DOCUMENT RESUME

ED 222 083 FL 013 265

AUTHOR McConnell, Beverly B.
TITLE The View from the Firing Line: Evaluation of Bilingual Education Programs. Professional Papers MC-1.
INSTITUTION National Center for Bilingual Research, Los Alamitos, Calif.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Jun 82
NOTE 32p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Bilingual Education; *Bilingual Education Programs; Bilingual Students; Elementary Secondary Education; *Evaluation Methods; Program Effectiveness; *Program Evaluation; Research Methodology

ABSTRACT

Evaluations of bilingual education programs have received criticism from many sources. This criticism has been partially due to difficulties in obtaining hard data in bilingual settings, such as the lack of test instruments for bilingual children, the lack of appropriate reference groups, and the deficiencies of traditional evaluation models. The special problems involved in evaluating bilingual programs are described along with emerging solutions. Major problems have been encountered with the three most common evaluation designs: the control group model, the comparison group, and the norm reference group. Alternative evaluation design models include multi-year studies and designs in which children serve as their own control group. Such time series type studies allow for the consideration of developing English skills among the subjects. In evaluating bilingual programs, it is important to avoid rigid models and standardized tests that are likely to underestimate program effectiveness. References are provided. (RW)

THE VIEW FROM THE FIRING LINE: EVALUATION OF BILINGUAL EDUCATION PROGRAMS

Beverly B. McConnell

June 1982

[redacted] MC-1

national center for bilingual research, 4665 lampson avenue, los alamitos, CA 90720, telephone (213) 598-0481

The material in this publication was prepared under Cooperative Agreement 00-CA 80 0001 from the National Institute of Education, Department of Education. Its contents do not necessarily reflect the views of the National Institute of Education or of any other agency of the United States government, and no official endorsement should be inferred.

THE VIEW FROM THE FIRING LINE: EVALUTION OF BILINGUAL EDUCATION PROGRAMS

Beverly B. McConnell

Introduction

If evaluators of bilingual education programs were actors on the Broadway stage, the terrible reviews they have received from the critics would surely have closed the show. The General Accounting Office (GAO) found local project evaluations "of little use" (GAO, 1976). A former director of the Office of Bilingual Education reported that seven years worth of local evaluations had been thrown away because they were "not worth the paper they were printed on" (Epstein, 1978).

Is all this criticism warranted? Does it indicate an indifference to the need for evaluation by those who administer bilingual programs? My answer would be "No." Actually, the emphasis in bilingual education on accountability has led the way for other programs. It was one of the first of the Federally funded education programs to require a detailed evaluation plan to be submitted with every proposal. It funded not only local evaluation with every program, but program auditors who came in to see if evaluation claims could be substantiated.

Then what went wrong? Are the people incompetent who were responsible for all of these "worthless" evaluations? When the Federal government jumped into education with both financial feet in the 1960's, requiring annual evaluations of every Title I, Title III, and Title VII program, it resulted in a need for trained evaluators which far exceeded the supply. There were many inexperienced people doing evaluations in the early years. But the technology of evaluation has seen tremendous growth since then. There are now many trained professionals working in the field; however, the satisfaction with evaluation efforts in bilingual education has not notably increased.

My perspective on this matter comes from being part of the maligned majority, one of the people on the firing line, an evaluator of bilingual education programs working at the local level. The "view from the firing line" is that there is no lack of concern about the need for evaluation on the part of administrators, either local or Federal. And there are competent people trying to provide useful program data. There are, however, major problems in trying to obtain a fair assessment of learning outcomes when it involves students who are linguistically and culturally different from the mainstream child. Are there appropriate tests for children who speak a great variety of non-English languages? If the test is in English, how can you tell if the child didn't respond because s/he didn't know the answer or simply because s/he didn't understand the question? Are there appropriate norms and reference groups for non-English speaking, ethnic minority children?

The explanation for why it has taken a long time to begin getting "hard data" on the effectiveness of bilingual education can be found in the answers to these questions. Evaluators have had to develop a whole new technology in order to overcome the scarcity of adequate test instruments for bilingual children, the lack of appropriate reference groups, and the deficiencies of traditional evaluation models.

In this paper I, as an evaluator of bilingual programs, describe what I see as some of the special problems that arise in evaluating bilingual programs. Solutions to these problems are surfacing; these solutions will be discussed as well. Despite the problems, we are by no means without hard data on the effectiveness of bilingual education, and I have incorporated some of these findings from my own and other evaluations to illustrate my points.

Evaluating the Effectiveness of a Bilingual Program: Evaluation Designs

In order to answer the question "Does bilingual education make a difference?", the first requirement of an evaluation is to find some means of judging how the children would have done in the absence of the program. In the traditional technology of evaluation, the three most common methods of doing this involve using: (1) A control group; (2) a comparison group; or (3) a norm reference group. However, evaluators are likely to encounter major problems in trying to use an evaluation design involving any of these three methods when the children involved belong to an ethnic or linguistic minority group, as they do in bilingual education programs. The reasons for this are discussed in reference to each of the three traditional evaluation models in the sections which follow. I have also included recommendations gleaned from experience on how some of these problems might be overcome. The final section of the discussion describes some less traditional evaluation designs that may be promising in view of the unique requirements of evaluating bilingual programs.

The Control Group Model. The control group represents the traditional research design and is considered the "ideal" method from a technical standpoint. In this method, children are randomly assigned to a "treatment" and a "control" group and their progress compared. However, there have been a number of court decisions mandating special educational treatment for children who do not speak English (Applewhite, 1979). Therefore, randomly assigning some children to a control group, particularly to a "no treatment" control group, would very likely be illegal, ruling out this option for all practical purposes.

