-added design could only be applied (according to its developers) to preschool bilingual education programs. There would have to be quite large numbers of preschool children entering each program to be evaluated, and their ages at the time of entry would have to span at least 12 months (if the program were to span a school year) before stable growth expectations could be generated. Even under these circumstances, there is a good chance that the design would fail to detect educationally significant treatment effects. Based on these considerations, our recommendation is to abandon the design.

Norm-Referenced Designs

The origins of norm-referenced evaluation methodology are difficult to trace. Flanagan's 1951 suggestion that a "year's growth" afforded a defensible basic unit for assessing relative academic progress, however, was almost certainly the precursor of early Title I evaluations where greater than month-for-growth became the hallmark of successful projects. The logic that disadvantaged children who gained more than a grade-equivalent month for each month of program participation were catching up to the national norm seemed impeccable.

Despite the logic, however, Tallmadge and Horst (1976) identified serious flaws in evaluations that used this early norm-referenced model. Problems with the scaling of grade-equivalent scores, with norms interpolation, and with the use of a single set of test scores for both selection and pretest purposes led these authors to reformulate the design and incorporate several restrictions on its implementation. The design was subsequently incorporated into Title I Evaluation and Reporting System (Tallmadge, Wood, & Gamel, 1981).

A review by Linn (1982, p. 24) concluded that the design has an inherent positive bias (attributable to statistical regression) of "only about 1 or possibly 2 NECs"[10] when used with an annual testing cycle. An empirical study by Tallmadge (1982) found the bias to be "on the order of 1 NCE when typical Title I groups are examined" (p. 110). Otherwise the model was found to be technically sound. The model can be subject to stakeholder bias, however, under conditions when testing and/or scoring are conducted by parties who are "interested" in the evaluation showing positive results. Tallmadge found that the design was less subject to random error than true experiments because students serve as their own controls and that, even with its bias, it produced more accurate treatment-effect estimates than the Posttest-Only Control Group Design (in six out of six large-scale tests) and the Pretest-Posttest Control Group Design with "covariance adjustment" (in four out of six large-scale tests). These investigations, however, all employed high quality instruments that had been carefully scaled and normed. The design would certainly not work as well with tests that were poorly standardized.

The norm-referenced design derives its no-treatment expectation from the "equipercentile assumption" which specifies that groups of students will maintain their status relative to a locally or nationally representative norm group from pre- to posttest in the absence of a special instructional intervention. Tallmadge (1982) found that this assumption was tenable for large heterogeneous groups of low-achieving students; for mid-size groups of low-achieving students in low-

---

10. One NCE equals approximately one-twentieth of a national standard deviation (see Hills, 1984).

socioeconomic-status schools in small, medium, and large school districts; for mid-size groups of low-achieving students in low-socioeconomic-status schools in rural, town, small city, city, and large city settings; and for project-size groups (grade within school) of low-achieving students across all of the above settings.

The strengths and weaknesses of the norm-referenced design that are mentioned briefly above have been well documented (e.g., Kaskowitz, 1982; Keesling, 1984; Linn, 1982; Tallmadge, 1985). The design, however, has additional and "fatal" shortcomings in bilingual education contexts. Deriving a no-treatment expectation from national norms for LEP children participating in a bilingual education program is exactly analogous to implementing a non-equivalent comparison group design where the treatment and comparison groups are very different from each other in educationally important ways. Thus, it seems clear that deriving growth expectations for LEP students from non-LEP populations is a fundamentally unsound practice (Baker & Pelavin, 1984). For this reason, the norm-referenced design cannot be recommended for use in bilingual settings.

The Gap-Reduction Design

The gap-reduction design is the first of two designs discussed here that do *not* generate no-treatment expectations. Both of these designs measure growth from pre- to posttest, but neither of them provides any information whatsoever as to how much better off students are after receiving the treatment than they would have been without it. While this shortcoming may appear fatal at first, many evaluation questions can be answered with good estimates of how much growth occurred--even if it is not possible to break that growth down into treatment-related and non-treatment-related components. And of course, the design can be implemented with groups that do not receive a treatment (if suitable groups can be found) to provide estimates of non-treatment-related growth.

Consider the question of which of two treatments is the more effective with particular target group. Reliable measures of total growth will enable us to answer that question. Given similar settings and similar students (random or random-in-effect assignment), we can assume equal non-treatment-related growth. Then,

whatever difference we observe in total growth is treatment-related. Not only can we determine which treatment is superior, we can quantify the difference, test it for statistical significance, and make judgments regarding its educational significance. It should be noted, however, that making such comparisons is, in effect, implementing a non-equivalent comparison group design. All the caveats and cautions discussed under that design are equally applicable here.

According to Perez and Horst (1982), gap-reduction designs of two types have been described in the bilingual education evaluation literature. In one design, gap reduction refers to the achievement levels of program participants getting closer to the national norm over time. In the other design, participants' achievement levels getting closer to those of some dissimilar comparison group. Since the national norm can be regarded as a dissimilar comparison group, there is, however, no real difference between these two types of gap-reduction designs.

A third kind of gap-reduction design has been discussed by Baker and Pelavin (1985) and considers the difference between *actual* and *potential* achievement levels. If a program were able to reduce this gap to zero, it would be clear that it had accomplished its objectives (had been successful) and that the students should be exited from the program (if that step had not already been taken). Baker and Pelavin refer to reducing the actual/potential achievement gap to zero as "fixing the problem" and draw an analogy to taking a hard-starting car into the garage for a tune-up. After the tune-up the car starts easily (up to its potential) and the treatment can be classified as successful. Baker and Pelavin go on to argue that the success of bilingual education programs can be determined in the same way. Unfortunately, determining a LEP student's achievement potential is probably a task that can *never* be accomplished with adequate precision. Thus, while Baker and Pelavin's formulation is quite attractive at the conceptual level, it may be unsound in real-world usage and have the potential of serious negative consequences if invalid test scores are misused.

It is interesting to note that, whenever normalized standardized scores (e.g., normal deviates, T scores, stanines, or NECs) are used, the particular gap we choose to work with has no effect whatsoever on the amount of gap reduction that is

achieved. In fact, the amount of gap reduction is mathematically equivalent to the growth made by the treatment group minus the growth made by the comparison (or norm) group:

$$
\begin{aligned}
\text{Gap reduction} &= (\text{pretest gap}) - (\text{posttest gap}) \\
&= (\text{pretest}_{comp} - \text{pretest}_{treat}) - (\text{posttest}_{comp} - \text{posttest}_{treat}) \\
&= \text{pretest}_{comp} - \text{pretest}_{treat} - \text{posttest}_{comp} + \text{posttest}_{treat} \\
&= (\text{posttest}_{treat} - \text{pretest}_{treat}) - (\text{posttest}_{comp} - \text{pretest}_{comp}) \\
&= (\text{treatment group growth}) - (\text{comparison group growth})
\end{aligned}
$$

Thus, it is clearly irrelevant whether the performance level of the comparison group is equivalent to that of the treatment group at pretest time; or one, two, or three standard deviations above or below it.

Both growth and gap reduction can be measured using other types of scores (e.g., raw or scale scores). When such scores are used, however, they must be standardized (divided by their respective standard deviations). Such standardization has been shown by Yen (1986) to compensate for the fact that the scale units of some tests (those developed using Thurstone scaling procedure) get smaller as age/grade levels increase, while the scale units of other tests (those developed using item response theory scaling techniques) get larger.

Positive estimates of growth always imply that the students are learning something. Positive indicators of gap reduction imply that they are not only learning something, but that they are learning more than students in the comparison group. The latter indicator tells us more about how well the students are doing than the former. Still, it does not provide us with any definitive information about how well the program is doing. It would probably be safe to infer, however, that positive gap reductions would only occur when programs were having beneficial effects on their participants. Without special help, the same students could be expected to fall further and further behind their non-LEP peers.

Sometime in the future we may collect enough sound, comparable data on bilingual program participants to generate at least crude norms on rates of growth and/or gap reduction. We might even be able to compile credible evidence regarding no-treatment expectations from evaluations that were able to implement some of the more rigorous designs. Both types of data would add substantially to the meaningfulness of findings obtained from the gap-reduction design. Until such data are compiled, however, our inferences about program effectiveness will be limited to relative rather than absolute impacts.

Given this limitation on data interpretability, evaluators will try to squeeze every bit of meaning out of the growth and gap-reduction indices they are currently able to generate. This search for meaningfulness brings us back to the Baker and Pelavin (1985) notion of potential. It *does* seem that knowing something about student aptitude levels would be helpful in interpreting the findings from gap-reduction evaluation studies. Two thoughts come to mind. First, all other things being equal, we would expect programs serving high-aptitude students to produce larger gains than programs serving low-aptitude students. Second, as students' actual achievement levels approach their aptitude levels, we would expect the rate of gap reduction to fall off--perhaps to zero when the two reach parity and "the problem is fixed."

Unfortunately, the usefulness of the ideas expressed above is entirely depending on the validity of whatever measure of potential we are able to obtain. If our measures are spuriously low (and they are certainly more likely to be too low than too high), then we could be mislead in our interpretation of evaluation findings. We might, for example, conclude that a finding of zero gap reduction was due to students' having reached their potential when, in fact, they had not. The alternative conclusion that the program was ineffective would have been more plausible had we had a more valid measure of the students' potential.

Despite hazards of this nature, we are of the opinion that aptitude measures would be nice to have if they could be obtained within a project's evaluation budget. Even if they cannot be used to predict absolute levels of student achievement, they are likely to be useful as relative predictors. In the same sense, they may also be

useful as covariates when attempting to make comparisons among programs serving (slightly) nonequivalent groups. Unfortunately, the aptitude measures that have the highest predictive validity are also the most expensive (e.g., the individually administered, Wechsler Intelligence Scale for Children, Revised).

Whether or not aptitude measures are used as interpretive aids, it should be noted that the gap-reduction design, unlike all of the designs discussed previously, has no significant implementation difficulties. Although it provides no estimate of treatment-related growth, it could if implemented simultaneously with a treatment and a no-treatment group. It can also be integrated with any of the designs discussed previously (except the norm-referenced design) with a resultant increase in the information return obtained from those designs. These several considerations lead us to recommend inclusion of the gap-reduction design in the prospective Title VII evaluation system.

Group Criteria-Mastery Designs

Although some may question whether group criteria-master designs can really be considered evaluation designs at all, the approaches described below are currently the most widely used in bilingual education evaluation and thus deserve consideration here. The evaluation process, using these designs, begins with specifying, in quantifiable, behavioral terms (see Mager, 1962), the objectives that the program intends to achieve. A criterion of success is then established (e.g., 80% of the program participants will master 80% of the program objectives), and a test is constructed to assess mastery of all objectives. If, in fact, the established criterion of success is met, the program is deemed successful.

