DOCUMENT RESUME

ED 299 328 TM 012 381

AUTHOR Carballo, Eduardo; And Others
TITLE Analyzing Test Data for Program Evaluation Purposes: Are There Procedures and Other Factors Which Alter the Results?
PUB DATE Apr 88
NOTE 36p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; Achievement Gains; *Achievement Tests; *Bilingual Education Programs; Bilingual Students; Educational Policy; Elementary Education; English; Limited English Speaking; *Program Evaluation; Second Language Instruction; Spanish; *Test Results
IDENTIFIERS Bilingual Education Act 1984; *Massachusetts

ABSTRACT

Ways in which an evaluation of bilingual programs could be appropriately conducted were studied, using two transitional bilingual education programs. In Program A, the first language (Spanish) is the initial medium for all instruction with a gradual phasing in of English. Program A is located in a large suburban school system in central Massachusetts, and provides full-time education to limited English proficient students. Proficiencies of 22 students were examined through interviews and written testing. The students were given the Metropolitan Survey Battery reading and mathematics tests and the Massachusetts Basic Skills Test for grade 6 as appropriate. Sixteen students were selected from Program B, with full-time instruction in both languages to make students comfortable and competent in either Spanish or English. Program B is located in a large urban school system in Massachusetts. Similar tests were given, and scores from each program were analyzed using: (1) Model A, a norm-referenced design comparing the rate of growth for these students and others not requiring services; (2) mean standard scores during and after mainstreaming; (3) gap reduction, the "catching up" of program students; and (4) results after mainstreaming. By methods 1, 2, and 3, Program A was judged effective; for method 4, differences were not statistically significant. Because of the small number of students, none of the methods gave conclusive results for Program B. The study illustrates that each model answers different questions; each clarifies some aspect. Difficulties in evaluating Program B show the need for alternative methods. Four tables and 17 graphs show test results. (SLD)

ANALYZING TEST DATA FOR PROGRAM EVALUATION PURPOSES: ARE THERE PROCEDURES AND OTHER FACTORS WHICH ALTER THE RESULTS?

Presented at the AERA Annual Conference, April 1988

Mr. Eduardo Carballo, Title VII Project Director
Massachusetts Education Department

Dr. Susan Reichman, Deputy Director
Evaluation Assistance Center - East

Dr. Gloria Zyskowski, Research Associate
Evaluation Assistance Center - East

ANALYZING TEST DATA FOR PROGRAM EVALUATION PURPOSES: ARE THERE PROCEDURES AND OTHER FACTORS WHICH ALTER THE RESULTS?

Presented at the AERA Annual Conference, April 1988

Mr. Eduardo Carballo, Title VII Project Director
Massachusetts Education Department

Dr. Susan Reichman, Deputy Director
Evaluation Assistance Center - East

Dr. Gloria Zyskowski, Research Associate
Evaluation Assistance Center - East

BACKGROUND ISSUES RELATING TO PROGRAM EVALUATION

The issue of conducting "good" evaluations has always been of concern in relation to any educational program, but recently the topic of suitable approaches to evaluating Title VII Part A as well as other bilingual education programs has received renewed interest. The importance of adhering to strict evaluation guidelines and procedures has not always been a priority with bilingual education, but the current emphasis on accountability in general has led to the development of systematic approaches to assessing the effectiveness of Title VII and state bilingual programs. With increasing pressures for school improvement and expanding information technology resources, this renewed focus on program evaluation seems highly appropriate (Burstein, 1984).

While the legislation empowering bilingual programs dates back to 1968 (The Bilingual Education Act - Title VII of the Elementary and Secondary Education Act) and, in Massachusetts, the State Chapter 71A of the Acts of 1971, no specific criteria for establishing program effectiveness were provided initially. Then, as a direct result of the landmark 1974 court case of Lau v. Nichols, impetus was provided for state and local educational agencies to design and implement bilingual education programs in a more systematic and accountable manner. Soon after this Supreme Court decision, the Bilingual Education Act was amended to specify in great detail the policies and procedures state and local educational agencies were expected to follow. Specifically, the amendments directed the Commissioner of Education to develop and disseminate bilingual education program models which would contain separate evaluation components (Holt & Arrellano, 1980).

From 1977 through the early 1980s, the Bilingual Education Act underwent a series of amendments, each time refining the extent and scope of evaluation activities. As a result of further amendments in 1984 and 1986, the regulations took the form they have today. The evaluation requirements (P.L. 98-511, section 733) presently specify that the evaluation design must include ". . .a measure of educational progress of Title VII project participants when measured against an appropriate nonproject comparison group." (34 CFR, section 500.50) In addition: (1) the evaluations must be representative of all persons, schools, or agencies served by the funded programs; (2) the instruments and procedures used in evaluations must provide reliable and valid measures of the program's progress toward accomplishing its objectives, taking into account the characteristics of the

population served; and (3) data collection procedures must 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 Title VII regulations make provision for using a variety of indicators of project effectiveness. In addition to using scores obtained from the administration of tests of academic achievement, program evaluators should also consider changes in the rate of student grade-retention, drop-out, absenteeism, referral to or placement in special education classes, placement in programs for the gifted and talented, and enrollment in post-secondary education institutions. Also, information must be collected on both current and former participants in Title VII Part A programs.

As can be seen by examining the history of evaluation requirements for federally and state funded bilingual education programs, concerns about program effectiveness have steadily increased over the years. A number of attempts have been made to develop systematic guidelines for evaluating bilingual programs, and numerous studies at the federal, state, and local levels have been undertaken to demonstrate the effectiveness of bilingual education. Desrite these efforts, it can safely be said that very little is known about the benefits, if any, that have accrued to program participants (Tallmadge, et a , 1987).

While educators and policy makers all agree that bilingual education services are needed to help language-minority students acquire an adequate education, there is little agreement regarding what instructional approach is most effective with these students. It has been noted 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." (Okada, et al., 1983, p. 4)

However, this lack of evidence regarding the effectiveness of bilingual programs should not lead one to conclude that the programs are not, in fact, effective and beneficial for the students participating. In one review of evaluation reports submitted to government funding agencies, 97% of the studies were rejected because they contained serious methodological flaws (Zappert and Cruz, 1977). In another examination of 176 evaluations of bilingual programs, only 39 of them were found to be methodologically sound, empirical studies (Baker and de Kanter, 1983). Therefore, one could more logically conclude that it is not the bilingual programs themselves that are at fault, but rather the poor quality of bilingual program evaluations.

It has long been recognized by practitioners that the assessment of a limited English-proficient (LEP) student's competence in English is not a simple and straightforward task. Aside from the general lack of psychometrically sound instruments necessary to measure program impact (Willig, 1985), there are a number of variables known to be related to learning a second language that need to be incorporated into an evaluation designed to assess the effectiveness of bilingual programs. Such factors as age, cognitive skills, parental attitudes, language use patterns in the school, home and community, socio-economic background, ethnicity, motivation and

self-esteem, immigrant status, degree of bilingualism, oral or written language skills, and school/teacher differences all have been identified as affecting learning in language minority children (Baker and Pelavin, 1984).

More importantly, perhaps, is the recognition that bilingual education programs may require an extensive period of time for their effects to become apparent. Several studies have concluded that the cumulative effects of bilingual programs on increasing achievement and IQ scores do not emerge until the fourth, fifth, or sixth years of bilingual instruction (Ovando and Collier, 1985). Further, it has been suggested that evaluating program effectiveness needs to take into account how successfully former LEP students function in mainstream classrooms or in society (Paulston, 1977).

Regardless of the difficulty encountered when attempting to assess the impact of bilingual education programs, a great deal of effort has been made to improve the quality of evaluation in bilingual education (O'Malley, 1984). A variety of evaluation designs and data analysis methods have been suggested over the years, each with particular advantages and limitations. While classically-trained evaluators might prefer to implement true experimental designs with students and teachers randomly assigned to either treatment or control groups, this condition is impractical in the educational setting (not to mention illegal under Title VII and Massachusetts Chapter 71A, which requires that all eligible students must be served). Consequently, more feasible approaches had to be found for determining how much of a student's observed growth can be attributed to the treatment provided by the bilingual education program.

One approach involves the non-equivalent control group design. The requirements for implementation of this design include: the use of program and comparison groups that are similar in nature; pre- and posttesting of both groups at approximately the same period of time, using the same test or tests with standardized norms; and, the assumption that there will be no statistical differences in the pretest scores for the two groups. One major limitation on the utility of this design with Title VII projects concerns the fact that, even if the pretest scores for the treatment and control groups are found to be equal, other crucial but probably unmeasured differences are likely to provide a large degree of error variance (Reichardt, 1979). Home language, family mobility, prior exposure to English, and prior schooling are examples of variables that should be addressed.