The Comparison Group Model: Identifying Key Variables. The second method usually proposed in the literature of evaluation is to have a comparison group that is not based on random assignment, but is

as nearly as possible like the group of children in the program being evaluated. Assuming that the two groups are matched on all key variables, the progress of children in the comparison group over a period of time can be taken as an indication of how the children in the bilingual program would have done in an alternative program or without any special program.

The hazard in applying the comparison group design to bilingual programs is that if the two groups are not matched on key variables, it will not only invalidate the results, but will also produce some very misleading and potentially harmful information. For example, most of the early research studies comparing bilingual and monolingual children failed to control for differences in socio-economic status in the two groups, resulting in a long string of research studies "documenting" that bilingualism was a "handicap." When the research community became aware that socio-economic status did have an important influence on test results, new research was undertaken with bilingual and monolingual children from the same socio-economic strata. These studies reversed the earlier findings, almost all of them finding a consistent advantage in bilingualism.

Few professionals today would carry out research or evaluations which did not take into account socioeconomic status as a factor that must be controlled in making comparisons between two groups. However, it took years of research to produce an awareness that socioeconomic status was a key variable. There are many other key variables that need to be controlled in order to obtain valid research or evaluation findings that involve linguistic minorities. The one most consistently ignored in evaluating bilingual programs is the initial level of linguistic competence in the groups being compared. The prime example of this can be found in the national impact study on the effectiveness of Title VII bilingual programs conducted by the American Institutes for Research and commonly referred to as the AIR Report, or the AIR Impact Study (AIR, 1977).

The AIR Impact Study selected 38 school districts from across the nation which had mature bilingual education programs. They asked each of these sites to nominate other classrooms within their own district or in a nearby district as a comparison group. It was specified that the comparison classrooms should have "students who would qualify for bilingual education and who were essentially the same "ethnic background, LINGUISTIC COMPETENCE (author's emphasis), and socioeconomic status as the students in Title VII project schools" (AIR, 1977; p. II-12).

The AIR research team ran into trouble with this evaluation design almost immediately. Of the 38 school districts included in the study, 18 indicated that they could not identify any comparison classrooms that met the criteria. This is not surprising. One reason is that by administrative mandate, any district receiving Federal funds for bilingual education is required to identify and provide special educational services for the children with the most limited English skills, i.e., those "most in need." Another reason is that before applying for funds, the district had to assess how many children needed the program; if the funds it received were near the amount requested, the district's bilingual program should be able to serve most of the limited English speaking children. This means that there would be no classrooms available within the district which would have a concentration of limited English speaking students to serve as a comparison group.

The option--i.e., to identify possible comparison classrooms in a "nearby district"--was most likely ignored. One reason may be that a school administrator does not have ready access to information about student characteristics in another school district. Another possible reason is that the administrator might hesitate to ask another district to take on this burden, since it is not a great privilege to be a

"comparison group"--there are no real advantages to the other district; and it takes the time of students, teachers and administrators, which is likely to be seen as a considerable disadvantage.

The other 20 districts in the AIR sample, however, attempted to be cooperative and did nominate comparison classrooms. These classrooms contained children of the same ethnic background and socioeconomic status as children in the bilingual program. It quickly became evident, however, that the comparison classrooms did not have children of the same level of "linguistic competence" as those in the bilingual program. The method used in the AIR Impact Study to judge the linguistic competence of the children was to have teachers classify students as English monolingual or Spanish monolingual; if bilingual, as English or Spanish dominant. On this basis it was found that, out of the entire national sample of over 1,600 Hispanic children in the comparison group, only 8 children were classified as "Spanish dominant bilingual," and 77 children were classified as "Spanish monolingual" (AIR, 1977). Since this was a national sample drawn from twenty different school districts, this represents an average of only 4 children per school district who were considered primarily Spanish speakers in the comparison group classrooms--hardly a viable sample on which to base any conclusions on how children "of limited English proficiency" would have fared in the school system without benefit of bilingual education.

If you add to the Spanish dominant children the 188 children classified as English dominant bilingual, the AIR study tested 273 children in the comparison classrooms across the nation who were either bilingual or Spanish monolingual. As shown in Figure 1, these 273 bilingual or Spanish monolingual children represent only 17% of the total sample in the comparison group; the remaining 83% were English monolingual children. In the Title VII bilingual classrooms, by contrast, 74% of the children tested were classified as either bilingual or Spanish monolingual.