A variation on the approach just described uses existing, often norm-referenced tests. Criteria are usually specified in terms of some percentage of the students served attaining some national percentile level of achievement (e.g., 80% of the students will attain the 40th percentile in reading as measured by the XYZ Achievement Test). Although this variation does, indeed, have some of the flavor of the group criteria-mastery design, it neither assesses mastery nor examines learning at the level of the small, discrete, *behavioral objectives* that are the hallmark of

criterion-referenced tests. It is, in fact, an evaluation approach that more closely resembles the gap-reduction design than criteria-mastery design we consider here. Perhaps it should be described as the project of a mixed marriage of the two designs. Unfortunately it appears to lack the strengths of either while possessing the weaknesses of both. Thus, although it appears to be the most widely used of all evaluation designs in bilingual education, we have elected not to consider it further.

The usefulness of the group criteria-mastery design appears to depend on the appropriateness of the objectives that are established. If each program is free to establish its own objectives, there is a recognized danger that they will be structured so as to guarantee success. If a program fails to achieve the established criterion of success one year, for example, it may simply lower its goals for the subsequent year rather than strengthening the treatment so that the original objectives can be achieved. Even in the case of a new program, ideas as to what ought to be achieved may be tempered by fears of failure. Thus there may be a gradual erosion of performance standards leading, in turn, to a lowering of performance, prompting a further lowering of standards, and so on in an ever descending cycle of mediocrity and lowered treatment and outcome construct validities.

Some authors (e.g., Glass, 1980) maintain that any attempts to measure success in terms of percentages of students mastering percentages of behavioral objectives are doomed to fail. As he describes the issue:

> This language of performance standards is pseudo-quantification, a meaningless application of numbers to a question not prepared for quantitative analysis. A teacher, or psychologist, or linguist simply cannot set meaningful standards for activities as imprecisely defined as "spelling correctly words called out during an examination period." (p. 186)

He goes on to say:

> To my knowledge, every attempt to derive a criterion score is either blatantly arbitrary or derives from a set of arbitrary premises. (p. 186)

In the context of minimum competency testing, he adds:

> Teachers and their consultants attempting to define "competencies" and writing test items intended to reflect minimal levels of acquisition...are likely to construct a competency-based test for graduation that, perhaps, only half of the seniors can pass; then they will be forced to back off and be accused publicly of either not knowing what students ought to know or else not teaching students what they ought to learn. (p. 187)

Others would regard Glass's position as extreme. Roudabush (1978), for example, clarifies the difference between norm-referenced and criterion-refereced tests as follows:

> The score on a norm-referenced test [derives meaning] from its relationship to the scores of other students in a norm group and has little meaning in any absolute sense...A criterion-referenced test, however, purports to give absolute information about a student with respect to the objectives measured by the test. Meaning is derived from the relationship of the objectives to the curriculum and, therefore, essentially [reflects] the status of the student with respect to that curriculum without reference to other students. (pp. 257, 258)

While acknowledging the problems associated with formulating objectives for a program and obtaining consensus approval of them, Roudabush makes a convincing argument that criterion-mastery evaluation approaches can be much more useful for local program improvement purposes than evaluations using norm-referenced instruments. He also acknowledges that the effectiveness of different educational interventions can only be compared (using a criterion-mastery approach) when the objectives of the interventions are nearly identical. Non-comparable objectives would preclude such effectiveness comparisons. Roudabush argues, however, that program objectives can be agreed upon in the basic skill areas of reading and math and points out that "successful statewide assessments and evaluations have been carried out using only criterion-referenced tests" (p. 268)

Peleg (1978) advocates use of the group criteria-mastery design for bilingual education because other models are very difficult or impossible to implement in bilingual settings. She suggests that the achievement objectives established of

program participants be comparable to those established for their non-participating peers in the same content areas. While this suggestion seems inappropriate for English language proficiency, it may well have merit in other academic subjects. If, for example, program participants are taught the same science curriculum in their native language as non-participants are taught in English, it would certainly seem reasonable to test them on the same content--possibly using a translation of the test used with the mainstream students (although this would be a variant of the model if the test were not of the mastery type).

Peleg also points out that bilingual projects often use "commercial" programs to teach basic skills. These programs generally have clearly stated, measurable objectives. The task of developing evaluation instruments would thus be straightforward. More importantly, it is a task that could be shared among all projects using particular program, thus lessening the burden on individual projects.

The idea of common objectives and master instruments could be extended to non-commercial programs as well and would remove some of the subjectivity that critics of the design find objectionable. It would also provide a basis for making the kinds of across-project comparisons that were discussed above under the gap-reduction design. Even so, as Boruch and Cordray (1980) point out, criterion-mastery standards

> ...are insufficient for judging program success. Testing level of competency before and after the program...is an improvement over the after-only strategy...but is still insufficient for attributing the gain to the program. Other competing explanations such as normal growth are *as plausible* in accounting for the gains, as the program. (pp. 5-12)

In summary, the group criteria-mastery design has serious deficiencies and is subject to abusive implementation. If well implemented, however, it can be especially useful for local curriculum improvement purposes. Because of the limited comparability of scores across different criterion-referenced tests, on the other hand, our recommendation is to use the design primarily as an adjunct to other

designs. Even at the local level there is a need to know how favorably one's program compares to others serving similar target groups in similar settings.

Summary

Table 5 summarizes the strengths, weaknesses, implementation requirements, and applicability to Title VII settings of each of the eight evaluation designs reviewed in this chapter.

TABLE 5

Characteristics of Eight Evaluation Designs Considered for Title VII Applications

Design	Strengths	Weaknesses	Implementation Requirements	Applicability to Bilingual Evaluations
True Experiments	Highest internal validity	Threats to validity associated with knowledge of group membership.	Random (or possibly random-in-effect) assignment to experimental and control conditions.	None, since current legislation mandates serving neediest children.
Non-Equivalent Comparison Group Design	High internal validity if groups are nearly identical.	No completely satisfactory way to adjust for differences between groups. Severity of this problem increases with with the difference between groups.	Treatment and comparison groups must be very similar on all educationally relevant characteristics.	Very limited, as available comparison groups will differ substantially from treatment groups.
Regression-Discontinuity Design	High internal validity. Consistent with assignment to conditions based on need or merit.	No clear-cut method to determine "correct" order of regression equation. Computationally complex. Needs large sample.	Assignment based on strict cutoff scores. Homogeneity of ethnicity and native language across cutoff score.	Very limited due to need for large numbers of ethnically and linguistically similar students both above and below cutoff score.
Time Series Design	High internal validity in most circumstances if there are many pre- and post-intervention data points.	Subject to history threat to internal validity. Requires as many as 50 data points to control for some extraneous influences.	Requires many pre and post-treatment data points.	Quasi time series designs are possible wherever appropriate pre-treatment data can be trained. Controlling the history threat to internal validity requires more pre- and postintervention data points than may be obtained.
Value-Added Design	High internal validity under very limited circumstances.	Only suitable for short-term evaluations. Requires linear regression and no prior treatments.	Requires sample of preschool children having a range of ages that exceeds the duration of the evaluation.	Very limited--preschool only. Probably too insensitive for use with small treatment groups.
Norm-Referenced Design	High internal validity under limited circumstances. Easy to implement.	Inherent small bias due to statistical regression. Can only be implemented using tests with high quality norms.	Requires use of standardized achievement tests.	None. Norms do not provide a valid no-treatment expectation for LEP students.
Gap-Reduction Design	Easy to implement. Works well in conjunction with other designs.	Provides no estimate of treatment-related growth (i.e., has no internal validity).	Can be implemented with either a live comparison group or with norms.	Suitable for all programs.
Group Criteria-Mastery Design	Can identify strengths and weaknesses in local curriculum (assuming use of an objectives-referenced test).	Provides no estimate of treatment-related growth. Growth estimates lack comparability across programs using different tests. Subject to misuse.	Requires development/adaptation of an objectives- (criterion-) referenced test.	Suitable for all programs. Adaptable to the measurement of nonacademic objectives (e.g., parent involvement).

7. COMPARABILITY, AGGREGATION, AND A COMMON GROWTH METRIC

Effect Size

On several occasions in the two preceding chapters, reference was made to effect size--although little attempt was made to clarify the exact meaning of that term. In fact, effect size is difficult to define in the "soft" sciences where measurement scales are typically relative (lack true"zero" values) and probably even lack equal intervals. Without additional information, for example, the statement that the treatment group outperformed the control group by 5 points on the XYZ Reading Test is virtually meaningless.

Although there were earlier attempts to define and quantify effect size, it was Cohen's seminal article in 1962 that first brought the importance of this concept to the attention of the social science community. He used the difference between the mean (or adjusted mean) posttest scores of the treatment and control groups divided by the pooled, within-group standard deviation as his index. He went on to use it, along with sample size and whatever statistical significance criterion was selected, to describe the power of standardized tests.

Glass (1976) adopted Cohen's index in his formulation of meta-analysis. Other investigators have proposed other indices--mostly estimates of the proportions of total outcome variance accounted for by the treatment--but these statistics have received less than wholehearted acceptance by the professional community (see Sechrest & Yeaton, 1982). They may have substantial merit when used in conjunction with complex experimental designs, but they do not appear to be superior to the Cohen/Glass index in less complex situations. For that reason, and because the Cohen/Glass index is generally used in meta-analyses, we have decided to restrict our discussion to that estimate of effect-size.

It is interesting to note that, when Cohen developed his index, he was not concerned with meta-analyses or with any form of comparability or aggregation of

data across multiple studies. His single concern was the relationship between effect size and the power of a statistical test. Since statistical tests employ local (sample) variances both within and between groups, it was entirely appropriate for him to use local means and local standard deviations in his formula for effect size. When we consider the aggregation of data across studies, however, we are interested in the *comparability* of effect-size estimates, *not* with considerations of the power of statistical tests. Given this focus, we find the Cohen/Glass metric somewhat deficient.

Consider the possibility that two entirely separate bilingual education programs serving equal numbers of children of the same age and ethnicity both employed the same form and level of the XYZ achievement test. The two evaluations produced identical observed posttest scores and, in both cases, the observed posttest scores exceeded the no-treatment expectation by 10 scale-score points. To us it seems logical to conclude that the two treatments had equal effect sizes.