A second suggested approach to evaluating bilingual programs is the regression-discontinuity design. The requirements for implementation in this case include: one pool of students who are similar in nature, have a range of capabilities, and are divided between treatment and control groups based on systematic differences; use of a designated cutoff score for assigning students to groups; and, a single population from which all students are drawn. There are several problems with this design that might make it inappropriate for use with most Title VII projects and state transitional bilingual education (TBE) programs. Large sample sizes are required, the design is computationally complex, and computers are required to carry out the analyses. More importantly, the model assumes that the students above and below the designated cutoff score will be representative of a single population. It is more likely with bilingual education programs that the students below the cutoff will be LEP, while the majority of

students above the cutoff will be native English speakers, two distinctly different populations (Tallmadge, et a.., 1987).

A third suggested design for bilingual evaluation is the quasi time series approach. The implementation requirements for this design include: the assurance that enough students will enter the program at each age/grade level to develop stable baseline entry data; pretesting at program entry followed by posttesting on a periodic basis; and, program stability over time. One advantage of a quasi time series design is that the students in the bilingual education program serve as their own control group, thus eliminating most of the problems associated with finding a suitable comparison group (McConnell, 1982). On the other hand, it might be quite difficult to obtain pre-treatment test scores on enough students at all levels served to make this design feasible. Also, this is particularly difficult in Massachusetts programs because most do not use standardized testing in English at the beginning of the program.

Another suggested approach for evaluating bilingual education programs is the Title I/Chapter 1 Model A design, frequently referred to as the norm-referenced design. This approach has been in widespread use with compensatory education programs for a number of years, and basically requires that students in a program be pre- and posttested on a periodic basis, using a test with national norms and standard scores which can be converted to normal curve equivalent (NCE) scores. While this design has the advantages of familiarity and relative ease of implementation, it has serious limitations for use in the context of bilingual education. There is a low probability that the treatment and control groups will have equivalent exposure to subject matter prior to pretesting (DeMauro, 1983) and, more importantly, deriving growth expectations for LEP students from the performance of a non-LEP population (the group from which national norms are obtained) has been cited as a fundamentally unsound practice (Baker and Pelavin, 1984).

The most recently developed approach to evaluating bilingual education programs is the gap-reduction design. The requirements for implementation of this design include: either pre- and posttesting students on a periodic basis using a test with national norms and standard scores; or, pre- and posttesting both a program group and an existing comparison group on a periodic basis using the same test. In this case, the comparison group does not have to be similar to the program group, but could consist of mainstreamed students, monolingual English speaking peers or other groups, depending on the specific evaluation question(s) of interest. While this design measures growth from pre-to posttest, it is not possible to break down that growth into treatment-related and non-treatment-related aspects. However, the gap-reduction design can be implemented simultaneously with a treatment and a no-treatment group to provide more information about the actual treatment effect, and, unlike the designs previously discussed, this evaluation model has no significant implementation difficulties (Tallmadge, et al., 1987).

It is obvious that a number of approaches have been suggested for evaluating bilingual education programs. Each design has a variety of limitations and advantages, many of which depend on the specific evaluation question(s) of interest and the various program constraints that may exist for a particular situation. Each of these various designs also has one charac-

teristic in common -- the use of pre and posttest scores to determine average student growth -- a metric that is then used for program evaluation purposes.

However, given the special characteristics of bilingual populations, test data alone will not render an accurate assessment of program effectiveness. Besides the difficulties inherent with using nationally normed tests with LEP populations, a number of additional variables affect the manner in which these students respond to bilingual education programs, any number of which could contaminate the evaluation data. Therefore, a comprehensive evaluation plan for bilingual education should incorporate a number of success indicators (Walsh, Carballo, 1986).

Why, then, does the investigation reported herein concentrate on an analysis of test data? Even though most individuals would readily agree that a thorough program evaluation should include much more than test scores, the reality of the situation is that the various test scores required by the Title VII Part A evaluation regulations may be the most straightforward data to collect. Unfortunately, while test scores can provide very useful data and are readily available, it is also true that these scores are easily misinterpreted and misused.

Those involved with the evaluation of Title VII Part A and state programs should therefore consider how different analysis procedures may affect the evaluation results. Furthermore, in order to properly interpret and use evaluation results, it is necessary to be aware of how differences in student characteristics may affect the overall evaluation outcomes. Finally, regarding the interpretation and use of evaluation results, it is necessary to be aware of the limitations of standardized test data when making mainstreaming decisions and overall program evaluation decisions.

MASSACHUSETTS' APPROACH TO PROGRAM EVALUATION

Under the state bilingual education law (Chapter 71A), the Commonwealth of Massachusetts requires the following:

> Chapter 71A Section 22 "Every school-age child of limited English-speaking ability not enrolled in existing private school systems shall be enrolled and participate in the program for a period of three years or until such time as he achieves a level of English language skills which will enable him to perform successfully in classes in which instruction is given only in English, whichever shall first occur. A child of limited English-speaking ability enrolled in a program in Transitional Bilingual Education may, at the discretion of the school committee and subject to the approval of the child's parent or legal guardian, continue in the program for a period longer than three years."

School districts must provide the program when 20 or more children of the same language group are found to be limited English-speaking. The classification of the children can only be done by a teacher of Transitional Bilingual Education qualified under the Act.

The districts are required to identify, place, and transfer students in

and out of the program by utilizing multiple criteria consisting of: interviews with students and parents, teacher recommendations, school records, language proficiency testing, and standardized testing. Although the process of carrying out these requirements varies from district to district, the basic criteria remain the same.

In addition, Massachusetts law contains two areas of evaluation requirements. The first is the annual evaluation of children enrolled in transitional bilingual education (TBE) programs to assess their attainment of English language skills. The second requirement is a self-evaluation conducted by the LEA via a locally established committee made up of staff and parents, and the results of this study are reported to the state on an annual basis. Since 1986, however, the Massachusetts State Board of Education has adopted eleven major goals and objectives for the improvement of the State's Bilingual Education Program. One of these major goals is the proper evaluation of TBE programs in the Commonwealth. In addition, during the same period of time, several bilingual education bills have been filed in the Massachusetts legislature to amend the existing law, and at least three of these bills recommend the establishment of program evaluation requirements.

PILOT STUDY

Rather than address the overall issue of the implementation of quality evaluation designs, this particular pilot study focused on factors which might affect overall program evaluation results, including various data analysis procedures conducted on student test scores while in the program and follow-up information on students who have been mainstreamed from that same program. The specific questions of interest were:

- When using the same set of test scores from students in a specific program, do the analysis procedures for various evaluation designs result in different program evaluation conclusions?
- Do the data analysis procedures affect whether student gains while in the program are reflected on various standardized test scores?
- If program effectiveness is evaluated using information collected after students have been mainstreamed from the program, how do the results of this type of evaluation compare to the more typical evaluation activities conducted while students are actually in the program?

It should be kept in mind that this particular pilot study focused on approaches to evaluating the effectiveness of an individual program and was not concerned with individual student growth or the individual exit decisions.

PROGRAM DESCRIPTION OF PILOT SITES

The data were collected at the state level, from two transitional bilingual education programs in schools in Massachusetts and included student

data over three academic years. An additional three years of follow-up descriptive data and test results were also collected on each student. The two programs which participated in the pilot study were as follows.

PROGRAM A

Program A is located in a large, suburban school system in Central Massachusetts and has been in operation since 1972. Approximately three percent of the students are classified as limited English proficient and are participating in TBE programs. These students represent one large group (Spanish). Program A provides services for students entering Kindergarten, first and the early part of second grade. Older students are provided with services through the Spanish Bilingual Program at the elementary, middle and secondary levels.

Program Goals. Program A is designed to provide a full-time meaningful education for children of Limited English Proficiency regardless of language, age or level of achievement. It is dedicated to utilizing, preserving, and directing the student's language, academic language and pride in his/her native history and culture while teaching them English and an understanding of the culture and history of the United States. The overall goals are as follows:

1. The student will be able to speak, comprehend, read and write English, thereby providing them with the tools to secure gainful employment or to continue their education in a predominantly English-speaking society.

2. TBE students will develop and/or maintain their competency in speaking, reading, and writing in their native language.

3. TBE students will use their native language and English to increase their knowledge and skills in all academic areas.

4. TBE students will learn about and have respect for their own cultural heritage and the cultural heritages of other groups.