LINGUISTIC CLASSIFICATION OF CHILDREN
TESTED FOR AIR REPORT

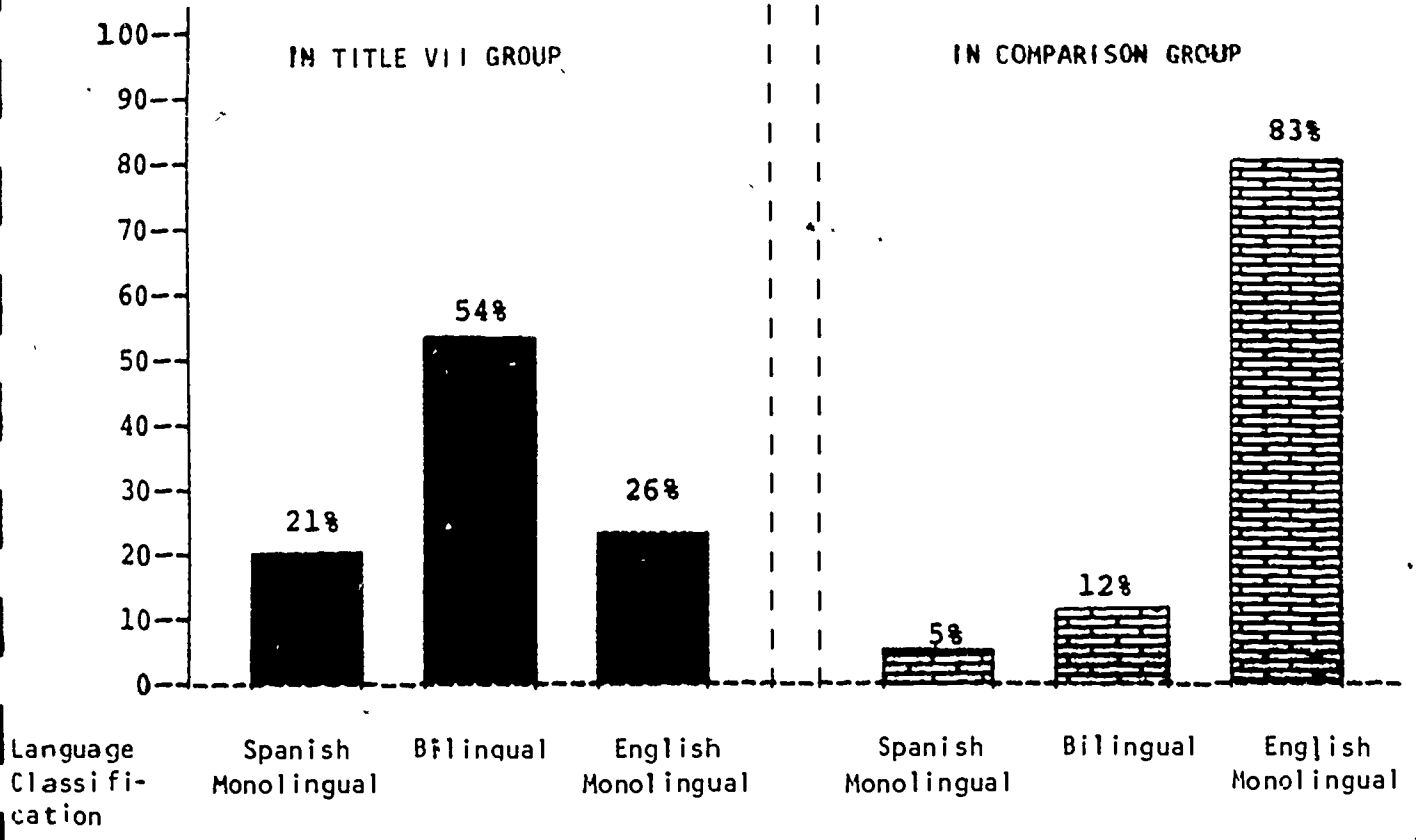

Figure 1: Percentage of children by language classification in the sample of children from Title VII classrooms and the sample from comparison group classrooms used in the national evaluation of bilingual education programs known as the AIR Report. (AIR, 1977, Appendix, pp. 123-127.)

To summarize the findings in Figure 1:

1. Of the children enrolled in the Title VII bilingual programs, 74% were classified as either bilingual or monolingual Spanish speakers. Only 17% of the children in the classrooms in the same districts which were nominated for use as a comparison group were classified as bilingual or monolingual in Spanish

2. The purpose of the comparison group was to provide a "credible estimate" of how non-English speaking or bilingual children would have progressed in traditional classrooms without benefit of bilingual education. The AIR study used for this purpose a sample of children, 83% of whom were identified by their teachers as speaking only English.

This presented a real dilemma for the AIR research team: How to answer the question, "Will children of limited English proficiency do better in bilingual classrooms than in regular, English-only classrooms?", when only the bilingual classrooms contained a significant number of children of limited English proficiency.

One way out of this dilemma would have been for the researchers to restrict their analysis to just those children in both groups who were limited in English. In fact, the original evaluation plan approved for the AIR study called for stratified analysis by children with comparable language classifications (AIR, 1977, pp. VI-17). However, there are limits to how small a group can be used for statistical analysis before the findings become so unstable that no valid inferences can be made. Out of the total sample of Hispanic children tested for the AIR study who were classified as Spanish Dominant Bilingual, 97% were in the Title VII group and 3% in the comparison group; for Spanish Monolingual children, 92% were in the Title VII group and 8% in the comparison group. Faced with this kind of a sample, the AIR research team abandoned the idea of stratified analysis by language groupings, stating only that "this approach was not feasible in view of the small numbers of students in the non-Title VII comparison classrooms who were given test questionnaires developed for Spanish monolingual students or for Spanish dominant bilingual students" (AIR, 1977, p. VI-17). Instead, the AIR researchers gave Spanish-speaking students tests that were below their actual grade levels; then, using expanded scale scores, they did their analysis across all language classifications for the total sample in both groups.

The other procedure researchers use when comparing groups that were not equal to begin with is to use analysis of co-variance. This was the procedure followed in the AIR study to attempt to compensate for differences in test scores resulting from the vastly different

composition of the two groups of children being compared. The assumptions on which this statistical procedure is based, however, make its use to compensate for differences in language competence highly questionable.

To use a race track analogy, analysis of co-variance makes a statistical adjustment only for differences in starting points (pre-test scores). If you start a race with one group 100 yards down the track ahead of the other, you can statistically "adjust" for this difference in reference to the final positions (post-tests) and "equalize" the starting points for the two groups. This assumes, however, that conditions are such that each group has an equal chance to make progress during the time the race is run. If you are teaching a subject area highly dependent on English skills, e.g., English language arts, and English is the first language of one group and the second language of the other, this is not a fair assumption. To return to my race track analogy, it is like letting one group take off running when the starting gun sounds, and making the other group stop and change their running shoes (from Spanish to English) before they can start. Statistical procedures cannot overcome differences that affect not only initial scores but learning processes as well. Comparison group evaluations in bilingual education which do not control for differences in language competence cannot produce a valid result, and no amount of statistical "magic" will overcome such a fundamental weakness in the evaluation design.