If we now learn that the comparison group[11] used in one of the evaluations was more homogeneous (standard deviation = 30) than the other (standard deviation = 50), should that factor alter our judgment that the two programs had equal effect sizes? We think not--but dividing the two 10-point gains by their corresponding comparison-group standard deviations yields quite different Cohen/Glass effect-size estimates of .33 and .20, respectively. While it *is* true that a 10-point gain will have a lower probability of occurring by chance in the more homogeneous group (and this relationship is important in computations of statistical power) it seems inappropriate to change presumably unbiased estimates of treatment effects on the basis of their non-chance probabilities of occurrence.

Unless the logic of the preceding paragraph is flawed, there would be no need to use any index other than observed posttest scores minus the no-treatment expectation to quantify effect sizes if all programs we wished to compare were

---

11. When one group is clearly the control group, it is common practice to use its standard deviation to compute effect size rather than the pooled, within-group standard deviation.

evaluated using the same test. It is only when we wish to make comparisons between effect sizes measured with different instruments (with scale units of different sizes) that we need to perform some kind of mathematical adjustments to effect comparability.

The most rigorous way to achieve comparability would be to perform equipercentile equatings (á la the Anchor Test Study described by Loret, Seder, Bianchini, & Vale, 1974) among all instruments. Such an equating study would be a major and very expensive undertaking. It could, however, be approximated, for standardized achievement tests, using publisher-provided national percentiles. To the extent that each publisher has succeeded in obtaining nationally representative samples, raw- or scale-score equivalencies could be established simply by finding the percentiles corresponding to each score on one test and then finding the scores that correspond to the same percentiles on all of the other tests. Using that procedure one could convert the scores on all tests to their equivalents on any one selected test. A somewhat simpler approach would be to convert all possible scores on all tests to normal deviates via area transformations of the corresponding percentiles. Subsequent analyses could simply use those normal deviates. Still simpler would be to use publisher-provided NCEs which are simply linear transformations of normal deviates.

A slightly less precise approach could be used for equating gains measured with different standardized tests. This approach would involve dividing the difference between the observed posttest scores and the no-treatment expectation by the national standard deviation of scores at the corresponding grade level. This approach would provide effect-size estimates similar to Cohen's but based on national rather than local standard deviations. As such, they would be immune to variations in the homogeneity of local treatment group scores and would have, we believe, significant advantages over the Cohen/Glass metric *when the goal is to achieve comparability across studies and instruments.*

While the metric just described has much to recommend it, (and is, in fact, exactly the type of metric employed in the TIERS Model A evaluation design), it can be adopted only in evaluations that employ nationally normed tests or tests for

which the standard deviations of nationally representative samples can be reasonably estimated. The restriction *may* not be critical in bilingual education evaluations, but it should also be noted that any modifications made to a test, its administrative instructions, or its time limits will alter its score-to-status-indicator relationship (e.g., raw scores to percentile conversions) and thus invalidate the entire quasi-equating procedure. In bilingual education applications, this restriction does seem sufficient to offset whatever advantages can be obtained by expressing effect sizes in terms of national-sample standard deviations.

Observed Growth, Relative Growth, and Treatment-Related Growth

The term, effect size, refers to treatment-related growth. In Chapter 6, however, we noted that there are likely to be situations in bilingual education where it is not possible to obtain valid no-treatment expectations and thus break observed growth down into its treatment-related and non-treatment-related components. In such situations the gap-reduction design appears to represent the best of the available choices for an evaluation strategy.

In the gap-reduction design we are not dealing with effect sizes but with gaps--and there is a *compelling reason* why those gaps must be expressed in terms of their corresponding comparison-group standard deviations. The need for such "standardization" stems from the fact that test-score standard deviations tend to either increase as a function of increasing age/grade levels (in the case of tests developed using Thurstone scaling procedures) or decrease (in the case of tests developed using item response theory procedures).

Suppose that an evaluation found a one-standard-deviation gap between the treatment-group and the comparison group on both pre- and posttests. That finding would indicate that the treatment group had exactly kept up with the comparison group--that there was neither gap reduction nor gap enhancement.

On the other hand, if the pre- and posttest gaps had been measured in terms of test score points instead of standard deviations, (apparently) different results would have been obtained. A test developed using Thurstone scaling procedures

would have shown that the gap increased from pre- to posttest, while a test developed using item response theory procedures would have shown that the gap decreased. Clearly, the appropriate way to present gap-reduction data is in terms of standardized, rather than raw- or scale-score measures. Yen (1986) offers convincing support for this conclusion.

What follows from the above is that gap-reduction measures must necessarily be expressed in standard deviation units--otherwise the artifacts of different scaling methods could be misidentified as differences between the growth of the treatment and comparison groups. But such gap-reduction measures are not comparable across studies employing comparison groups of varying degrees of homogeneity. They need no adjustment for interpretation at the local level, but they can bias comparisons between projects, and distort aggregations across projects.

Another metric--which we call he Relative Growth Index or RGI--controls for the heterogeneity of the comparison group and is thus preferable to the gap-reduction index for purposes of comparisons and aggregations.

To conclude this chapter (and the report) the authors would like to recommend that the growth of project participants always be measured using the gap-reduction model and the RGI metric. This recommendation applies equally to situations where nothing more can be done *and* to situations where evaluation designs enabling growth to be broken down into treatment-related and non-treatment-related components can be employed. Implementing this recommendation may seem like an unnecessary additional burden in evalution settings where internally valid estimates of treatment-related growth can be obtained. We believe, however, that the additional effort will pay significant dividends in the future by providing baseline data that will enhance the interpretability of growth measures in settings where such measures are all that can be obtained.

REFERENCES

Abbott, M. M. (1985, April). Theoretical considerations in the measurement of the English-language proficiency of limited-English-proficient students. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chica .

Adams, M., & Frith, J. (Eds.). (1979). Testing kit: French and Spanish. Washington, D.C.: U.S. Department of State, Foreign Service Institute.

Albert, M. L., & Obler, L. K. (1979). The bilingual brain. New York: Academic Press.

Alexander, L., Frankiewicz, R., Williams, K. E. Facilitation of learning and retention of oral instruction using advance and post organizers. Journal of Educational Psychology, 1979, 71.

Alkin, M., Kosecoff, J., Fitz-Gibbon, S.C., & Seligman, R. (1974). Evaluation and decision-making: The Title VII experience. CSE Monograph Series in Evaluation, No. 4. Los Angeles: University of California, Center for the Study of Evaluation.

Anderson, J. G., & Johnson, W. H. (1971). Stability and change among three generations of Mexican-Americans: Factors affecting achievement. American Educational Research Journal, 8(2), 285-309.

Anderson, L., Evertson, C., & Brophy, J. An experimental study of effective teaching in first grade reading groups. Elementary School J., 79 (4) 193-223.

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

Asher, J. & Price, B. (1969). The learning strategy of total physical response: Some age differences. Child Development, 38, 1219-1227.

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.

Bain, B. C. (1975). Toward an integration of Piaget and Vygotsky: Bilingual considerations. Linguistics, 160(5), 20.

Baker, K. A., & de Kanter, A. A. (1981). Effectiveness of bilingual education: A review of the literature. Washington, D.C.. 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. Lexington, MA: Lexington Press.

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

Baker, K. A., & Pelavin, S. (1985). Implications for evaluation models from bilingual evaluations. Washington, D.C.: Unpublished manuscript.

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

Balasubramonian, K. (1983, Winter). Not on test score alone. Bilingual Journal, 7(2) 17-21.

Balkan, L. (1970). Les essets du bilinguise francais-anglais sur les aptitudes intellectuelles. Bruxelles. AIMAV.

Baral, D. P. (1979). Academic achievement of recent immigrants from Mexico. Journal of the National Association for Bilingual Education, 3, 1-13.

Beckerman, T. M., & Good, T. L. (1981). The classroom ratio of high- and low-aptitude students and its effect on achievement. American Educational Research Journal, 18, 317-327.

Berman, P. (1978). Designing implementation to match policy situation: A contingency analysis of programmed and adaptive implementation. Santa Monica, CA: unpublished manuscript.

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

Berman, P., McLaughlin, M. W., Bass, G. V., Pauly, E., & Zellman, G. (1977). Federal programs supporting educational change, Vol. VII: Factors affecting implementation and continuation. Santa Monica, CA: Rand Corporation. (R-158917-HEW)

Berman, P., & Pauly, E. (1975). Federal programs supporting educational change, Vol. II: Factors affecting change agent projects. Santa Monica, CA: Rand Corporation.

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

Bhatnagar, J. (1980). Linguistic behavior and adjustment of immigrant children in French and English schools in Montreal. International Review of Applied Psychology, 29, 141-159.

Bilingual project evaluators: An overview. (1978, Fall). 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.

Bland, M., & Keisler, E. (1966). A self-controlled audio-lingual program for children. French Review, 40, 266-276.

Boruch, R. F. (1974, May). Regression-discontinuity designs: A summary. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Boruch, R. F. (1978). Comments of Tallmadge's paper. In M. J. Wargo & D. R. Green (Eds.), Achievement testing of disadvantaged and minority students for educational program evaluation. Monterey, CA: CTB/McGraw-Hill.

Boruch, R. F., & Cordray, D. S. (1980). An appraisal of educational program evaluations: Federal, state, and local agencies. Evanston, IL: Northwestern University. (U.S. Department of Education Contract No. 300-79-0467)

Boruch, R. F., & De Gracie, J. S. (1977, April). The use of regression-discontinuity models with criterion referenced testing in the evaluation of compensatory education. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Box, G. E. P., & Jenkins, G. M. (1970). Time-series analysis: Forecasting and control. San Francisco: Holden-Day.

Brislin, R. W. (Ed.). (1976). Translation: Application and research. New York: Gardner Press.

Brislin, R. W., Loner, W. J., & Thorndike, R. M. (1973). Cross-cultural research methods. New York: John Wiley & Sons.

Brookover, W. G., Lezotte, L., Brown, L., & Greenberg, L. Changes in school characteristics coincident with changes in student achievement. Michigan Department of Education, 1977.

Brookover, W., Schweitzer, J., Beady, C., Flood, P. and Wisenbaker, J., Elementary School Climate and School Achievement, College of Urban Development, Michigan State University, 1976.

Brophy, J. Reflections on research in elementary schools. Journal of Teacher Education, 1976, 27, (1), 31-34.

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

Brown, R., & Epstein, J. Interaction of achievement level and reinforcing properties of daily grading system. Education, 98, (2), 132-134.

Bryk, A. S., & Weisberg, H. I. (1976). Value-added analysis: A dynamic approach to the estimation of treatment effect. Journal of Educational Statistics, 1, 127-155.