5. TBE students will develop positive attitudes toward school and school-related activities.

Staffing. The Program A staff are a stable teaching force of 20 bilingual and ESL teachers. Only five teachers have left the program in the last five years. Of the 12 bilingual teachers, 11 are fully certified and one is on waiver while completing an internship. Most of the eight ESL teachers are grandfathered (they were teaching ESL before certification was required in 1982). Two of them also hold bilingual certification and another three have fluency in Spanish. Fifty percent of the staff hold Master's Degrees. The staff is provided with regularly scheduled release days throughout the academic year for inservice training.

Instructional Methodology. Program A utilizes the first language of the student as the medium of instruction in all subject areas to the extent necessary to ensure academic progress and concept attainment. Kindergarten

bilingual class is not counted towards the three years required for transition to the mainstream. Since the instructional aim is to convert gradually from the use of the native language to the English language medium of instruction, students are provided opportunities, such as mathematics, to transfer skills from the native language to English.

Program A does not subscribe to the usage of precise percentages of native language instruction; the student's program schedule is dependent upon the individual student's language and total academic needs. However, while no specific percentage of time can be regarded as fixed, the general guidelines for the non-English speaker provide initially for 80-90% native language instruction and 10-20% ESL instruction, with a gradual decrease in the proportion of L1 and an increase in L2 as the student progresses along the continuum of English language acquisition.

Within the self-contained classrooms, students are grouped for instruction according to their language and ability levels in each of the academic subject areas. Content material is adapted and presented at the students' varying levels to permit each student to perform at his/her current skill level. The English as a Second Language/English Language Development (ELD) is a pull-out component. There are three levels of CORE offerings: Beginner, Intermediate and Advanced.

Entry-Exit Criteria. At the time of entry, NEP students are assessed through locally developed instruments. In addition, an individual interview is conducted with the student in his/her native language by a qualified bilingual person to determine the student's language dominance/preference, provide a cross-check with the Home Language Survey Information, and to obtain information regarding attitudes and interests. An interview is also provided in English by a qualified ESL person(s) to assess the students' proficiency in the aural, oral, reading and writing aspects of the English language. Once the student documents have been evaluated, the interview completed, and the results of the examination of student records and language dominance and proficiency levels completed, the student is placed in the appropriate chronological grade and level of the TBE Program.

During the Spring (April/May) of each academic year, all TBE students are evaluated for proficiency/achievement in English by means of a combination of classroom observation by the teacher(s) and formal assessment instruments as described below:

	UNDERSTANDING	SPEAKING	READING	WRITING
K	Teacher Classroom Assessment	Teacher Classroom Assessment		
1st - 12th	ESL Test and Teacher Classroom Assessment	ESL Test and Teacher Classroom Assessment	Metropolitan Reading Survey (1978)	Writing Based on Picture Stimulus

All students in the TBE Program who have achieved a level of English proficiency as measured by academic performance, achievement, and assessment will be eligible for transfer to the standard curriculum and all students completing three years not counting K are eligible for transfer also. However, no student may be mainstreamed from the TBE Program to the standard curriculum prior to three years of enrollment, except under the following conditions:

1. Upon teacher recommendation, Parent(s)/Guardian(s) input and written consent, and approval of the Bureau of TBE, Massachusetts Department of Education.

2. The Parent(s)/Guardian(s) signs a Waiver of Services Form which is placed in the student's cumulative folder, and a copy maintained in the TBE Program office.

Should a student who transfers out of the TBE Program prior to three years continue to demonstrate limited English proficiency, he/she may be re-enrolled in the program for a length of time equal to that which remained at the time of transfer.

All mainstreaming must take into account the following:

- o Teacher recommendation
- o Result of annual examination
- o Parent input
- o On-going academic achievement and classroom performance
- o Years in TBE Program
- o Pupil attitude

Unlike many other districts, Program A offers an extra year of ESL support to first year mainstream students (providing students remain in the same school they attended for TBE).

Information on Program A Students Used in Pilot Study

Because the intended purpose of this pilot study was to consider various ways in which a program evaluation might be conducted, student level data were systematically selected to provide at least five years of information on each student. Within each site, student cumulative folders were reviewed, with student level data selected on the basis of whether: (1) an individual had been in the program for at least three years; and (2) followup data for at least two years was available in the cumulative folders.

A total of 22 students were identified as matching the necessary selection criteria. Of these students, all were born in the United States but were classified as non-English speakers, with a native language of Spanish. Of the students selected, 21 students entered the program in kindergarten and one in first grade. Because the Program A school district does not count kindergarten towards the three-year requirement for exit, all students used in the sample had two years of program treatment by the time they were first tested. For 20 of the students this was the only school district attended, while two students briefly attended schools elsewhere. During the three years in the program, 68% received Chapter 1 services and 64% (duplicated count) received other types of services (e.g., speech,

resource room, social services). All of the students were exited at the end of their third year in the transitional bilingual program. At the time of identification for this pilot study, all students were currently placed in the mainstream at grade level.

Additional information collected on these students included the method used to determine proficiency in L1 and L2. In regard to proficiency in L1, 100% of the students were interviewed, 100% of the students were administered an oral language proficiency measure, 27% were administered a language proficiency test and previous school records were reviewed for 9% of the students. As to proficiency in L2, interviews, oral language proficiency measures and bilingual/ESL teacher judgement were used with all students. Additionally, a language proficiency test was used with 18% of the students and previous school records for one student (5%).

As to exit information, all students were exited from the program at the end of their third year of participation. Assessment team recommendations, language proficiency tests and standardized achievement test results were used for all 22 students. Additionally, 77% of the students were interviewed prior to exit.

Where available, the following standardized test scores were collected for each student while in the program: Metropolitan Survey Battery - Reading (Pre Primer, Primer, Primary 1, Primary 2). Follow-up test information included the SRA Reading and Math Tests (Levels D and E) and results of the Massachusetts 6th Grade Basic Skills Test (reading, math and writing).

Further information collected for each student included the average number of days attended for each year in the program and for two years after exit; and grades in five subject areas (reading, English, math, science and social studies) for two and one half years after mainstreaming. This information is summarized in Tables 1 and 2.

PROGRAM B

Program B is located in a large, urban school system in Massachusetts and has been in operation since 1974. Approximately six and one-half percent of the students in the district are classified as limited English proficient (LEP) and participate in Transitional Bilingual Education programs. The primary language group is Spanish (61% of the LEP students in the district) and, of this group, approximately 65% are of Puerto Rican descent.

Program Goals. This two-way program is designed to provide a full-time program in both the native language and English. The major goal is that all students will be comfortable and competent using either Spanish or English. Other major goals are:

- To instill in all the students an enjoyment of learning.
- To develop basic and critical thinking skills via reading, language arts, math, science, social studies, computers, physical education, art, and music.
- To create a profound cultural experience and to promote the kind of

awareness, sensitivity, officiation and understanding that comes from mastering a second language and second culture.

- To foster and develop a positive self-image and social interpersonal skills.

Staffing. The Program B staff are a veteran and stable teaching force of 13 bilingual and ESL teachers. No staff has left the program over the last two years, and the turnover rate is very small. Of the 12 bilingual teachers, ten are certified and two are on waiver. The one ESL teacher is certified, as well as five of the bilingual teachers who also teach ESL. Fifty percent of the staff hold Master's Degrees. The staff is provided with regularly scheduled release days throughout the academic year for inservice training. In addition, the staff are working with a nearby college developing additional literature for reading program improvements.

Instructional Methodology. Program B is a K-5 program for NEP, LEP and fully-English proficient (FEP) students. The program is a two-way model that utilizes both the first and second language of the students as the medium of instruction in all subject areas. As a two-way bilingual program, it has the responsibility of providing instruction in both English and Spanish as it helps its students achieve the curriculum objectives that have been designated by the Program B Public Schools. The goal is that all students will be comfortable and competent using either Spanish or English. In the two-way program, language is a means, as well as an object, of instruction and so both English and Spanish are used as the medium of instruction in a variety of subjects. Since it is recognized that it is not always appropriate nor instructionally sound to switch from language to language in any given class period, the following language guidelines are maintained:

- Students are taught reading and language arts in their primary language.

- Students receive either an English or a Spanish as a second language class in their secondary language.

- Both English and Spanish are used in all other subjects and school related activities. The overall language balance is approximately 50-50, and Spanish or English is selected as the primary or dominant language for any given class period. Subjects like math, social studies, and science are taught using primarily one language for one particular unit and then using the other language for the next. The actual languages used during any unit are planned carefully to ensure the maximum academic development of the students, as well as their becoming effective learners in two languages.

Entry/Exit Criteria. Identification of the LEP student is carried out using the LEA procedures specified in the Voluntary Lau Plan. First, all parents of new students in the school system complete a Home Language Survey. Upon review of the completed survey, those which indicate that the students speak a language other than English in the home are assessed in listening, speaking, reading, and writing in English and their native language using LEA-developed structured interviews, tests, and CLOZE reading passages. The combined results of the assessment procedures are used to

identify LEP students and to assign them to an appropriate category. The Lau Plan further establishes criteria for the instruction of these students within a bilingual program.