In stating their conclusions, the AIR Report was careful to eliminate "linguistic competence" from the final list of variables on which a match was found, indicating only that the Title VII and comparison classrooms were matched in terms of "(a) ethnicity, (b) socio-economic status, and (c) grade level" (Danoff, 1978, p. 3). This distinction, however, got lost "in the fine print" by most people who have quoted the study's findings. Since the purpose of Title VII funding was to provide special language services for children of

limited English proficiency, it is not unreasonable that readers of the AIR report would assume that a national study of its effectiveness would have considered "limited English proficiency" as a key variable on which both groups would be alike. It was taken as a shocking indictment of bilingual education, therefore, when the AIR Report stated, "In general, across grades, when total Title VII and non-Title VII comparisons were made, the Title VII students were performing in English worse than the non-Title VII students" (Danoff, 1978, p. 14). Probably few people would have been surprised if they had reported, "When the English language arts scores of Title VII children (74% of whom were bilingual or non-English speaking) were compared with scores of non-Title VII children (83% of whom spoke only English), the Title VII students were performing worse in English than the non-Title VII students."

How can a local program evaluator avoid the problems encountered by the AIR study in using a comparison group design? The syphoning off of the children with the least amount of English skills into the bilingual classrooms will be an obstacle to using a comparison group evaluation design in nearly every district receiving funds for bilingual education programs. The best alternative would be to test in another district that does not have a special bilingual program. Because this is an imposition on another school district, however, bilingual programs frequently continue to use the same comparison group test data for a number of years rather than testing each year. This procedure introduces some ambiguity into the findings from external events that might have influenced children tested in different time periods, and from the differences in outcomes that might be related to differences in the type of administrative support given in the two districts. As a tradeoff, however, these drawbacks to using a comparison group in another school district are probably more than offset by the advantage of having a comparison group more similar in linguistic competence than could probably be obtained by testing within the same school district.

In larger school districts bilingual funds may be concentrated into certain target schools with a larger population of limited English speaking students. This leaves a possible comparison group to be found among the limited English speaking students in schools where there isn't a sufficient concentration to justify a bilingual program. In this case it is probably better to test pockets of children of limited English proficiency from a number of different schools, making up a composite comparison group, instead of trying to select an intact classroom from another school which would probably contain very few bilingual students. It would also be better to stratify the analysis by levels of linguistic competence within both groups instead of using the classroom as a unit of analysis.

The most important guide to local program evaluation, however, is that a language test should be given at the same time as the academic achievement tests. The language test provides the basis for subgrouping children's tests based on their language classification to make a stratified analysis possible.

When an evaluation design subgroups children by language classification, the number in each analysis unit becomes smaller. In order to obtain a sufficiently large number of tests to make possible stratification by levels of language competence, it may help to eliminate some other types of subgroupings which are less important. Separate analyses by school, by individual classrooms, or by sex are not essential in measuring total program effect. It is, in fact, helpful to randomize teacher effect by using a unit of analysis larger than single classrooms. It is also possible to do a cumulative analysis, combining test scores for more than one program year. Any of these suggestions will increase the number of tests available for program analysis, making it possible to create subgroupings based on language differences and other critical variables.

With an appropriate match between program children and a comparison group, the comparison group model can be an effective means of measuring the effectiveness of bilingual programs. Without these precautions to assure an appropriate match, the evaluation can produce very misleading imformation that can do great damage in undermining the credibility of the bilingual instruction.

The Norm Reference Group. The third and most popular of the traditional methods used to establish a "no treatment" reference group is called "norm based" or "norm referenced" evaluation. There are actually several models of norm based evaluations with somewhat different statistical processes applied. They all have in common that they use standardized tests for which there are published norms based on the testing of some representative sample of children whose scores are taken to represent what is "normal" for children at a given age or grade level.

The basic premise of a norm based evaluation is that one can estimate how children would have done without the program by assuming that without any special intervention they would receive scores that bore the same relationship to the national norms on their post-tests as they did on their pre-tests. Since children's raw scores on a test would increase with age and grade levels, if a particular child's raw score was at the 50th precentile in reference to other children on the pre-test, his/her score would have to increase in proportion to that of all other children to remain at the 50th percentile on the post-test at a later age or higher grade level. Simple reporting of gains in raw scores from pre- to post-test will answer the question, "Have the children learned something?". Use of standard scores makes it possible to answer the question, "Have children learned more or less than they would have been expected to in the period of time between pre- and post-tests?". It provides a "standard" for judging the adequacy of the gain in terms of some reference group.

The reference group, however, is the source of the problem for using this model with linguistically and culturally different children in bilingual education. Unless the two groups are basically alike, what represents "normal" progress in one group may be quite different from what "normal" progress would be for the other group. To borrow an analogy used by Mercer (1973), men and women represent two groups that each have a different "normal" distribution of height, although there is some overlap. If the height of women as they grow taller is measured by the norm that was established by measuring a representative sample of men, the result will be that most women's growth will not represent "normal progress" over the years, and from our evaluation we will be forced to conclude that women, as a group, are slow "growers" and end up abnormally short.

Most nationally standardized tests base their norms on a representative sample of U.S. children, which would mean primarily English-speaking children from the majority culture. By definition, such a norm group will not resemble children who have been selected for a bilingual program because these children represents a linguistic and ethnic minority group. Thus, the basic assumption on which the norm based evaluation models rest is greatly weakened; i,e., that in the absence of a special program, children's academic growth would be the same as that of the norm group, so that at two different times their relative achievement level in reference to the norm group would remain unchanged. As in the analogy on the distribution of height among men and women, if two groups have a normal distribution of scores that do not fit the same curve, then projections from one to the other will not fit either.

How much bias will be introduced into the evaluation of bilingual program if standardized test norms are used as the basis for judging what children would have achieved without the program? It depends upon (1) how much the test that is used relies on knowledge of English; and (2) the degree to which the content of the test can be specifically

learned in a classroom situation and how much it requires "background" in language or cultural information that the majority group has access to and the minority group does not.