Bryk, A. S., & Weisberg, H. I. (1976). Use of the nonequivalent control group design when subjects are growing. Psychological Bulletin, 85, 950-962.

Bryk, A. S., & Woods, E. (1980, December). An introduction to the value-added model and its use in short-term impact assessment. Cambridge, MA: The Huron Institute.

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). Los Angeles: California State University; Evaluation, Dissemination and Assessment Center.

Burry, J. (1982, April). Evaluation and documentation: Making them work together. Bilingual Education Paper Series, 5(9), Los Angeles: California State University; Evaluation, Dissemination and Assessment Center.

Burstall, C. (1975). Primary French in the balance. Educational Research, 17, 193-198.

Bye, T. (1977). Tests that measure language ability: A descriptive compilation. Berkeley: Bay Area Bilingual Education League.

Cabello, B. (1983). A description of analysis for the identification of potential sources of bias in dual language achievement tests. Journal of the National Association for Bilingual Education, 7(2), 33-51.

California State Department of Education. (1974). Studies on immersion education: A collection for United States educators. Sacramento: Author.

California State Department of Education. (1981). Schooling and language minority students: A theoretical framework. Los Angeles: California State University; Evaluation, Dissemination and Assessment Center.

California Test Bureau/McGraw-Hill. (1982). Criteria (No. 15). Monterey, CA: Author.

Campbell, D. T. (1963). From description to experimentation: Interpreting trends as quasi-experiments. In C. W. Harris (Ed.), Problems in measuring change. Madison, WI: University of Wisconsin Press.

Campbell, D. T. (1979). "Degrees of freedom" and the case study. In T. D. Cook & C. S. Reichardt (Eds.), Qualitative and quantitative methods in evaluation research. Beverly Hills, CA: Sage.

Campbell, D. T., & Borouch, R. F. (1975). Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), Evaluation and experience: Some critical issues in assessing social programs. New York: Academic Press.

Campbell, D. T., & Erlebacher, A. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), Compensatory education: A national debate. New York: Bruner/Mazel.

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

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.

Cantrell, R. P., Stenner, A., Jackson, & Katzenmeyer, W. G. Teacher knowledge, attitudes, and classroom teaching correlates of student achievement. Journal of Educational Psychology, 1977, 69, (2), 172-179.

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

Carey, S. T., & Cummins, J. (1985). Achievement, behavioral correlates and teachers' perceptions of Francophone and Anglophone immersion students. Alberta Journal of Educational Research, 29, 159-167.

Carrasco, R. L. (1981). Expanded awareness of student performance: A case study in applied ethnographic monitoring in a bilingual classroom. In H. T. Trueba, G. P. Guthrie, & K. H. Au (Eds.), Culture and the bilingual classroom. Rowley, MA: Newbury House.

Carter, L. F. (1984). The sustaining effects study of compensatory and elementary education. Educational Researcher, 13(7), 4-13.

Carter, T. P. (1970). Mexican Americans in school: A history of educational neglect. New Jersey: College Entrance Examination Board.

Carver, R. P. (1975). The Coleman Report: Using inappropriately designed achievement tests. American Educational Research Journal, 12(1), 77-86.

Cazden, C. B. (1985). Effective instructional practices in bilingual education. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Cazden, C. B., John, V. P., & Hymes, D. (Eds.). (1972). Functions of language in the classroom. New York: Teachers College Press.

Center for Applied Linguistics. (1977). Bilingual education: Current perspectives/law. Arlington, VA: Author.

Center for Educational Field Studies. (1970). An evaluation of a project for the analysis, development, implementation, and diffusion of the new social studies curricula. St. Louis, MO: Washington University, Author. (ERIC Document No. ED 054 996)

Center for the Study of Evaluation. (undated). Bilingual evaluation technical assistance workshop text IV: Planning for implementation evaluation. Los Angeles: University of California, Author.

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

Cervantes, R. A. (1979). Exemplary consafic chingatropic assessments: The AIR report. Bilingual Education Paper Series, 2(8), Los Angeles: California State University, National Dissemination and Assessment Center.

Chan, K. S., & Rueda, R. (1979). Poverty and culture in education: Separate but equal. Exceptional Children, 45, 422-428.

Chapman, D. W., & Carter, J. F. (1979). Translation procedures for the cross cultural use of measurement instruments. Educational Evaluation and Policy Analysis, 1(3), 71-76.

Charters, W. W., & Pellagrin, R. (1973). Barriers to the innovation process: Four case studies of differentiated staffing. Administrative Science Quarterly, 9(1), 3-14.

Chavez, E. L. (1982). Analysis of a Spanish translation of the Peabody Picture Vocabulary Test. Perceptual and Motor Skills, 54, 1335-1338.

Chesarek, S. (1981, March). Cognitive consequences of home or school education in a limited second language: A case study in the Crow Indian Community. Paper presented at the Language Proficiency Assessment Symposium, Airlie House, VA.

Chesterfield, R. P., Moll, L. C., & Perez, R. (1982, Fall). A naturalistic approach for evaluation. Bilingual Journal, 7(1), 23-26.

Cholewinski, M., & Holliday, S. (1979). Learning to read: What's right at home is right at school. Language Arts, 56(6), 671-680.

Christian, C. C. (1976). Social and psychological implications of bilingual literacy. In A. Simoes (Ed.), The bilingual child. New York: Academic Press.

Clayton, C. A., Drummond, D. J., Alexander, B. ., and Cameron, B. F. (1980). Validation of student counts used to allocate funds for the ESEA Title 1 migrant education program - technical report. Research Triangle Park, N.C.: Research Triangle Institute.

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

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

Cohen, E. G., Deal, T., Meyer, J., & Scott, W. R. (1979). Technology and teaming in the elementary school. Sociology of Education, 52, 20-33.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. Journal of Abnormal and Social Psychology, 65, 145-153.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev.). New York: Academic Press.

Cole, H. (1971). Implementation of a process curriculum by the campus team strategy. Syracuse, NY: Eastern Regional Institute for Education.

Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., & York, R. (1966). Equality of educational opportunity. Washington, D.C.: U.S. Government Printing Office.

Consalvo, R. W., & Orlandi, L. R. (1983). Principles and practices of data collection and management. Bilingual Journal, 7, 13-16.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally.

Cook, T. D., & Gruder, C. L. (1978). Metaevaluation 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.

Cooley, H. (1979). Multiple measures of second language acquisition among Hispanic children in a bilingual program. Doctoral dissertation, University of Wisconsin.

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, D.C.: Georgetown University Press.

Cordray, David S. (1986). Quasi-experimental analysis: a mixture of methods and judgment. In W. M. K. Trochim (Ed.). Advances in quasi-experimental design and analysis. San Francisco: Jossey-Bass, Inc.

Crawford, J., Brophy, J. E., Evertson, C. M., & Coulter, C. L. Classroom dyadic interaction: Factor structure of process variabies and achievement correlates. Journal of Educational Psychology, 1977, 69, 761-772.

Crespo, O. I., & Louque, P. (1984). Parent involvement in the education of minority language children. A resource handbook. Rosslyn, VA: InterAmerica Research Associates.

Cronbach, L. J. (1963). Evaluation for course improvement. Teachers College Record, 64, 672-683.

Cronbach, L. J. (1970). Essentials of psychological testing (3rd ed.). New York: Harper & Row.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., et al. (1980). Toward reform of program evaluation: Aims, methods, and institutional arrangements. San Francisco: Jossey-Bass.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change"--or should we? Psychological Bulletin, 74, 68-80.

Cronbach, L. J., & Suppes, P. (Eds.) (1969). Research for tomorrow's schools: Disciplined inquiry for education. New York: MacMillan.

Crowther, F. (1972). Factors affecting the rate of adoption of the 1971 Alberta social studies curriculum for elementary schools. Unpublished Master's thesis, University of Alberta.

Cummins, J. (1976). The influence of bilingualism on cognitive growth: A synthesis of research findings and explanatory hypotheses. Working Papers on Bilingualism, 9. Toronto: Ontario Institute for Studies in Education.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. Review of Educational Research, 49, 222-251.

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

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In California State Department of Education, Schooling and language minority students: A theoretical framework. (pp. 3-49). Los Angeles: California State University; Evaluation, Dissemination and Assessment Center.

Cummins, J. (1984). Bilingualism and special education: Issues in assessment and pedagogy. Clevedon, Avon, England: Multilingual Matters, Ltd.

Cummins, J. & Gulutsan, M. (1974). Some effects of bilingualism on cognitive functioning. In S. Carey (Ed.), Bilingualism, biculturalism and education. Edmonton: The University of Alberta Press.

Cummins, J., & Mulcahy, R. (1978). Orientation to language in Ukrainian-English bilingual children. Child Development, 49, 1239-1242.

Cummins, J. (1983). Conceptual and linguistic foundations of language assessment. In S. Seidner (Ed.). Issues of language assessment, Volume II: Language assessment and curriculum planning. Illinois: Illinois State Board of Education.

Danoff, M. N. (1978). Evaluation of the impact of ESEA Title VII Spanish/English bilingual education program. Overview of the study and findings. Palo Alto, CA: American Institutes for Research.

Darcy, N. T. (1953). A review of the literature on the effects of bilingualism upon the measurement of intelligence. Journal of Genetic Psychology, 82, 21-57.

De Avila, E. A., & Havassy, B. E. (1974). The testing of minority children: A neo-Piagetian approach. Today's Education, 63, 71-75.

De Avila, E. H. (1981, February). Improving cognition: A multi-cultural approach. Stanford, CA: Stanford University School of Education. (Final report: National Institute of Education Grant No. NIE-G-78-0158)

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

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

De Gracie, J. S., & Fuller, W. A. (1972). Estimation of the slope and analysis of covariance when the concomitant variable is measured with error. Journal of the American Statistical Association, 67, 930-937.

De Mauro, G. E. (1983, Winter). Models and assumptions for bilingual education evalution. Bilingual Journal, 7(2), 8-12.

De Mauro, G. (1985, April). Issues in the placement of limited-English-proficient students and in evaluation of their instructional programs. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Dervin, B., & Greenberg, B. (1972). The communications environment of the urban poor. In G. Kline & P. Tichenor (Eds.), Current perspectives in mass communications research. Berkeley, CA: Sage Publications.

Deutsch, C. P. (1973). Social class and child development. In B. M. Caldwell & H. N. Ricciuti (Eds.), Review of child development research (Vol. 3). Chicago: The University of Chicago Press.

Development Associates, Inc. (1974). A process evaluation of the Bilingual Education Program, Title VII, Elementary and Secondary Education Act. Washington, D.C.: Author.