The Lau Plan details five steps that define the process by which LEP students move from bilingual instruction to a totally mainstreamed education program. The following is a description of each step:

STEP ONE shall indicate a schedule whereby a student receives all academic instruction, and all or most non-academic instruction in bilingual education classes. The student is integrated, if at all, only for physical education, study hall, and/or lunch. Note: due to the minimal amount of integration, this step shall be used (if at all) only for diagnostic purposes for new students who enter with no records from their previous school.

STEP TWO shall indicate a schedule whereby a student receives all academic instruction, and some non-academic instruction, in bilingual education classes. The student is integrated for non-academic subjects as physical education, art, music, library, industrial arts, home economics, study hall, typing, etc.

STEP THREE shall indicate a schedule whereby a student receives most academic instruction and some non-academic instruction in bilingual education classes. The student is mainstreamed for one or two academic subjects. The elementary student is also mainstreamed for one or two of the following: mathematics, science, social studies, English, reading.

STEP FOUR shall indicate a schedule whereby a student is mainstreamed for most or all academic and non-academic subjects with the option to take elective bilingual subjects as space is available.

STEP FIVE shall indicate a student who has achieved successfully in STEP FOUR and is totally mainstreamed for all academic and non-academic subjects. This shall be a monitoring step whereby the student's progress is reviewed periodically and bilingual support services such as counseling are provided as needed. If an individual secondary level student wishes, he/she may select one or two electives from among the courses offered in the bilingual program.

Information on Program B Students Used in Pilot Study

A total of 16 students were identified as matching the necessary selection criteria. Nine of the students were in 6th grade at the time of identification, four in 7th grade, and three were in 8th grade. Of these students, all have Spanish as their native language, although almost all were born in the U.S. but were classified as non or limited English-speaking. Of the 16 students from Program B, 10 students entered the program in kindergarten, five in the first grade, and one in the 2nd grade. For 15 of the 16 students, this was the only school district attended, while one student had attended school elsewhere in the U.S. as well.

Ordinarily, at the point a student becomes FEP, in most Massachusetts' districts they are mainstream. However, because the Program B school district allows for fully-English proficient students to continue in two-way programs, the point after which students became FEP was the point utilized in the mainstream for the purpose of this pilot study. While in the program, 11% of the 6th graders, 25% of the 7th graders, and 66% of the 8th graders received Chapter 1 services. At the time of this pilot study, all students were currently placed at grade level.

In addition, information collected on these students included the method used to determine proficiency in L1 and L2. In regard to proficiency in L1, 100% of the students, regardless of grade level, were interviewed, and 100% were administered an oral language proficiency measure, 27% of the 7th graders were administered a language proficiency test, and previous school records were received for 33% of the 6th graders and 66% of the 8th graders.

Exit Information. All students, regardless of grade level, were evaluated by an assessment team for the purpose of establishing the student's level of fluency in English. All received language proficiency tests, standardized achievement tests, and all but 22% of the 6th graders' records of academic partial mainstream data were reviewed.

Where available, the following standardized test scores were collected for each student while in the program: Metropolitan Survey Battery, Reading and Math (Primary 1, Primary 2, Elementary, Intermediate). Follow-up test information included the MAT-6 Multi-Level Battery, Forms L & M. Spring norms were used by the districts for all tests. Also, results of the Massachusetts 6th Grade Basic Skills Test (reading, math and writing) were reviewed where available.

Additional information collected for each student included the average number of days attended for each year in the program and for two years after exit; and grades in five subject areas (reading, English, math, science, and social studies) for two and one-half years after exit. This information is summarized in Tables 3 and 4.

DATA ANALYSIS APPROACH

In order to obtain a picture of how various types of evaluation designs and test analysis procedures can affect program evaluation results, the test scores for each program were analyzed in the following ways:

(1) Model A - The Chapter 1 norm-referenced design (called "Model A") was used to measure growth as reflected by standardized test scores administered during program participation. With this particular design, the major evaluation issue of interest is how rate of growth for students in the program compares to rate of growth for those students who do not receive services, as reflected by the standardized test norms. The unit of analysis for the test scores is a normal curve equivalent (NCE).

(2) Mean Standard Scores - The second evaluation design used was a basic analysis of mean test scores to determine growth in standard score

units from standardized tests administered both during program participation and after mainstreaming. Using this particular approach, the major evaluation issue of interest is simply whether growth, as reflected by test scores, occurs over the testing period. The unit of analysis is the standard score (SS) of the test(s) administered.

(3) Gap Reduction - The gap reduction design was used to measure growth during program participation, again as reflected by standardized test scores. With this particular design, the major evaluation question of interest is whether program participation helps students catch up to mainstream peers, as reflected by test scores. Additionally, follow-up test scores were used to determine whether any gap that existed at time of exit stayed relatively the same, increased or decreased.

Gap reductions are calculated as follows: "The pretest gap is the comparison group's mean pretest score minus the project group's mean/median pretest score divided by the comparison group's pretest standard deviation. The posttest gap is the comparison group's mean posttest score minus the project group's mean/median postest score divided by the comparison group's posttest standard deviation. The gap reduction is the pretest gap minus the posttest gap" (Tallmadge et al., 1987, p. H-4). The unit of analysis for the test scores is the relative growth index (RGI), a newly proposed metric. An RGI is the "... project group's growth minus the comparison group's growth divided by the comparison group's growth and multiplied by 100 (Tallmadge et al., 1987, p. H-4).

(4) Out-of-Program Results - The final evaluation design used was one which focused on student performance after mainstreaming. Using this particular approach, a variety of data were collected on students after they had left the program. For this pilot study, standardized test scores, results on the Massachusetts Basic Skills Test and grades in the mainstream were used as indicators for establishing program effectiveness. Using this particular approach, the major evaluation issue of interest is whether the program provides students with the skills necessary to succeed in the mainstream setting.

EVALUATION RESULTS

Each of the evaluation results used in this pilot study yielded its own units of analysis, resulting in somewhat diverse views of program effectiveness. As the results for the two programs show, each model addresses somewhat different issues and so, evaluates the programs in various ways.

PROGRAM A RESULTS

Model A: If it is assumed that students in need of bilingual services would score below their mainstream peers on a standardized achievement test and continue to stay behind (or even fall further behind) unless they received additional services, then it could be argued that any increase in the rate of learning over time is an indicator of program effectiveness. Three data points using the Metropolitan Achievement Test, Total Reading scores from students in the program were used to evaluate Program A from this point of view.

The average test scores, expressed in normal curve equivalents (NCEs) were as follows: Program A students tested in the spring of 1983 were all in grade 1 and had an average NCE score of 35. One year later, in the spring of 1984, all students were in grade 2 and had an average NCE score of 43. At the third data point, all students were in grade 3 and had an average NCE score of 46.

Looking at gains over two years of services, students in Program A showed an average gain of eight NCEs from the spring of 1983 to the spring of 1984 and an average NCE gain of three from the spring of 1984 to the spring of 1985 (see Figure 1). As the data reflect, students in Program A showed an increased rate of learning over what would have been expected had they not received special services. Had students, on the average, maintained their relative position, the expected NCE "gain" would have been zero, reflecting growth at the same rate as students in the mainstream. Students in Program A showed an average growth rate for the two-year period which was greater than the norming group. From a program evaluation point of view, it would appear that Program A is effective.

Mean Standard Scores: Using this approach for evaluating services, the only assumption made is that the program should contribute to students' showing growth over time, as reflected by an increase in test scores. To apply this model, the same three years of scores from the Metropolitan Achievement Test, Total Reading were used for students in the program.

The average pretest/posttest scores, expressed in standard scores were as follows: Program A students tested in the spring of 1983 (grade 1) had an average standard score of 453. In the spring of 1984 these students, now in grade 2, had an average standard score of 595. By the spring of 1985 (grade 3) the average standard score was 645. Looking at gains over the two years of services, students in Program A showed an average gain of 140 standard score units from the spring of 1983 to the spring of 1984 and an average gain of 50 standard score units from the spring of 1984 to the spring of 1985 (see Figure 2). Both gains were significant as indicated by two-tailed t-tests on each set of test scores ($t = 11.21$, $p < .001$; $t = 6.72$, $p < .001$, respectively). Using this approach, the program would also be evaluated in a positive light.