An illustration from three different tests that are part of the test battery used by the Individualized Bilingual Instruction (IBI) program (McConnell, 1981) will illustrate this point. Figure 2 shows the average score achieved by several hundred Spanish-dominant children in the IBI program over a period of years on the mathematics and reading subtest of the Wide Range Achievement Test (Jastak & Jastak, 1965) and on the Peabody Picture Vocabulary Test (Dunn, 1965). The first score represents children's test level when they first entered the program; then, left to right, their test scores after one, two and three years of bilingual instruction. All three tests are reported in terms of standard scores which, in this case, means that native English-speaking children with which the tests were normed would have a mean score of 100, with a standard deviation of 15. Since more people are familiar with percentiles than with standard scores, the percentage of children in the national norm group who had a score as low or lower than the standard scores reported for project children is shown in parentheses next to the test scores of children in the bilingual program.

The math test at this level was computational and, therefore, required little English; the skills it covered were all skills that would be specifically taught in the classroom, so there was little cultural bias relating to differences in background. This test would be the one on which project children would be most nearly on an equal footing with children in the norm group, and which could, therefore, be taken as the truest example of tneir basic ability to learn an academic skill.

The reading test was in English, which was a second language that had to be learned by all the children included in this particular sample of IBI children; for the children with whom this test was

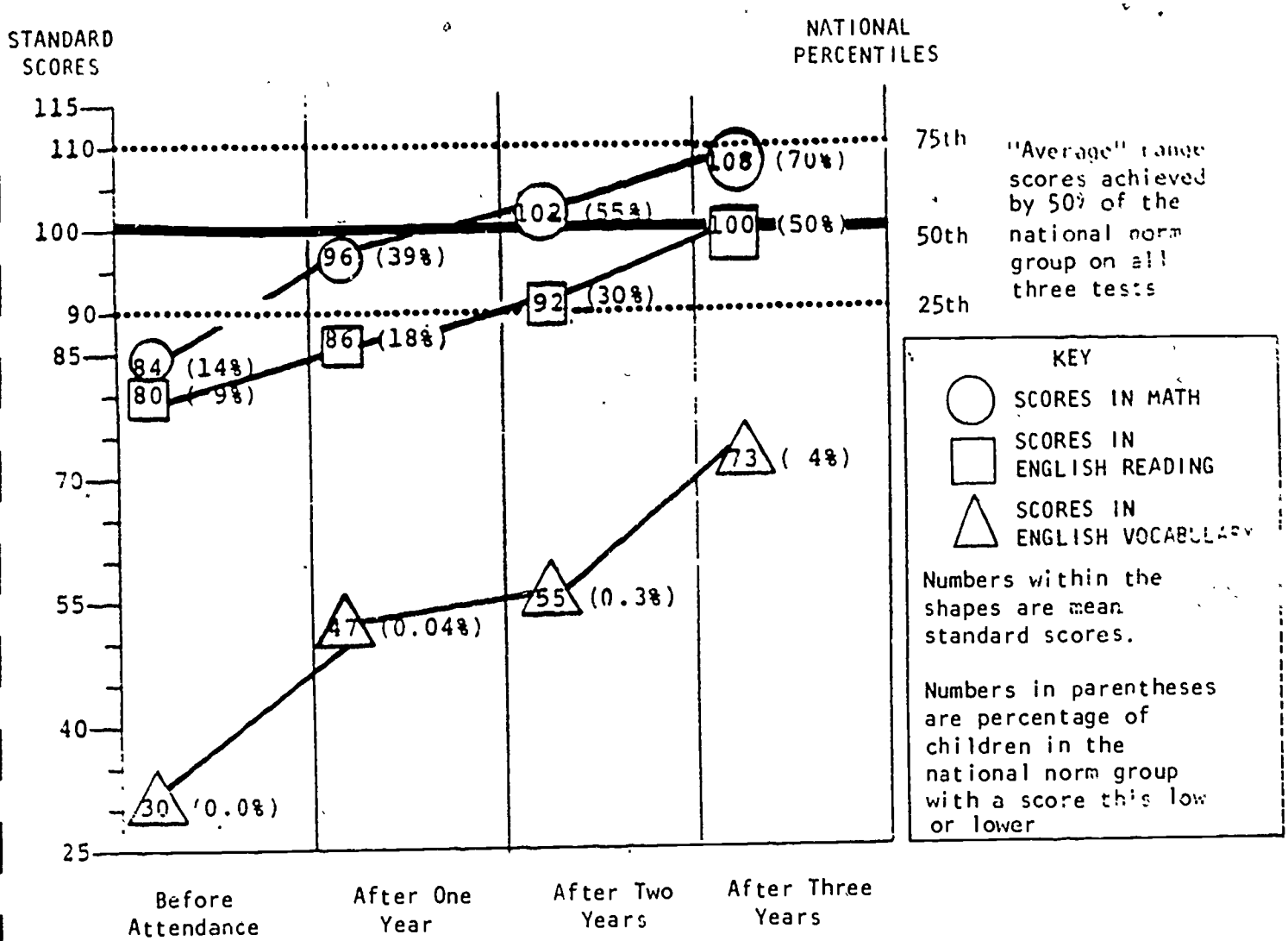

Figure 2: Mean standard scores and percentile ranks of Spanish speaking children enrolled in the IBI program in math and English reading from the Wide Range Achievement Test, and in English vocabulary from the Peabody Picture Vocabulary Test, after differing periods of time in the IBE bilingual program.

To summarize the findings in Figure 2:

1. In math, which does not depend much on English skills, IBI children were able to achieve scores in the "average" range based on national norms after only one year of bilingual instruction. After two and three years, their scores exceeded that of 55% and 70% of the children in the national norm sample, respectively.

2. In English reading, which was a second language for the IBI children, it required two years to achieve scores in the "average" range, and three years to reach a score achieved by 50% of the native English speaking children with whom the test was normed.