Development Associates, Inc. (1979). Evaluation of California's educational services to limited and non-English-speaking students. Executive summary, interim reports 3 and 4. San Francisco: Author.

Development Associates, Inc. (1983). Request for forms clearance with supporting statement for national longitudinal evaluation of services for language minority limited-English proficient students. Arlington, VA: Author.

Diebold, A. R. (1968). The consequences of early bilingualism in cognitive development and personality formation. In E. Norbeck, D. Price-Williams, & W. M. McCord (Eds.). The study of personality. New York: Holt, Reinhart & Winston.

Dominguez, D., Tunmer, W. E., & Jackson, S. L. (1980, April). Measuring degree of implementation of bilingual education programs: Implications for staff development and program evaluation. Paper presented at the Annual Conference of the National Association for Bilingual Education, Anaheim, CA.

Douglas, D., & Johnson, D. M. (1981). An evaluation of Title VII evaluations: Results from a national study. Paper presented at the Annual Conference of the National Association for Bilingual Education, Boston.

L. Downey Research Associates. (1975). The social studies in Alberta-1975. Edmonton, Alberta: Author.

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

Duncan, S. E., & De Avila, E. A. (1979). Bilingualism and cognition: Some recent findings. Journal of the National Association for Bilingual Education, 4, 15-50.

Ekstrand, L. H. (1979). Replacing the critical period and optimal age theories of second language acquisition with a theory of ontogenetic development beyond puberty. In Educational and psychological interactions, Lund University, Malmo School of Education, Sweden.

Elman, A. (1981). Quality control in Title I: Manual versus computer conversions of test scores. Palo Alto, CA: American Institutes for Research.

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

Ervin-Tripp, S. (1974). Is second language learning like the first? TESOL Quarterly, 8(2), 111-127.

Evaluation, Dissemination and Assessment Center. (1976). Evaluation instruments for bilingual education: An annotated bibliography. Austin, TX: Author.

Evaluation, Dissemination and Assessment Center. ((1982). Bilingual program planning, implementation, and evaluation: Foundations for evaluating bilingual programs. Austin, TX: Author.

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 Title VII projects in bilingual education. Austin, TX: Author.

Fantini, M. D. (1970). Community control and quality education in urban school systems. In H. M. Levin (Ed.), Community control of schools. Washington, D.C.: Brookings Institute.

Fathman, A. (1975). The relationship between age and second language learning productive ability. Language Learning, 25(2).

Fisher, C. W., N. N. Filby, R. S. Marliave, L. A. Chaen, M. M. Dishaw, J. Moore, & D. C. Berliner (1978). Teaching behaviors, academic learning time and student achievement: Final report of Phase III-B, beginning teacher evaluation study. San Francisco: Far West Laboratory for Educational Research and Development.

Fisher, C. W. (1983). Significant bilingual instructional features: Final recommendation report. San Francisco: Far West Laboratory for Educational Research and Development.

Fisher, C. W., & Guthrie, L. F. (1983). Significant bilingual instructional features: Executive summary. San Francisco: Far West Laboratory for Educational Research and Development.

Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.), Educational measurement. Washington, D.C.: American Council on Education.

Frederick, W. C., Walbert, H. J., & Rasher, S. P. Time, teacher comments and achievement in urban high schools. Journal of Educational Research, 1979, 73, (2), 63-65.

Freeman, E. B. (1982). The Ann Arbor decision: The importance of teachers' attitudes towards language. The Elementary School Journal, 83, 41-47.

Fullan, M., & Pomfret, A. (1977). Research on curriculum and instruction implementation. Review of Educational Research, 47(1), 335-397.

Gallimore, R., Bogg , J., & Jordan, C. (1974). Culture, behavior, and education. Beverly Hills: Sage Publications.

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

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

Genesee, F. (1978). A longitudinal evaluation of an early immersion school program. Canadian Journal of Education, 3(4), 31-50.

Genesee, F. (1984). Historical and theoretical foundation of immersion education. In Studies of immersion education: A collection for United States educators. Sacramento: California State Department of Education, 32-57.

Genesee, F. (in press). Second language learning through immersion: A review of U.S. programs. Review of Educational Research.

Genesee, F., & Lambert, W. (1983). Trilingual education for majority language children. Child Development, 54, 105-114.

Genesee, F., Polich, E., & Stanley, M. (1977). An experimental French immersion program at the secondary school level: 1969-1974. Canadian Modern Language Review, 33, 318-332.

Genesee, F., Tucker, G. R., & Lambert, W. (1975). Communication skills of bilingual children. Child Development, 46, 1013-1018.

Georgetown BESC and BUENO BEMSC Services. (1985). Forum, 8(5), 5.

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

Gillmore, D., & Dickerson, A. D. (1980). The relationship between instruments used for identifying children of limited-English-proficiency in Texas. Bilingual Resources, 3(3), 16-29.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5(10), 3-8.

Glass, G. V. (1980). When educators set standards. In L. E. Baker & E. S. Quallmalz (Eds.), Educational testing and evaluation: Design, analysis, and policy. Beverly Hills, CA: Sage.

Glass, G. V., Wilson, V. L., & Gottman, J. M. (1975). Design and analysis of time-series experiments. Boulder, CO: Colorado Associated University Press.

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

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

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

Gonzalez, J. M. (1979). Coming of age in bilingual/bicultural education: A historical perspective. In H. T. Trueba & C. Barnett-Mizrahi (Eds.), Bilingual multicultural education and the professional. Rowley, MA: Newbury House.

Good, T. L., Ebmeier, H., & Beckerman, T. Teaching mathematics in high and low SES classrooms: An empirical comparison. Journal of Teacher Education, 1978, 29, (5), 85-90.

Good, T. L., & Grouws, D. Teaching effects: A process/product study in fourth grade math classes. Journal of Teacher Education, 1977, 28, (3), 49-54.

Good, T. L., & Grouws, D. The Missouri mathematics effectiveness project: An experimental study in fourth grade classrooms. Journal of Educational Psychology, 1979, 71, (3), 355-362.

Good, T. L. (1981). Teacher expectations and student perceptions: A decade of research. Educational Leadership, 38, 415-423.

Good, T. L. (1982). How teachers' expectations affect results. American Education, 18(10), 25-32.

Gordon, I. (1978). Parent and community involvement in compensatory education. Urbana, IL: University of Illinois.

Green, D. R. (1983, April). Content validity of standardized achievement tests and test curriculum overlap. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal.

Gross, N. C., Giacquinta, J. B., & Bernstein, M. (1971). Implementing organizational innovations: A sociological analysis of planned educational change. New York: Basic Books.

Hakuta, K. (1984). Causal relationship between the development of bilingualism, cognitive flexibility, and social-cognitive skills in Hispanic elementary school children. New Haven, CT.

Hakuta, K., & Diaz, R. M. (1983). The relationship between degree of bilingualism and cognitive ability: A critical discussion and some new longitudinal data. In K. E. Nelson (Ed.), Children's language. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hall, G. E., & Louchs, S. F. (1977, Summer). A developmental model for determining whether the treatment is actually implemented. American Education Research Journal, 14(3), 263-276.

Hall, G. E., & Louchs, S. (1978). Teacher concerns as a basis for facilitating and personalizing staff development. Teachers College Record, Vol. 80, No. 1.

Hamingson, D. (Ed.). (1973). Toward judgment. Norwich, England: University of East Anglia.

Hansen, J. C., & Fouad, N. A. (1984). Translation and validation of the Spanish form of the Strong-Campbell Interest Inventory. Measurement and Evaluation in Guidance, 16(4), 192-197.

Hanson, R. A., Schutz, R. E., & Bailey, J. D. (undated). What makes achievement tests tick: Investigation of alternative instrumentation for instructional program evaluation. Los Alamitos, CA: Southwest Regional Laboratory for Educational Research and Development.

Hazen, M. (1980). An argument in favor of multimethod research and evaluation in CAI and CMI instruction. Association for Educational Data System Journal, 14(4), 275-284.

Henderson, A. (Ed.). (1981). Parent participation--student achievement: The evidence grows. Columbus, MD: National Committee for Citizens in Education.

Hess, R. D. (1970). Social class and ethnic influence on socialization. In P. Mussen (Ed.), Carmichael's manual of child psychology. New York: John Wiley.

Hess, R., & Buckholdt, D. (1974). Degree of implementation as a critical variable in program evaluation. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Hestand, D. (1973). Strategies and procedures used, and problems encountered in implementing differentiated staffing: A case study. Unpublished doctoral dissertation, University of Houston.

Heyman, A. G. (1973). Invandrarbarn, Studiehandledning. Stockholms invandrarnamnd.

Hills, J. R. (1984). Interpreting NCE scores. Educational Measurement: Issues and Practices, 3(3), pp. 25-26, 31.

Hirata, L. C. (1975). Youth, parents, and teachers in Chinatown. Urban Education, 10, 279-296.

Holt, D. D., & Arellano, J. A. (1980). Federal and state legal cases for bilingual education. In Bilingual program, policy, and assessment issues. Sacramento, CA: California State Department of Education.

Holtzman, W. (1980). Bilingual program outcomes. Paper presented at the regional conference R & D Speaks: Bilingual/Multicultural Education, Austin, TX.

Hoover, T., & Kamm, M. (1981). A guide to processing student information. Bilingual Journal, 6, 16-19.

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

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. Mountain View, CA: RMC Research Corporation. (RMC Report No. UR-460)

House, E. (1975). The politics of educational innovation. Berkeley, CA: McCutchan.

Hubert, J. A. (1981). Impact evaluation in bilingual and compensatory education programs with high retention rates. New England Educational Research Organization Annual Best Paper Monograph, 11-22.

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

Hubert, J. A. (in press). Improving the evaluation of bilingual education programs through public policy. In H. LaFontaine, B. Persky, & Golubchick (Eds.), Bilingual education.

Hurwitz, N. (1975). Communications networks and the urban poor. Equal Opportunity Review.

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

InterAmerica Research Associates, Inc. (1979). Development of evaluation models for ESEA Title VII bilingual education projects. Technical proposal in response to RFP 79-117. Rosslyn, VA: Author.

InterAmerica Research Associates, Inc. (1985). The 1984 Bilingual Education Act. Rosslyn, VA: Author.

Ironson, G. H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. Journal of Educational Measurement, 16(4), 209-225.

Izzo, S. (1981). Second language learning: A review of related studies. Rosslyn, VA: National Clearinghouse for Bilingual Education.