Gap Reduction: Using the gap reduction for program evaluation purposes rests on the assumption that an effective program is one which helps students in a program close any gap which may exist between their performance and their mainstream peers. This model was applied using the same three data points from the Metropolitan Achievement Test, Total Reading for students in the program. In order to evaluate the program using the gap reduction model, analysis procedures were conducted to determine both the amount of gap reduction and the relative growth index.

Focusing first on the test results from the spring of 1983 (end of grade 1) to the spring of 1984 (end of grade 2), a gap reduction of .25 was obtained (see Figure 3). Over this time period, students in the program learned at a rate that was about one-quarter faster than their mainstream peers, as reflected by the test norming group. The relative growth index for this period was 18%. Students did, in fact, narrow the gap that existed between them and the norming group.

Test scores for the period from the spring of 1984 (end of grade 2) to the spring of 1985 (end of grade 3), resulted in a gap reduction of .14 (see Figure 3). Again students showed a learning rate that was faster than the norming group. The relative growth index for this period was 25%. As with the 1983 to 1984 period, test results indicate that students in the program narrow the gap between themselves and the norming group.

Evaluating Program A using the gap reduction approach again results in positive conclusions regarding the effectiveness of the program. Students who participated in the program for the three years did decrease the gap that existed between their average test score and that of the norming group. The program was effective in regard to helping these students learn at a faster rate than the norming group, as reflected by the reading test scores.

Out-of-Program Results: Since one of the main objectives of transitional bilingual education programs is to prepare students for mainstream classroom work, it could be argued that TBE program effectiveness be measured by examining student performance after mainstreaming has taken place. A variety of indicators were examined to obtain an illustration of students' performance after they had exited the program. While it is recognized that these indicators will also be affected by mainstream classroom events, it is nevertheless believed that student performance in the mainstream is directly related to TBE program effectiveness.

Part of the statewide school assessment effort in Massachusetts involves the annual administration of a basic skills test to all students in grades 3, 6, and 9. The test battery covers reading, writing, and mathematics skills. A score of 65% correct or higher is considered "mastery" on the reading and mathematics components, while the writing section is scored and reported as either "pass" or "fail." Students in Program A took the Massachusetts Basic Skills test at the beginning of grade 6 (in the mainstream), with the following results:

- o 54% achieved mastery in reading with a score of 68% or higher;
- o 23% achieved mastery in reading with a score of 80% or higher;
- o 72% achieved mastery in math with a score of 68% or higher;
- o 46% achieved mastery in math with a score of 80% or higher; and,
- o 95.5% passed the writing section of the test.

To put these scores in some perspective, of the district-wide population of 6th grade students, 530 children were eligible to take the basic skills test. Eighty-six percent of these students attained mastery on the reading section of the test, 89% achieved mastery on the math portion, and 93% of the students passed the writing section.

Scores on the basic skills reading test for students in this pilot study were found to be significantly correlated with standard scores on a standardized reading test administered prior to exit from the TBE program ($r = 0.66$, $p < .01$) and with standard scores on a standardized reading test administered in the mainstream ($r = 0.78$, $p < .001$). Based on the results

of the statewide test, it would appear that skills acquired in the TBE program have a significant effect on mastery of basic skills in the mainstream and on test scores obtained after exit from the program.

Another indicator of program success is grades in the mainstream. The students in Program A who were included in this pilot study received average grades throughout their three years to date in the mainstream. The mean grade obtained was C+ in all subject areas investigated (Reading, English, Math, Science, and Social Studies). Grades in Reading and English from school year 1987-1988 were found to be significantly related to scores on the reading portion of the state basic skills test administered during the fall of that year ($r = 0.69$ and 0.68 respectively, $p < .01$), and this could be considered an indirect measure of program effectiveness.

One further indicator of program success is test scores in the mainstream. By examining Figure 4, it can be seen that, once they had been exited, students in this pilot study did not maintain the pattern of steadily increasing mean scores on standardized tests that they exhibited while in the program. Nevertheless, the mean standard scores obtained on standardized reading tests administered in school years 1985-86 and 1986-87 indicated that some growth had occurred. However, the results were not found to be either statistically or educationally significant, so test scores in the mainstream would not appear to be a useful indicator of program effectiveness.

EVALUATION RESULTS - PROGRAM B

Model A: As was done with Program A, it is assumed that the use of Model A for program evaluation purposes will provide an indication of program effectiveness. Two years of testing with the Metropolitan Reading Test, Total Reading were used to evaluate Program B from the Model A point of view. As the data reflect, some students in Program B showed an increased rate of learning over what would have been expected had they not received special services. In most cases, the number of students was too small to make any reliable conclusions regarding the program.

The average test scores, expressed in normal curve equivalents (NCEs) were as follows: Of the students tested in the spring of 1983, the average NCE for those in the first grade was 48 (n=3) and the average NCE for those in the third grade was 32 (n=3). For students tested in the spring of 1984, the average NCE for those in the second grade was 50 (n=7) and the average NCE for those in the fourth grade was 40 (n=3). For students tested in the spring of 1985, the average NCE for those in the third grade was 43 (n=7) and the average NCE for those in the fifth grade was 60 (n=3).

Looking at gains over two years of services (see Figure 5), students in Program B showed an average gain of: 1.5 NCEs for first graders tested in the spring of 1983 and then as second graders in the spring of 1984; 7.8 NCEs for third graders tested in the spring of 1983 and then as fourth graders in the spring of 1984; a loss of 7 NCEs for second graders tested in the spring of 1984 and then as third graders in the spring of 1985; and a gain of 18 NCEs for fourth graders tested in the spring of 1984 and then as fifth graders in the spring of 1985. While gains show a general positive trend, given the small number of students for which matched test

scores were available, no real conclusions regarding program effectiveness can be made.

Mean Standard Scores: Using the approach of simply looking for growth over time for students who participated in Program B, as reflected by standard scores on the Metropolitan Achievement Test, Total Reading, a slightly different evaluation picture would emerge. As can be seen by the data presented in Figure 6, all grades tested during the period of 1983 to 1985 showed some gains, as indicated by the increase in standard scores. However, these increases showed a broad range, from an inconsequential gain of 2 to moderate increases of about 121 standard scores and, additionally, were based on very small numbers of students. Again, a general positive trend can be seen, but not verified given these limitations.

Gap Reduction: The final evaluation approach used with the Program B data was that of the gap reduction approach. In this case gap reductions ranged from a -45% to a +42%. The emerging pattern in regard to program evaluation shows a conflicting picture of both increasing and decreasing gaps over time. As with the previous models, the number of students with matching data was very low, making it impossible to actually draw conclusions as to the overall effectiveness of the program.

Out-of-Program Results: As was previously noted, students in Program B who were included in this pilot study were not all in the same grade, and the data were analyzed separately by grade level. Because of the testing cycle for the Massachusetts Basic Skills Test, only those students who were in Grade 6 during the 1987-88 school year were administered this test, so it was not possible to examine this indicator of program effectiveness for all pilot study participants in Program B.

Of the students in Grade 6 who had scores reported for the basic skills test, the following results were obtained:

- 83% achieved mastery in reading with a score of 65% or higher;
- 50% achieved mastery in reading with a score of 80% or higher;
- 100% achieved mastery in math with a score of 65% or higher;
- 83% achieved mastery in math with a score of 80% or higher; and,
- 83% passed the writing section of the test.

Scores on the basic skills reading test for students in this pilot study in Program B were not found to be significantly correlated with standard scores on a standardized reading test administered prior to exit from the TBE program, or with standard scores on a standardized reading test administered in the mainstream. While this small group of students performed well on the statewide test, the data do not indicate that the skills acquired in the TBE program necessarily have a significant effect on mastery of basic skills in the mainstream or on test scores obtained after exit from the program. Therefore, scores on the statewide basic skills test in and of themselves would appear to be an indicator of program effectiveness, but in this case, no direct relationship can be established between these scores and skills acquired in the TBE program.

Another indicator of program success is grades in the mainstream. The students at all three grade levels in Program B who were included in this pilot study received average or above average grades throughout their years in the mainstream. The mean grades obtained ranged from C+ to B+ in all subject areas investigated, with the exceptions of a D+ average in Reading (1986-87 school year) for the group who were in Grade 7 at the time of this pilot study, and a D+ in Math (1987-88 school year) for the students who were in Grade 6 at the time of this study. Grades in Reading from school year 1986-87 were found to be significantly related to scores on the reading portion of the state basic skills test ($r = 0.99$, $p < .01$), and this could be considered an indirect measure of program effectiveness.