3. In English vocabulary, a test affected by both language and cultural differences, the IBI children had an initial average score of 30. After three years, the average standard score, or verbal "IQ" score, had risen to 73, a score exceeded by all but 4% of the norm group sample. This can be taken as an indication of test bias rather than limited ability since these are the same children who are achieving at or above national norms in reading and math.

normed, English was, of course, a first language. There is relatively little cultural bias in this particular test, however, in that it measures specific skills which can be taught in a classroom--matching, letter recognition, and decoding skills.

The English vocabulary test represents the most extreme difference between project children and the test norm group. The language difference is compounded by the content, which represents identification of objects and actions that a child would be familiar with from "life experience" rather than specific classroom teaching. The standard scores for this test are also reported as verbal "IQ" scores. A great deal of damage has been done to linguistic and ethnic minority children from misuse of verbal IQ scores, which is one reason why the data shown in Figure 2 are very important to an understanding of the issues involved in using standardized tests with linguistic minorities.

As shown in Figure 2, these Spanish-speaking migrant children scored very low on all three tests when they first enrolled in the bilingual program. In all cases, their scores were more than one standard deviation (i.e., 15 points) below the mean established by the norm group. The band between the standard scores of 90 and 110 in the Figure represents the "avarage range" for the middle 50% of children in the norm group between the lowest 25% and the top 25%. It took project children one year to get into this "average range" in math, and by the second and third year their average scores were higher than 55% and 70% of the norm group children, respectively. For English reading, it required two years for IBI children to reach the "average range," and three years for them to reach the national norm, i.e., the score that was achieved by 50% of the norm group children.

In English vocabulary, the children started with a standard score, or "IQ," of 30 and it took them three years to reach an average standard score of 73--a score exceeded by all but 4% of the middle-class, majority culture, English-speaking children with whom the test

was normed. For a majority group child, a verbal IQ score below 70 would be taken as an indication that special education classes should be considered. All of the children in the IBI program would have been considered candidates for special education on this basis. Their educability, however, is demonstrated by the academic gains they made in math and reading, when given access to bilingual instruction.

If the vocabulary test alone had been the basis for evaluating the bilingual program, would the U.S. public consider a program successful if it reported that, after three years, children's English vocabulary was only as good as the lowest 4% of U.S. school children, as represented by the test norm group? On the other hand, would a program be considered successful that enabled the sons and daughters of Spanish-speaking migrant farm workers to excell 70% of U.S. school children in math, and 50% of U.S. school children in English reading skills after three years of bilingual instruction? Same program. Different tests, with different degrees of linguistic and cultural bias.

Many of my colleagues who are evaluating bilingual programs were not native speakers of English. They have received advanced university degrees, but many of them still speak with some bitterness of having their permanent school record show that at kindergarten they had an IQ score of zero. Some were assigned to classes for the educable mentally retarded, and only got back into mainstream classes when some adult had courage enough to demand their reassignment. It is probably not surprising that many of them have refused to use standardized tests in their program evaluations, devising local tests and criterion-referenced tests designed only to see if children are learning what they are being taught. These evaluations are often very helpful at the local level to provide feedback to teachers on what children are or are not learning. At the national level they are considered to be "so

what" evaluations because they provide no standard of what the children might have learned under different circumstances or with no special program at all.

Norm based evaluations are by far the most common type used by schools to evaluate all sorts of programs. This is because they are the most economical, in that only program children need to be tested (the reference group has already been tested, as represented by the published norms). Standard scores are also very useful in comparing one program to another, in that even if different tests are used, if the scores are all reduced to a common metric (variance from a mean), it is possible to compare children's gains from one project to another. It seems unlikely, therefore, that the use of standardized tests and norm reference groups can be totally avoided in evaluating bilingual programs. That being the case, the evaluator can only take steps to minimize the distortion of project findings that may result from their use.

One such step is to use a number of different measures in evaluating a program, if this is at all possible. A range of findings will make it easier to tell if one particular test reflects an unusual amount of test bias.

If it is possible to combine the use of standardized tests with a comparison group evaluation design, or a design such as the one described in the following section which allows project children to serve as their own controls, most of the effects of test bias can be overcome. All statistical analysis in such an evaluation compares children who share the same ethnic and language background; any test bias, therefore, affects all groups equally. The question, "How would the children have done without the program?", in this case would be answered by comparison to the children representing the same ethnic and language group. The question, "How does the achievement of children in the program compare to that of native English speaking children?", may

be answered by reference to the test norms. This adds to the interpretive power of the evaluation but does not base the "no treatment" estimate on an inappropriate reference group.

There may be circumstances in which there is no basis for evaluation except the test norms, such as a statewide evaluation which aggregates the findings from several programs. An example of this is shown in Figure 3, which illustrates data from a 1980 evaluation of bilingual education programs in Colorado. Since several different tests had been used by participating schools, all scores were converted into normal curve equivalent (NCE) standard scores. This standard score scale uses a mean of 50 and a standard deviation of 21. The NCE scale was the one either required or recommended for use by most Federally funded programs under Title I of the Elementary and Secondary Education Act. The U.S. Department of Education established an arbitrary guideline that a gain in a school year of 3 or more NCE's would be considered an educationally meaningful gain, and a gain of 7 NCE's (i.e., one third of a standard deviation) would constitute a gain worthy of labeling the program an "exemplary" program.

The Colorado evaluators, Egan and Goldsmith (1980), therefore, reported any classroom in which the mean score was within plus or minus 3 NCE's of the pre-test NCE level as one in which there had been "no change," a post-test that dropped by more than 3 NCE's as a negative effect, and one that gained more than 3 NCE's as one that had brought about significant improvement in children's educational performance. As Figure 3 illustrates, almost two thirds of the bilingual classrooms (64 out of 100 statewide) reported significant improvement in test scores, with more than half of these (34 out of 64) showing a gain of 7 or more NCE's, the measure of an "exemplary" program. They reported 87% of the programs as "successful" based on either showing "no change" or "significant improvement."