Jencks, C., Smith, M. S., Acland, H., Bane, M., Cohen, D., Gintis, H., Heynes, B., & Michelson, S. (1972). Inequality: A reassessment of the effect of family and schooling in America. New York: Basic Books.

Jense, J. V. (1962a). Effects of early childhood bilingualism I. Elementary English, 39, 132-143.

Jensen, J. V. (1962b). Effects of early childhood bilingualism II. Elementary English, 39, 358-366.

Johnson, D. W., & Johnson, R. T. Instructional goal structure: Cooperative, competitive, or individualistic. Review of Educational Research, 1974, 44, 213-240.

Johnson, D. W., Johnson, R. T., & Scott, L. The effects of co-operative and individualized instruction on student attitudes and achievement. Journal of Social Psychology, 1978, 104, 207-216.

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

Jolly, S. J., & Gramenz, G. W. (1984). Customizing a norm-referenced achievement test to achieve curricular validity: A case study. Educational Measurement: Issues and Practices, 3(3), 16-18.

Judd, C. M., & Kenny, D. A. (1981). Estimating the effects of social interventions. London: Cambridge University Press.

Kaskowitz, D. H. (1982). An examination of differences in gain estimates across testing cycles. Mt. View, CA: RMC Research Corp.

Kaskowitz, D. H., Binkley, J. L., & Johnson, D. M. (1981). A study of teacher training programs in bilingual education. Volume II: The supply and demand for bilingual education teachers. Mountain View, CA: RMC Research Corp.

Katz, I. (1967). Some motivational determinants of racial differences in intellectual achievement. International Journal of Psychology, 2, 1-12.

Keesling, J. W. (1984). Differences between fall-to-spring and annual gains in evaluating Chapter 1 programs. Oxnard, CA: Advanced Technology.

Kenny, D. A. (1975). A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. Psychological Bulletin, 82(3), 345-362.

Kerr, D. M., Kent, L., & Lam, T. C. M. (1985). Measuring program implementation with a classroom observation instrument: The Interactive Teaching Map. Evaluation Review, 9(4), 461-482.

Kessler, C., & Quinn, M. E. (1980). Bilingualism and science problem-solving ability. Bilingual Education Paper Series, 4(1). Los Angeles: California State University; Evaluation, Dissemination and Assessment Center.

Kimball, W. L. (1968). Parental and family influences on academic achievement among Mexican-American students. Dissertation Abstracts International, 29, 1965A. (University Microfilms No. 68-16, 550).

Klienfeld, J. S. (1979). Eskimo school on the Andreafsky. New York: Praeger Publications.

Krashen, S., Long, M. A., & Scarcella, R. C. (1979). Age, rate, and eventual attainment in second language acquisition. TESOL Quarterly, 13(4), 573-582.

Krashen, S. D. (1981). Bilingual education and second language acquisition theory. In Office of Bilingual-Bicultural Education, California State Department of Education (Ed.), Schooling and Language Minority Students: A Theoretical Framework. Los Angeles: California State University, Evaluation, Dissemination, and Assessment Center.

Krashen, S. D. (1982). Principles and Practice in Second Language Acquisition. Oxford, England: Pergamon Press, Ltd.

Lambert, W. E. (1975). Culture and languages as factors in learning and education. In A. Wolfgang (Ed.), Education of immigrant students. Toronto: Ontario Institute for Studies in Education.

Lambert, W., & Tucker, G. R. (1972). Bilingual education of children: The St. Lambert experiment. Rowley, MA: Newbury House Publishers.

Lambert, W. E. The effects of bilingualism on the individual: cognitive and sociocultural consequences. In Bilingualism: Psychological, social, and educational implications, edited by P.A. Hornby. New York, Academic Press, 1977.

Language Proficiency Instrument Review Committee. (1982, January). Summary review of the Language Proficiency Instrument Review Committee: BINL, BSM I, BSM II, IDEA Proficiency Test (IPT), LAS I, LAS II. Sacramento, CA: Author.

Laosa, L. M. (1975). Bilingualism in three United States Hispanic groups: Contextual use of language by children and adults in their families. Journal of Educational Psychology, 67(5), 617-627.

Laosa, L. M. (1977). Inequality in the classroom. Observational research on teacher-student interactions. Aztlan International Journal of Chicano Studies Research, 8, 51-67.

Laosa, L. M. (1982). The sociocultural context of evaluation. In B. Spodek (Ed.), Handbook of research in early childhood education. New York: The Free Press, 501-520.

Laosa, L. M. (1985). Letter to G. Kasten Tallmadge dated October 25, 1985.

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

Law, A. I. (1978, September). Proceedings of the Bilingual Instrument Review Committee (AB 3470). Sacramento: California State Department of Education, Office of Program Evaluation and Research.

Lawton, J. and Powell, N. Effects of advanced organizers on preschool children's learning of math concepts. Journal of Experimental Education, 1979, 47, (1), 76-81.

Lee, B. (1985, Spring). Statistical conclusion validity in ex post facto designs: Practicality in evaluation. Educational Evaluation and Policy Analysis, 7(1), 35-45.

Lega, L. I. (1981). A Colombian version of the Children's Embedded Figures Test. Hispanic Journal of Behavioral Sciences, 3(4), 415-417.

Legarreta-Marcaida, D. (1981). Effective use of the primary language in the classroom. In Schooling and language minority students: A theoretical framework. Los Angeles: California State University; Evaluation, Dissemination and Assessment Center.

Leibowitz, A. H. (1982). Federal recognition of the rights of minority language groups. Rosslyn, VA.: InterAmerica Research Associates, Inc.

Leinhardt, G., & Seewald, A. M. (1981). Overlap: What's tested, what's taught? Journal of Educational Measurement, 18(2), 85-94.

Levin, H. M. (1970). Community control of schools. Washington, D. C.: Brookings, Institute.

Levin, H. M. (1975). The staff of life. Paper presented at the Annual Convention of the American Psychological Association, New York.

Levin, J. R. Inducing comprehension in poor readers: A test of a recent model. Journal of Educational Psychology, 1973, 65, (1), 19-24.

Liedke, W. W., & Nelson, L. D. (1968). Concept formation and bilingualism. Alberta Journal of Educational Research, 14, 225-232.

Linn, R. L. (1982). The validity of the Title I evaluation and reporting system. In E. R. Reisner, M. C. Alkin, R. F. Boruch, R. L. Linn, & J. Millman (Eds.), Assessment of the Title I evaluation and reporting system. Washington, D.C.: U.S. Department of Education.

Lockheed, M. E., & Harris, A. M. (1984). Cross-sex collaborative learning in elementary classrooms. American Educational Research Journal, 21(2), 275-294.

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

Lord, F. M. (1967). Elementary models for measuring change. In C. W. Harris (Ed.), Problems in measuring change. Madison, WI: University of Wisconsin Press.

Loret, P. G., Sedar, A., Bianchini, J. C., & Vale, C. A. (1974). Anchor test study: Equivalence and norms tables for selecting reading achievement tests (grades 4, 5, 6). Washington, D.C.: U.S. Government Printing Office.

Lucker, G. W., Rosenfield, D., Sikes, J., & Aronson, E. Performance in the interdependent classroom: A field study. American Educational Research, 13, (2), 115-123.

Macnamara, J. (1966). Bilingual and primary education. Edinburgh: Edinburgh University Press.

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

Marascuilo, L. A., & Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias on Chi Square statistics. Journal of Educational Measurement, 18(4), 229-248.

March, J. G., & Simon, H. A. (1958). Organizations. New York: John Wiley & Sons.

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

McCain, L. J., & McCleary, R. (1979). The statistical analysis of the simple interrupted time series quasi-experiment. In T. D. Cook & D. T. Campbell (Eds.), Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally.

McCauley, D. E., & Colberg, M. (1983). Transportability of deductive measurement across cultures. Journal of Educational Measurement, 20(1), 81-92.

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. San Francisco: Jossey-Bass.

McConnell, B. B. (1983). Needed: A new technology for evaluating bilingual education programs. Bilingual Education Paper Series, 7(1). Los Angeles: California State University; Evaluation, Dissemination, and Assessment Center.

McConnell, B. B. (1985). Memorandum to G. K. Tallmadge dated November 4, 1985.

Mehrens, W. A. (1984). National tests and local curriculum: Match or mismatch? Educational Measurement: Issues and Practices, 3(3), 9-15.

Mercer, J. R. (1977). Implications of current assessment procedures for Mexican-American children. Bilingual Education Paper Series, 1(1). Los Angeles: California State University; Evaluation Dissemination and Assistance Center.

Mercer, J. R., Gomez-Palacio, M., & Padilla, E. (in press). The development of practical intelligence in cross-cultural perspective. In R. J. Sternberg & R. K. Wagner (Eds.), Practical intelligence: Origins of competence in the everyday world. Cambridge: Cambridge University Press.

Merino, B. J., Politzer, R. L., & Ramirez, A. G. (1979). The relationship of teachers' Spanish proficiency to pupils' achievement. Journal of the National Association for Bilingual Education, 3(1), 21-38.

Merino, B. J., & Spencer, M. (1983). The comparability of English and Spanish versions of oral language proficiency instruments. Journal of the National Association for Bilingual Education, 7(2), 1-31.

Miller, T., & Dhand, H. (1973). The classroom teacher as curriculum developer for project Canada West. Canada: Saskatchewan Teachers' Foundation.

Millman, J. (1975). Selecting educational researchers and evaluators. Princeton, NJ: ERIC Clearinghouse on Tests, Measurement, and Evalution, TM Report 48.

Mohanty, A. K. (1982). Bilingualism among Kond tribals in Orissa (India). Consequences, issues, and implications. Uktal University, unpublished research report.

Molina, H., & Shoemaker, D. M. (1973). A preliminary evaluation of a bilingual Spanish/English program using multiple matrix sampling. Paper presented at the International Seminar on Language Testing, San Juan, P.R.

Moll, L.C. (1981). The microethnographic study of bilingual schooling. In R. V. Padilla (Ed.), Ethnoperspectives in bilingual education research: Bilingual education technology. Tempe, AZ: Arizona State University.

Moore, F. B., & Parr, G. D. (1979). Models of bilingual education: Comparisons of effectiveness. Elementary School Journal, 79(2), 93-97.

Morrison, F. et al. (1979). French proficiency status of Ottawa and Carleton students in alternative programs: Evaluation of the second language learning (French) programs in the schools of the Ottawa and Carleton Board of Education. Sixth annual report. Toronto: Ministry of Education.

Mougeon, R., & Canale, M. (1978-79). Maintenance of French in Ontario: Is education in French enough? Interchange, 9, 30-39.