An additional indicator of program success is test scores in the mainstream. Once they had been exited, students in this pilot study from Program B did not maintain the pattern of steadily increasing mean scores on standardized tests that they had tended to exhibit while in the program. However, the mean gain in standard scores obtained on the standardized reading tests administered in school years 1986-87 and 1987-88 to the 7th grade group of students was found to be statistically significant ($t = 7.61$, $p < .01$), indicating that some meaningful growth had occurred. Again, it is recognized that small numbers of students participated in this portion of the pilot study, so these results should be interpreted with great caution. However, in this case, it would appear that test scores in the mainstream could be used as an indicator of program effectiveness.

DISCUSSION

As can be seen by the application of various evaluation models on the Program A data, in each case the results were similar in a certain sense -- regardless of the model used, the program appeared effective. However, each model attempts to answer different types of questions and so, in this way, provides very different evaluation results. On the basis of the Model A evaluation results, it can be said that students in Program A learn more than would have been expected had they not received any special services. On the basis of the mean standard scores evaluation results, it can be said that students show growth from pretest to posttest intervals. On the basis of the gap reduction evaluation results, it can be said that students in Program A narrow the gap between themselves and their mainstream peers from pretest to posttest time. Using out-of-program evaluation results, it can be said that program services appear to provide students with the skills necessary to compete in the mainstream, as reflected by scores on the basic skills test and by class grades. However, mainstream test scores do show a drop in performance.

Applying the various evaluation models to Program B results in a somewhat confusing and conflicting picture regarding the effectiveness of the services. However, this appears to be due primarily to the small number of matched test scores available for evaluation purposes. In fact, this particular program exemplifies the difficulty of using many evaluation models with bilingual programs. Often, due to problems such as mobility, models which stipulate matched pretest/posttest scores on annual testing by grade level result in very low numbers. This points to the need for alternative evaluation models which can accommodate such problems.

LIMITATIONS

It should be noted that some unfortunate limitations were imposed on the data analyses undertaken for this pilot study, all of which have implications for interpretation of the results. First, it should be emphasized that data collected on only a small number of students were used in the analyses. As the pilot study activities progressed, it became apparent that it would not be possible to obtain complete records on very many students given the variables of interest in this investigation. In part this was due to the inconsistent quality of student records encountered in the various school districts that were considered for inclusion in this pilot study, resulting in very few students having all the desired information available. Also, in some cases it was discovered that students participating in TBE programs are exempt from taking standardized tests administered as part of a district's regular program evaluation cycle, thus eliminating test scores that formed a major portion of the original data analyses.

Regardless of the reasons for the small numbers, the results of the analyses reported in this paper should be interpreted with caution. While many of the findings of this pilot study seem to verify the authors' hypotheses, it is believed that the reported results are most useful as indicators of trends that bear further investigation. Rather than take the numbers as absolute values, the reader is advised to consider them as evidence that a variety of approaches to evaluation of bilingual programs need to be studied, and to use the results of this pilot study to provide direction for future research into the most appropriate methods for evaluating these programs.

RECOMMENDATIONS

(1) There is no question that the evaluation procedures used with bilingual education programs need to be improved. Whether at the federal or state level, suitable evaluation approaches must be implemented to monitor accurately the quality of the education that language minority students are receiving. A complete evaluation of a program's effectiveness would entail examining all major components and addressing the different goals and objectives which guide the program's design and implementation.

Clearly, evaluation designs will vary depending upon the particular program goals being evaluated. But there are some key areas which should be addressed in any complete evaluation of the type being discussed here. One is the need to include a measure of educational progress of the participants as measured against an appropriate non-program comparison group. One very important aspect of bilingual education relates to preparing students to continue their education in the regular mainstream curriculum. In this case it is recommended that the most appropriate comparison group to use when evaluating these programs should consist of the program participants' mainstream peers.

Another equally valid and important evaluation question concerns individual student growth while in the program. If this is the evaluation

issue of interest, it is recommended that students be assessed on an annual basis to determine growth, using the program participants as their own comparison group. The important point is that a program should be evaluated on the basis of all major components, using a variety of comparison groups depending on the particular aspect of the program being assessed.

(2) While most educators would agree that a thorough program evaluation should include more than test scores, it is nevertheless the case that the majority of evaluation models proposed for use with bilingual education programs use the analysis of pretest/posttest scores as an integral part of the evaluation process. It is, after all, much more convenient to report, discuss, and compare scores obtained from tests that are routinely administered in school districts across the country than it is to collect other types of evaluation-related data. While it is recognized that scores from standardized testing programs can provide very useful data, it is nevertheless true that these scores can also be easily misinterpreted and misused.

It is recommended that those involved with the evaluation of Title VII Part A programs, or bilingual programs in general, should consider how the use of different analysis procedures may affect evaluation results as reflected by test scores. Furthermore, in order to interpret and use test results properly, it is necessary to be aware of the limitations of standardized test data in regard to making mainstreaming decisions and overall program evaluation decisions.

(3) This pilot study highlights the need for careful use of test data as part of the overall program evaluation process. However, it is imperative that additional evidence of program effectiveness be collected to supplement test data, and that all results be interpreted in light of the situation and analysis activities undertaken. Regardless of the specific evaluation model(s) used, it is recommended that the following data items be collected as part of a complete program evaluation:

- o information on language proficiency, including when and how it was determined for both English and the student's native language;
- o date of placement in the particular bilingual program;
- o educational history of the student, including schools previously attended and grades completed in the native country, where appropriate;
- o length of time spent in the program and a description of same, including program goals, staffing, instructional methodology, and entry-exit criteria; and
- o post program data covering a period of at least two years after mainstreaming occurred, including academic grades, attendance, and the results of any standardized tests administered during that time.

(4) One of the major limitations placed on data analysis undertaken for this pilot study related to the inconsistent quality of student records maintained by school districts considered for inclusion in this investigation. Complete information could be obtained on only a very small number of students, so the analyses could not be conducted as thoroughly as would have been desirable. It is recommended that a more systematic approach be taken to data collection and maintenance of student records. Increased efforts should be made to ensure that complete background, in-program, and follow-up information is gathered and then maintained in some easily accessible fashion. Not only would this expedite future investigations of program effectiveness but, even more importantly, this would also assist educators in designing programs that best meet the needs of their students.

TABLE 1

AVERAGE ATTENDANCE

Percent of Days Present

	In Program			In Mainstream	
	1982-83	1983-84	1984-85	1985-86	1986-87
Days Present					
170-180	19%	33%	55%	36%	55%
160-169	33%	43%	23%	46%	27%
140-159	34%	24%	18%	18%	18%
139 and below	9%	--	4%	--	--

TABLE 2

GRADES IN MAINSTREAM

Average Grades

Subject Areas	1985-86	1986-87	1987-88 (half year)
Reading	2.1	2.0	2.1
English	2.1	2.2	2.2
Math	2.3	2.2	2.3
Science	2.2	2.4	2.3
Social Studies	2.3	2.3	1.6

KEY A = 4
B = 3
C = 2
D = 1

TABLE 3

AVERAGE ATTENDANCE

Grade 6 - Percent of Days Present

Days Present	In Program 1982-83	1983-84	1984-85	In Mainstream 1985-86	1986-87
170-180	33%	11%	67%	56%	33%
160-169	33%	56%	11%	33%	33%
140-159	22%	33%	22%	11%	33%
139 and below	11%	--	--	--	--

Grade 7 - Percent of Days Present

Days Present	In Program 1982-83	1983-84	1984-856	In Mainstream 1985-86	6986-87
170-180	50%	50%	50%	25%	--
160-169	50%	25%	50%	75%	50%
140-159	--	25%	--	--	25%
139 and below	--	--	--	--	25%

Grade 8 - Percent of Days Present

Days Present	In Program 1982-83	1983-84	1984-85	In Mainstream 1985-86	1986-87
170-180	33%	33%	33%	100%	50%
160-169	33%	33%	66%	--	--
140-159	33%	33%	--	--	--
139 and below	--	--	--	--	50%

TABLE 4

GRADES IN MAINSTREAM

Grade 6 Average

Subject Areas	1985-86	1986-87	1987-88 (half-year)
Reading	2.4	2.3	2.5
English	2.6	2.6	3.0
Math	2.7	2.9	1.8
Science	2.8	2.7	2.3
Social Studies	2.4	2.9	2.4