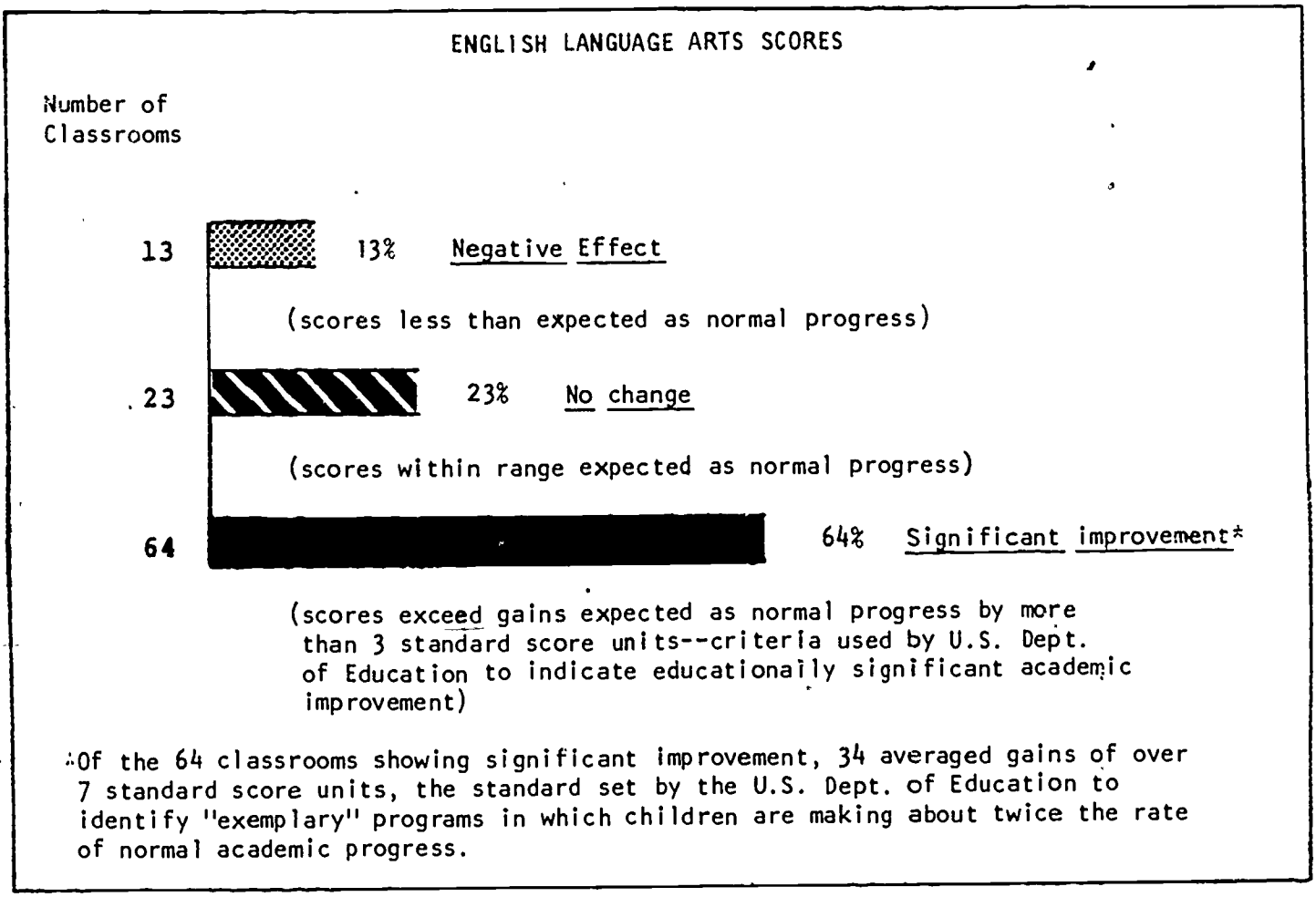

Figure 3: Pre- to post-test gains (1979-80 school year) in English language arts of linguistically different children from Colorado bilingual education classrooms (Egan & Goldsmith, 1980).

To summarize the findings in Figure 3:

1. Of 100 Kindergarten through third grade classrooms in Colorado's bilingual education programs, the majority, or nearly two thirds, produced significant gains in English language arts for linguistically different children enrolled.

2. Over one third of the classrooms (34) produced gains nearly twice the expected rate of normal progress, equaling a standard set by the U.S. Dept. Of Education to identify "exemplary" educational programs.

3. Gains within the range expected as normal progress were found in 23% of the classrooms, and 13% of the classrooms reported test scores indicating children made less than normal progress for the year.

Inclusion of "no change" classrooms as successful is soundly based on research that has shown that children from a poverty level or from ethnic and linguistic minorities do not, in fact, "hold their own" in traditional classrooms in reference to the norm group for most standardized tests. Without special programs, they tend to have lower standard scores as their grade levels increase (Coleman, 1966; Linn, 1979).

Use of a range of scores to define categories of program success or failure as was done in the Colorado evaluation also helps to minimize some of the lack of "fit" between the normal distribution of scores in a minority and a majority student population.

Overall, the aggregation of test scores across programs is a poor way to measure program effectiveness. Distortions built into statistical analysis, test bias and testing error, and lack of test comparability are all compounded when this approach is used. Also, language of instruction is not the only element involved in children's learning. It would have been considered ludicrous if some of our more famous research studies in education, such as the first grade reading studies (Stauffer, 1966), had been carried out with no further attention to curriculum differences and classroom variables than the information that all classes were conducted "in English." It should be equally obvious that not all programs of bilingual education are alike. However, if legislators require an indication of whether a class of programs is, in general, successful or unsuccessful, the approach demonstrated in the Colorado evaluation is probably as appropriate as any that could be devised. It is certainly much better than one that combines test scores from all programs into a statistical blender that produces a "mean gain" or loss across all participating children, regardless of language background, and across all classes, regardless of differences in instructional approach.