National Assessment of Educational Progress. (1983). Students from homes in which English is not the dominant language: Who are they and how well do they read? (No. 11-R-50). Denver, CO: Education Commission of the States.

National Center for Bilingual Research. (1982). Synthesis of reported evaluation and research evidence on the effectiveness of bilingual education basic projects (Vol. 1). Los Alamitos, CA: Author.

National Clearinghouse for Bilingual Education. (1983). Information and technical assistance needs of Title VII demonstration projects: Validation and evaluation. Washington, D.C.: 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.

O'Brien, M. L. (1985, April). Psychometric issues relevant to selecting items and assembling parallel forms of language-proficiency instruments. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Ogbu, J. U. (1978). Minority education and caste. New York: Academic Press.

Okada, M., Besel, R., Bachelor, P., Glass, G. V., & Montoya-Tannatt, L. (1983). Syntheses 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. Washington, D.C.: Georgetown University Press.

Oller, J. W., & Nagato, N. (1974). The long term effect of FLES. The Modern Language Journal, 58(1-2), 15-19.

Olson, L., & Samuels, S. (1973). The relationship between age and accuracy of foreign language pronunciation. Journal of Educational Research, 66(6), 263-268.

O'Malley, J. M. (1984). Options for improving local evaluations. Rosslyn, VA: National Clearinghouse for Bilingual Education.

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

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

Overall, J. E, & Woodward, J. A. (1977). Common misconceptions concerning the analysis of covariance. Journal of Multivariate Behavioral Research, 12, 171-185.

Oxford, R., Pol, L., Lopez, D., Stupp, P., Gendell, M., & Peng, S. (1981). Projections of non-English language background and limited English proficient persons in the United States to the year 2000: Educational planning in the demographic context. Journal of the National Association for Bilingual Education, 5(3), 1-30.

Oyama, S. (1976). A sensitive period for the acquisition of a non-native phonological system. Journal of Psycholinguistic Research, 5, 261-285.

Parsons, H. M. (1974). What happened at Hawthorne? Science, 193, 922-932.

Patton, M. Q. (1979). Evaluation of program implementation. In Evaluation studies review annual (Vol. 4). Beverly Hills, CA: Sage Publications.

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

Peal, E., & Lambert, W. E. (1962). The relation of bilingualism to intelligence. Psychological Monographs, 76, 546.

Pedhazur, E. J. (1982). Multiple regression in behavioral research: Explanation and prediction (2nd ed.). New York: Holt, Rinehart & Winston.

Peleg, Z. R. (1978). Impact assessment in the evaluation of bilingual programs: It is feasible? Educational Technology, 18, 19-23.

Penaloza-Stromquist, N. (1980). Teaching effectiveness and student achievement in reading in Spanish. The Bilingual Review, 7(2), 95-104.

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

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

Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. Educational and Psychological Measurement, 40, 397-404.

Pletcher, B. P., Locks, N. A., Reynolds, D. F., & Sission, B. G. (1978). A guide to assessment of limited English speaking students. New York: Santillana Publishing Company, Inc.

Politzer, R. L., & Hoover, M. R. (1976). Teachers' and pupils' attitudes toward black English speech varieties and black pupils' achievement. Stanford, CA: Stanford University, Center for Research and Development in Teaching.

Popham, W. J. (1975). Educational evaluation. Englewood Cliffs, NJ: Prentice-Hall.

Porter, A. C. (1968). The effects of using fallible variables in the analysis of covariance. Dissertation Abstracts International, 28, 3517B. (University Microfilms No. 67-12, 147, 144)

Pottinger, J. S. (1970, May). Identification of discrimination and denial of services on the basis of national origin. Memorandum, U.S. Department of Health, Education and Welfare; Washington, D.C.

Provus, M. (1971). Discrepancy evaluation. Berkeley, CA: McCutchan.

Purkey, S. C., & Smith, M. S. (1983). Effective school: A review. The Elementary School Journal, 83, 427-452.

Ramirez, A. G. (1978). Teaching reading in Spanish: A study of teacher effectiveness. Stanford, CA: Stanford University, Center for Educational Research.

Ramirez, A. G., Arce-Torres, E., & Politzer, R. L. (1976). Language attitudes and the achievement of bilingual pupils (Research and Development memo No. 146). Stanford, CA: Stanford University, Center for Research and Development in Teaching.

Ramirez, A. G., & Politzer, R. L. (1975). The acquisition of English and the maintenance of Spanish in a bilingual education program. TESOL Quarterly, 9(2), 113-124.

Ramirez, A. G., & Stromquist, N. (1979). ESL methodology and student language learning in bilingual elementary schools. TESOL Quarterly, 8(2), 145-158.

Ramirez, J. D., Merino, B. J., Bye, T. T., & Gold, N. C. (1982). Assessment of oral English proficiency: A status report. Paper submitted to the Eighth International Conference on Testing, University of Edinburgh, Scotland.

Ramirez, J. D., Wolfson, R., Tallmadge, G. K., & Merino, B. (1984). Study design of the longitudinal study of immersion programs for language-minority children. Mountain View, CA: SRA Technologies, Inc.

Raven, J. C. (1940). Matrix tests. Mental Health, 1, 10-18.

Reichardt, C. S. (1979a). The design and analysis of the nonequivalent group quasi-experiment. Unpublished doctoral dissertation, Northwestern University.

Reichardt, C. S. (1979b). The statistical analysis of data from nonequivalent group designs. In T. D. Cook & D. T. Campbell (Eds.), Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally.

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

Rivera, C., & Simich, C. (1981). Language proficiency assessment: Research findings and their application. In S. S. Seidner (Ed.), Issues of language assessment: Foundations and research. Springfield, IL: Illinois State Board of Education.

Rodriguez, A. M. (1980). Empirically defining competencies for effective bilingual teachers: A preliminary study. Bilingual Education Paper Series, 3(12). Los Angeles: California State University; Evaluation, Dissemination and Assessment Center.

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

Rodriguez-Brown, F. V, (1980). Do's and don'ts of bilingual program evaluation. Bilingual Education Paper Series. 3(6). Los Angeles: California State University; Evaluation, Dissemination and Assessment Center.

Rogosa, D. R. (1980). A critique of cross-lagged correlation. Psychological Bulletin, 88(2), 245-258.

Rogosa, D. R., & Willet, J. B. (1983). Documenting the reliability of the difference score in the measurement of change. Journal of Educational Measurement, 20(4), 335-343.

Rose, C., & Nyre, G. F. (1977). The practice of evaluation. ERIC/TM Report 65. Princeton, NJ: Educational Testing Service.

Rosenbluth, A. R. (1976). The feasibility of test translation between unrelated languages: English to Navajo. TESOL Quarterly, 10(1), 33-43.

Rosenfeld, G. (1971). Shut those thick lips! A study of slum school failure. New York: Holt, Rinehart, & Winston.

Rosenshine, B. Recent research on teacher behaviors and student achievement. Journal of Teacher Education, 1976, 27, (1), 61-64.

Rosenthal, A. S., Milne, A. M., Ellman, F. M., Ginsburg, A. L., & Baker, K. A. (1983). A comparison of the effects of language background and socioeconomic status on achievement among elementary-school students. In K. A. Baker & A. A. de Kanter (Eds.), Bilingual education. Lexington, MA: Lexington Press.

Rosenthal, A. S., Milne, A. M., Ginsburg, A. L., & Baker, K. A. (1981). A comparison of the effects of language background and socioeconomic status on achievement status among elementary school students. Washington, D.C.: U.S. Department of Education.

Rosier, P., & Holm, W. (1980). The Rock Point experience: A longitudinal study of a Navajo school (Saad Naaki Bee Na'nitin). Bilingual Education Series, 8. Arlington, VA: Center for Applied Linguistics.

Roudabush, G. E. (1978). Program evaluation using criterion-referenced tests. In M. J. Wargo & D. R. Green (Eds.), Achievement testing of disadvantaged and minority students for educational program evaluation. Monterey, CA: McGraw-Hill.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased item detection techniques. Journal of Educational Statistics, 3(2), 213-233.

Rutter, M., Maughan, B., Mortimore, P., & Ouston, J. Fifteen Thousand Hours, Cambridge, MA: Harvard University Press, 1979.

Santiago, M., & de Guzman, E. (1977). A child's step forward in reading: The effect of language of materials and other factors on reading comprehension among grade four pupils. Research series. Philippine Normal College.

Santiago, R. (1985). Academic success for LEP students in American schools: Putting education back into bilingual education. Presented at the Second Annual LAU Center Conference, Columbus, OH.

Saretsky, G. (1972). The OEO P.C. experiment and the John Henry effect. Phi Delta Kappan, 53, 579-581.

Scheuneman, J. (1979). A method of detecting bias in test items. Journal of Educational Measurement, 16(3), 143-152.

Schimmel, D., & Fisher, L. (1977). The rights of parents in the education of their children. Columbia, MD: National Committee for Citizens in Education.

Schinke-Llano, L. (1984). Programmatic and instructional aspects of language immersion programs. Mountain View, CA: SRA Technologies.

Scott, S. (1973). The relation of divergent thinking to bilingualism: Cause or effect? Unpublished research report, McGill University.

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

Sechrest, L., West, S. G., Phillips, M. A., Redner, R., & Yeaton, W. H. (1979). Some neglected problems in evaluation research: Strength and integrity of treatments. In Evaluating studies review annual (Vol. 4). Beverly Hills, CA: Sage Publications.

Sechrest, L., & Yeaton, W. H. (1982). Magnitudes of experimental effects in social science research. Evaluation Review, 6(5), 579-601.

Seidner, S. (1981). Political expedience or educational research: An analysis of Baker and de Kanter's review of the literature of bilingual education. Rosslyn, VA: National Clearinghouse for Bilingual Education.

Shafer, R. (1978). Home learned language functions: How they assist beginning reading. Pape. presented at the Ninth World Congress on Sociology, Uppsala, Sweden.

Shaycoft, M. F. (1979). Handbook of criterion-referenced testing. New York: Garland STPM Press.

Shipman, J., & Bussis, A. (1968). The impact of the family. In Disadvantaged children and their first school experience. ETS-OEO longitudinal study. Princeton, NJ: Educational Testing Service.

Shipman, M. (1974). Inside a curriculum project. London: Methuen & Co.

Silverman, R. J., Noa, J. K., & Russell, R. H. (1977). Oral language tests for bilingual students: An evaluation of language dominance and proficiency instruments. Portland, OR: Northwest Regional Educational Laboratory.