Grade 7 Average

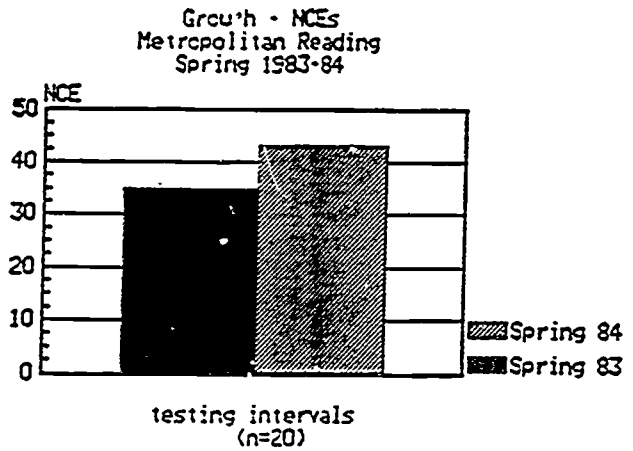
Subject Areas	1985-86	1986-87	1987-88 (half-year)
Reading	3.0	1.5	2.8
English	2.5	3.0	2.25
Math	3.5	3.25	2.25
Science	3.0	2.5	2.5
Social Studies	3.3	2.7	2.25

Grade 8 Average

Subject Areas	1985-86	1986-87	1987-88 (half-year)
Reading	2.3	2.5	3.0
English	3.0	2.0	2.5
Math	3.0	2.0	3.0
Science	2.7	2.0	3.0
Social Studies	2.9	2.0	4.0

KEY A = 4
B = 3
C = 2
D = 1

FIGURE 1



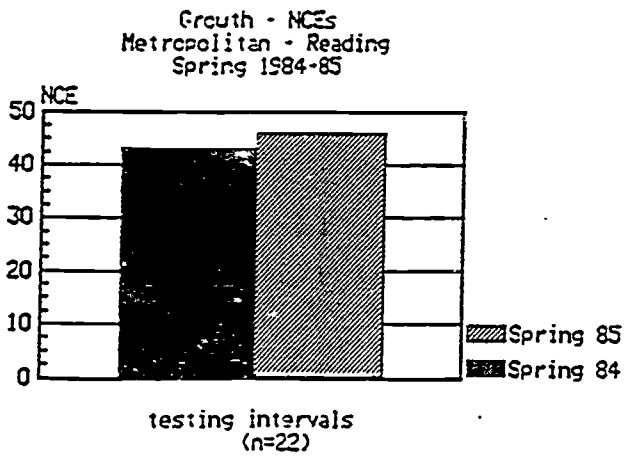
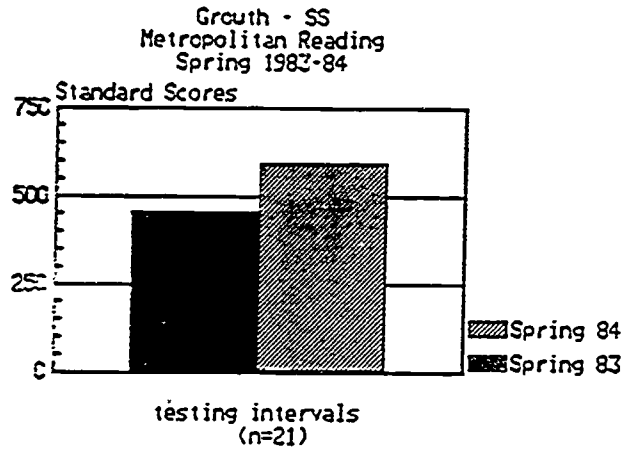


FIGURE 2



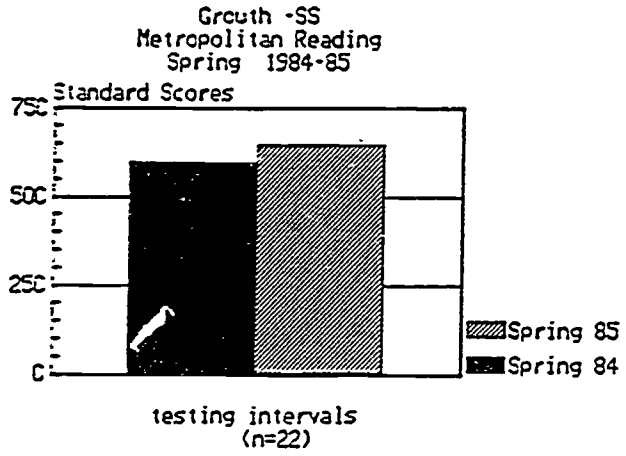
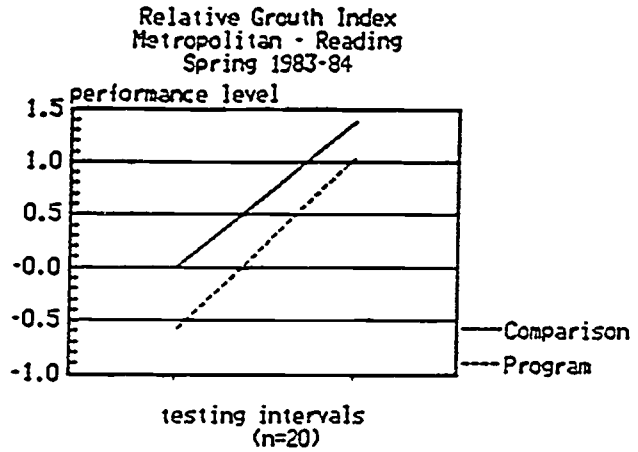


FIGURE 3


Relative Growth Index
Metropolitan - Reading
Spring 1983-84
performance level
1.5
1.0
0.5
-0.0
-0.5
-1.0
Comparison
Program
testing intervals
(n=20)

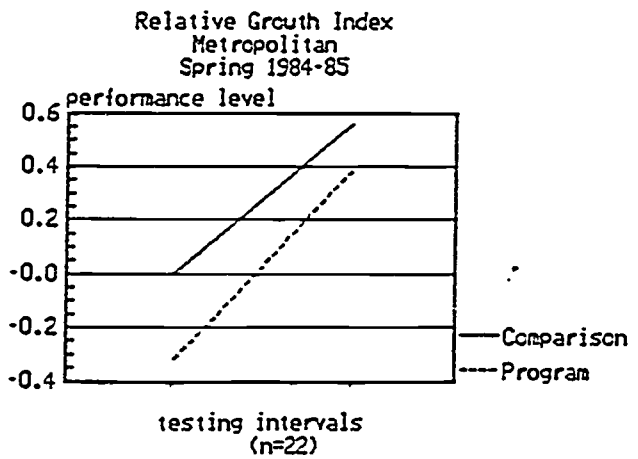
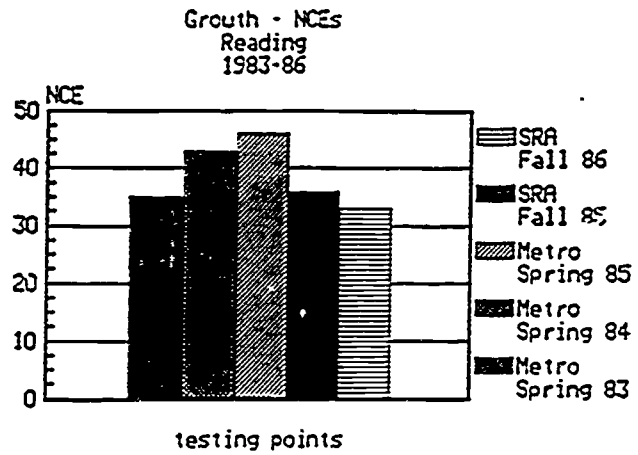

Relative Growth Index
Metropolitan
Spring 1984-85
performance level
0.6
0.4
0.2
-0.0
-0.2
-0.4
Comparison
Program
testing intervals
(n=22)

FIGURE 4


Grouth - NCEs
Reading
1983-86
NCE
50
40
30
20
10
0
SRA Fall 86
SRA Fall 85
Metro Spring 85
Metro Spring 84
Metro Spring 83
testing points

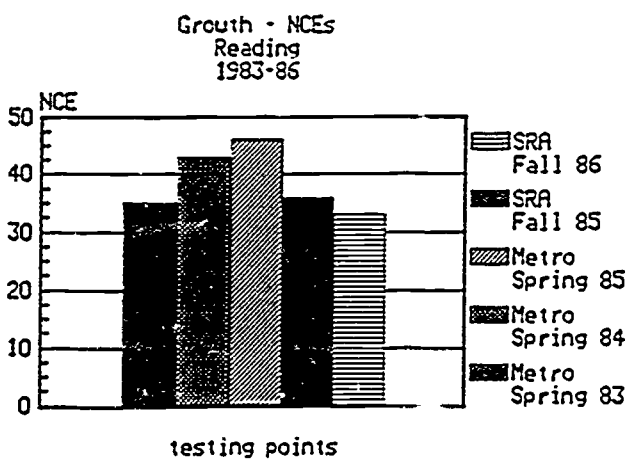
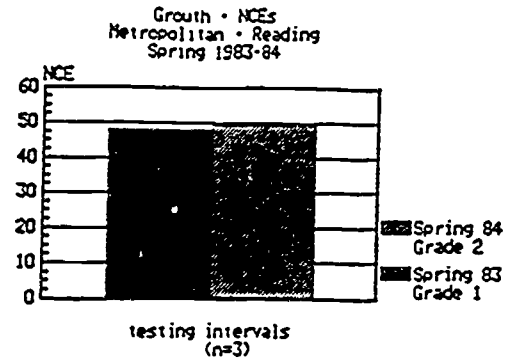
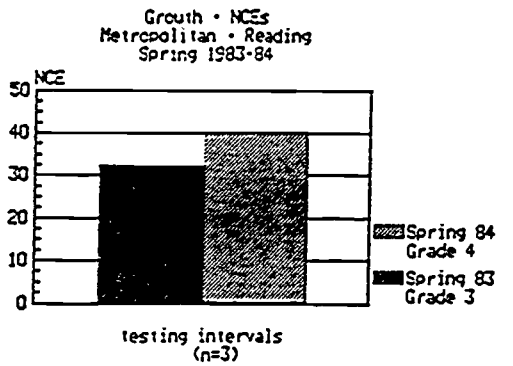
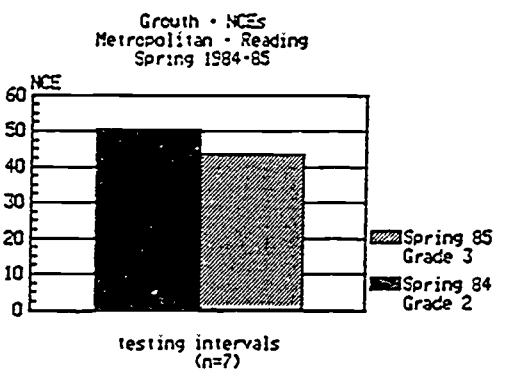
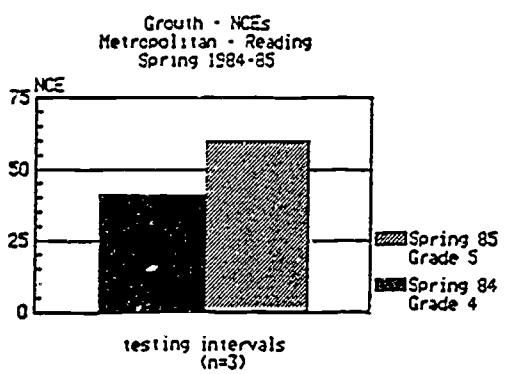

Grouth - NCEs
Reading
1983-86
NCE
50
40
30
20
10
0
SRA Fall 86
SRA Fall 85
Metro Spring 85
Metro Spring 84
Metro Spring 83
testing points

FIGURE 5



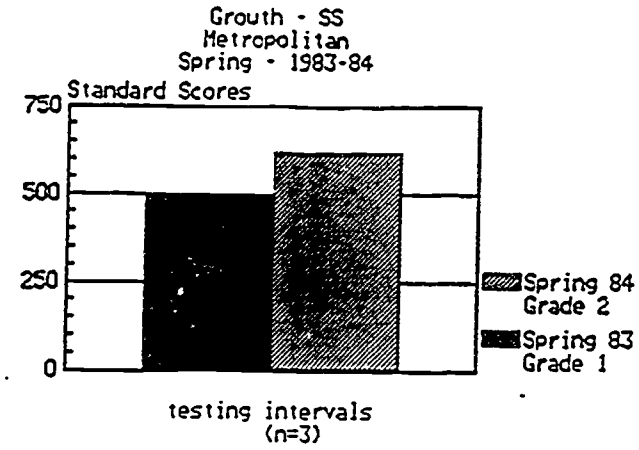


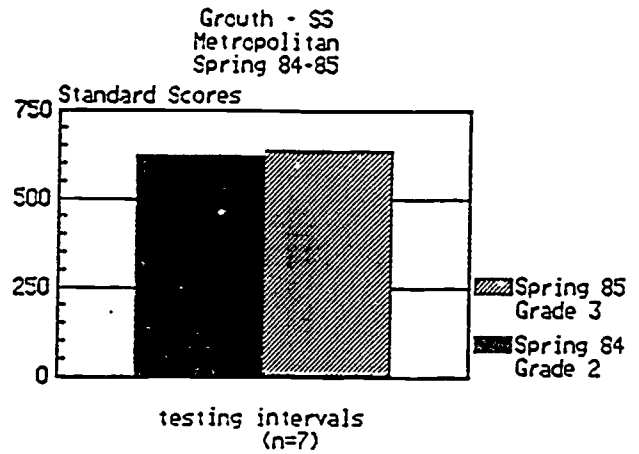


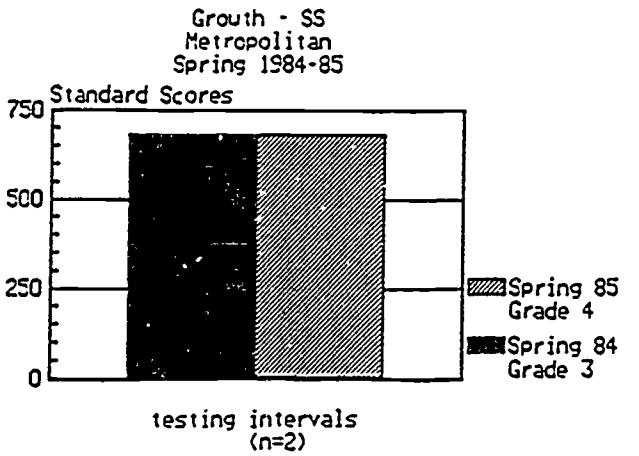


BEST COPY AVAILABLE

FIGURE 6

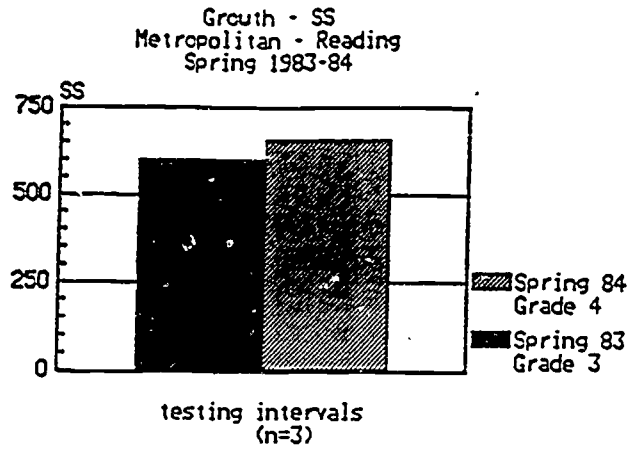


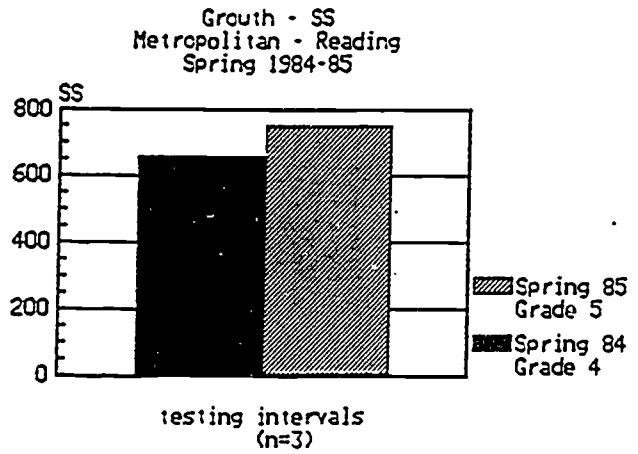




BEST COPY AVAILABLE

FIGURE 6 continued





REFERENCES

Baker, K.A., & de Kanter, A.A. (1983). Federal policy and the effectiveness of bilingual education. In K.A. Baker and 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.

Burstein, L. (1984). The use of existing data bases in program evaluation and school improvement. Educational Evaluation and Policy Analysis, 6, (3), 307-318.

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

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.

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

Okada, M., et al. (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.

O'Malley, J.M. (1984). Options for improving local evaluations. Rosslyn, VA: National Clearinghouse for Bilingual Education.

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

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

Reichardt, C.S. (1979). 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.

Tallmadge, T.K., Lam, T.C.M., & Gamel, N.N. (1987). The evaluation of bilingual education programs for language-minority, limited-English-proficient students: A status report with recommendations for future development. Washington, D.C.: U.S. Department of Education.

Walsh, E., & Carballo, E. (1986). Transitional bilingual education in Massachusetts: a preliminary study of its effectiveness. Quincy, MA: State Department of Education, Bureau of Transitional Bilingual Education.

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

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.