Skutnabb-Tangas, T., & Toukomaa, P. (1976). Teaching migrant children's mother tongue and learning the language of the host country in the context of the socio-cultural situation of the migrant family. Helsinki: Finnish National Commission for UNESCO.

Slavin, R. E. Student teams and comparisons among equals: effects on academic performance and student attitudes. Journal of Educational Psychology, 1978, 70, (4), 532-538.

Snow, C., & Hoefnagel-Hohle, M. (1978). Age differences in second language acquisition. In E. Hatch (Ed.), Second language acquisition. Roley, MA: Newbury House.

So, A. Y., & Chan, K. S. (1984). What matters? The relative impact of language background and socioeconomic status on reading achievement. Journal of the National Association for Bilingual Education, 8(3), 27-41.

Solomon, R. L. (1949). An extension of control group design. Psychological Bulletin, 46, 137-150.

Solomon, W., Ferritor, D., Hearn, J., & Myers, E. (undated). The development, use, and importance of instruments that validly and reliably assess the degree to which experimental programs are implemented. St. Louis: CEMREL, Inc.

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

Sorbom, D. (1978). An alternative to the methodology for analysis of covariance. Psychometrica, 43, 381-396.

Sorbom, D., & Joreskog, K. G. (1981, June). The use of structural equation models in evaluation research. Paper presented at the Conference on Experimental Research in the Social Sciences, Gainesville, FL.

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.

Spolsky, B., & Cooper, R. (Eds.). (1977). Frontiers of bilingual education. Rowley, MA: Newbury House.

Stallings, J. A. How instructional processes relate to child outcomes in a national study of Follow Through. Journal of Teacher Education, 1976, 27, (1), 43-47.

Stearns, M. S., Peterson, S., Robinson, M., & Rosenfeld, A. (1973). Parent involvement in compensatory education programs: Definitions and findings. Menlo Park, CA: SRI International.

Stewart, B. L. (1980). Investigating the technical adequacy of Model C in Title I evaluation. Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

Stoll, L. J. Reading Program Administration: Does it make a difference? Administrator's Notebook, 1978-79, 27, (3).

Swain, M. (1980). French immersion programs in Canada. Multi-culturalism, 4(2), 3-6.

Swain, M. (1984). A review of immersion education in Canada: Research and evaluation studies. In Studies on immersion education: A collection for United States educators. Sacramento: California State Department of Education.

Sween, J. A. (1971). The experimental regression design: An investigation into the feasibility of nonrandom treatment allocation. Unpublished doctoral dissertation, Northwestern University.

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

Tallmadge, G. K. (1982). An empirical assessment of norm-referenced evaluation methodology. Journal of Educational Measurement, 19(2), 97-112.

Tallmadge, G. K. (1985). Rumors regarding the death of the equipercentile assumption may have been greatly exaggerated. Journal of Educational Measurement, 22(1), 33-39.

Tallmadge, G. K., & Horst, D. P. (1976). A procedural guide for validating achievement gains in educational projects (revised). Washington, D.C.: U.S. Government Printing Office. (Stock No. 017-080-01516)

Tallmadge, G. K., Wood, C. T., & Gamel, N. N. (1981). Users' guide: ESEA Title I evaluation and reporting system (revised). Mountain View, CA: RMC Research Corporation.

Teitlebaum, H., Hiller, R., Gray, T. C., & Bergin, V. (1982). Changing schools: The language minority student in the eighties. Washington, D.C.: Center for Applied Linguistics.

Texas Education Agency. (1977). Report from the committee for the evaluation of language assessment instruments. Austin: Author.

Thistlewaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. Journal of Educational Psychology, 51, 309-317.

Thonis, E. W. (1976). Literacy for America's Spanish-speaking children. Newark, DE: International Reading Association.

Thonis, E. W. (1980). Speech, print, and thought in bilingual bicultural education. Sacramento: California State Department of Education, Office of Bilingual Bicultural Education.

Thonis, E. W. (1981). Reading instruction for language minority students. In Schooling and language minority students: A theoretical framework. Los Angeles: California State University; Evaluation, Dissemination and Assessment Center.

Thorndike, R. L. (1942). Regression fallacies in the matched groups experiment. Psychometrika, 7, 85-102.

Tikunoff, W. J. (1982). Consequences for students in successful bilingual instructional settings. San Francisco: Far West Laboratory for Educational Research and Development.

Tikunoff, W. J. (1983). Utility of the SBIF features for the instruction of LEP students. San Francisco: Far West Laboratory for Educational Research and Development.

Tikunoff, W. J. (1984). Five significant bilingual instructional features: A summary of findings from Part I of the SBIF descriptive study. The Second Annual Symposium on Evaluation and Interdisciplinary Research in Bililngual Education Monograph Series, 5(3), 1-17.

Tikunoff, W. J. et al. (1981). Success indicators and consequences for limited English language proficient students in the SBIF study. San Francisco: Far West Laboratory for Educational Research and Development.

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

Topacio, C. (1979). Special community needs in bilingual education. The Bilingual Journal, 3(4), 17-19.

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

Trochim, W. M. K. (1980). The regression-discontinuity design in Title I evaluation: Implementation, analysis, and variation. Unpublished doctoral dissertation, Northwestern University.

Trochim, W. M. K. (1980). Research designs for program evalution: The regression-discontinuity approach. Beverly Hills, CA: Sage.

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

Troike, R. C. (1981). Zeno's paradox and language assessment. In S. S. Seidner (Ed.), Issues of language assessment: Foundations and research. Springfield, IL: Illinois State Board of Education.

Tucker, G. R., & Cziko, G. A. (1978). The role of evalution in bilingual education. In J. E. Alatis (Ed.), Georgetown University round table on languages and linguistics. Washington, D.C.: 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.

Ulibarri, D. M., Spencer, M. L., & Rivas, G. A. (1981). Language proficiency and academic achievement: A study of language proficiency tests and their relationship to school ratings as predictors of academic achievement. Journal of the National Association for Bilingual Education, 5(3), 47-80.

U.S. Commission on Civil Rights. (1972). The excluded student: Educational practices affecting Mexican Americans in the Southwest. Mexican American Education study, Report III. Washington, D.C.: U.S. Government Printing Office.

U.S. Commission on Civil Rights. (1975). A chance to learn: Bilingual-bicultural education. Washington, D.C.: 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, D.C.: Author.

U.S. Department of Health, Education and Welfare. (1971). Manual for project applicants and grantees: Programs under Bilingual Education Act (Title VII, ESEA). Washington, D.C.: Author.

Valdes, G., Barrera, R., & Cardenas, M. (1984). Constructing matching texts in two languages: The application of propositional analysis. Journal of the National Association for Bilingual Education, 9(1), 3-19.

Van de Vijver, F. J. R., & Poortinga, Y. H. (1985). A comment on McCauley and Colberg's conception of cross-cultural transportability of tests. Journal of Educational Measurement, 22(2), 157-161.

Veltman, C. J. (1980, October). Relative educational attainment of Hispanic-American children, 1976. Paper presented at the Aspira Hispanic Forum for Responsible Educational Policy, Washington, D.C.

Voldomec, V. (1963). Multilingualism. Leyden: A. W. Sythoff.

Vocolo, J. (1967). The effect of foreign language study in the elementary school upon achievement in the same foreign language in high school. The Modern Language Journal, 51(8), 463-469.

Wahab, Z. (1974, March). Teacher-pupil transaction in bi-racial classrooms: Implications for instruction: Paper presented at the Annual Convention of the Pacific Sociological Association, San Jose, CA.

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

Wargo, M. J., & Green, D. R. (Eds.) (1978). Achievement testing of disadvantaged and minority students for educational program evaluation. Monterey, CA: CTB/McGraw-Hill.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). Unobtrusive measures: Nonreactive research in the social sciences. Chicago: Rand McNally.

Weber, M. The effect of learning environment on learner involvement and achievement. Journal of Teacher Education, 1978, 29, (6), 81-85.

Weick, K. E. (1976). Educational organizations as loosely coupled systems. Administrative Science Quarterly, 21(1), 1-19.

Weiss, C. H. (1972). Evaluation research: Methods of assessing program effectiveness. Englewood Cliffs, NJ: Prentice-Hall.

Wells, G. (1979). Describing children's linguistic development at home and at school. British Educational Research Journal, 5, 75-89.

Werner, O., & Campbell, D. I. (1970). Translating, working through interpreters, and the problem of decentering. In R. Naroll & R. Cohen (Eds.), A handbook of methods in cultural anthropology. New York: American Museum of Natural History, 398-420.

Werts, C. E., Linn, R. L., & Joreskog, K. G. (1977). A simplex model for analyzing academic growth. Educational and Psychological Measurement, 37, 745-756.

Williams, F., and Associates. (1976). Explorations of the linguistic attitudes of teachers. Rowley, MA: Newbury House.

Willig, A. C. (1984). A meta-analysis of selected studies on the effectiveness of bilingual education. Doctoral dissertation, University of Illinois, Urbana, IL.

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

Wilson, S. M., & Hiscox, M. D. (1984). Using standardized tests for assessing local learning objectives. Educational Measurement: Issues and Practices, 3(3), 19-22.

Wolcott, H. F. (1977). Teachers versus technocrats: An educational innovation in anthropological perspective. Eugene: University of Oregon, Center for Educational Policy and Management.

Wong-Fillmore, L. (1983). "Effective Language Use in Bilingual Classrooms." In Compatibility of SBIF Features with other research on instruction for limited English-proficient students, edited by W. J. Tikunoff, 43-61. San Francisco, Far West Laboratory for Educational Research and Development.

Wong-Fillmore, L. (1983). The language learner as an individual: Implications of research on individual differences for the ESL teacher. In M. A. Clarke & J. Handscombe (Eds.), On TESOL'82: Pacific perspectives on language learning and teaching. Washington, D.C.: TESOL.

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

Worthen, B. R., & Gagne, R. M. (1969). The development of a classification system for functions and skills required of research and research-related personnel in education. Boulder, CO: American Educational Research Association, Task Force on Research Training, Technical Paper No. 1.

Wortman, P. M. (1983). Evaluation research, a methodological perspective. Annual Review of Psychology, 34, 223-260.

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

Yee, L. Y., & LaForge, R. (1974). Relationship between mental abilities, social class, and exposure to English in Chinese fourth graders. Journal of Educational Psychology, 66, 826-834.

Yen, W. M. (1986). The choice of scale for educational measurment. An IRT perspective. Journal of Educational Measurement, 23(4), 299-325.

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

Zimmerman, D. W. & Williams, R. H. (1982). Gain scores in research can be highly reliable. Journal of Educational Measurement, 19(2), 149-154.