Alternative Evaluation Design Models

An extensive discussion of evaluation design alternatives is beyond the space limitations of this paper. There are two design factors which should be mentioned, however, because they have particular relevance to the unique requirements of evaluating bilingual programs. The first is the importance of building a multi-year dimension into the evaluation design.

Most educational evaluations involve pre- and post-testing at the beginning and end of one school year. Evaluation that is done over such a short time period is likely to be quite misleading as to the value of the bilingual education program. One reason is that in the bilingual classroom the scheduling of class time and the sequencing of instruction are often different than in the traditional classroom. For example, English reading instruction begins in kindergarten or first grade in regular classrooms; in the bilingual program, if reading is started in the child's primary language (other than English), it might mean that instruction in English reading will not begin until second grade. The initial comparison between children in the bilingual program and their counterparts in a regular program would logically show a negative effect on English reading in first or second grade comparisons, whereas in the long run the child's literacy in a second language might enhance English reading skills.

This appears to be the case in the data from the bilingual program at Rock Point School on the Navajo reservation (Rosier & Holm, 1980). As shown in Figure 4, children in the bilingual program who began reading in Navajo and did not start English reading until second grade were initially behind children in the comparison group taught only in English. The advantage shifted to the bilingual program by third grade and increased each year thereafter. But it was not until fifth and sixth grade that the superiority of children in the bilingual program was enough to be statistically significant. By sixth grade the

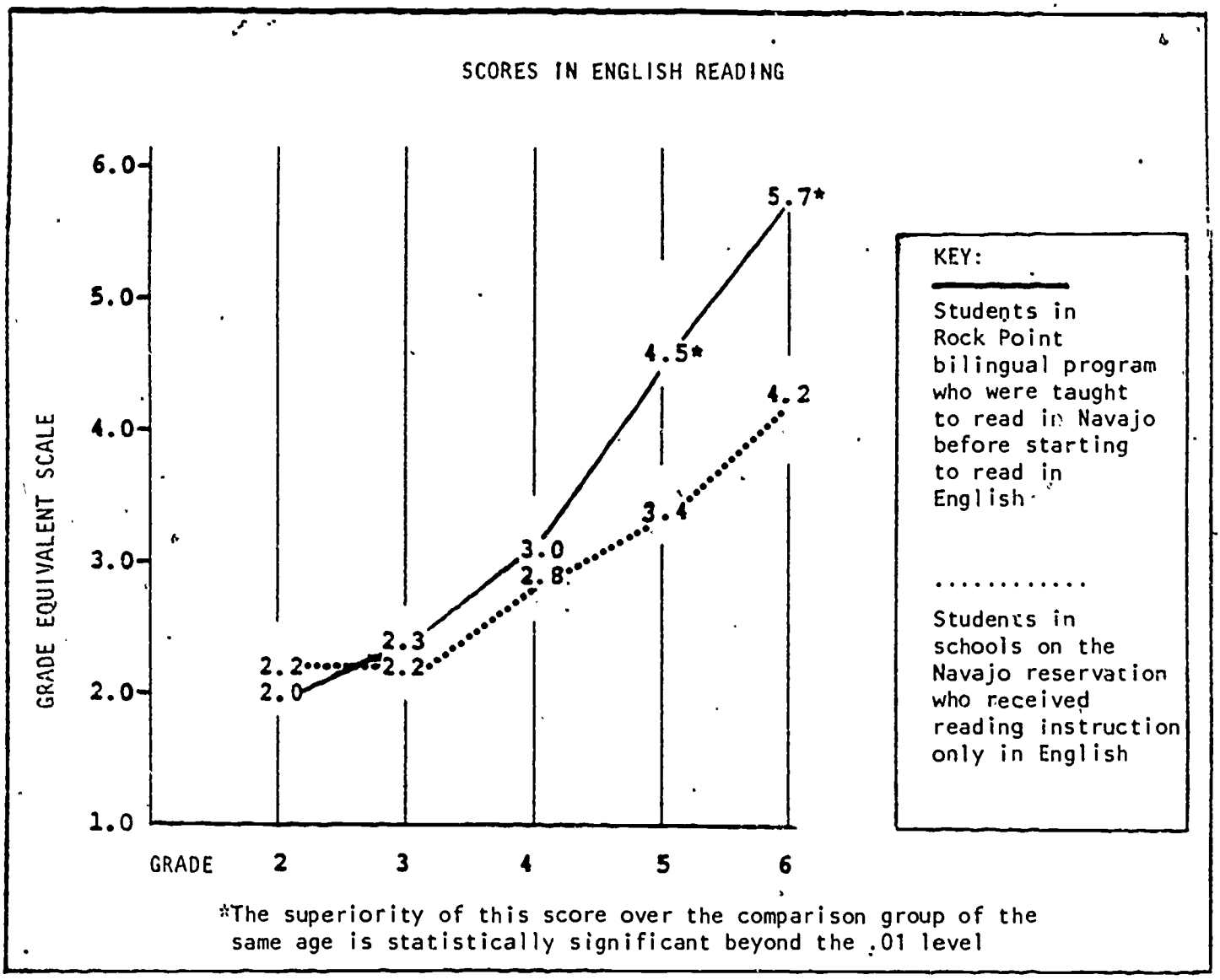

Figure 4: 1977 mean grade point equivalent scores in total reading for children in the Rock Point bilingual program and 1975 scores for a comparison group on reservation schools using only English instruction, by grade level, using the Stanford Achievement Test. (Source: Rosier & Holm, 1980)

To summarize the findings in Figure 4:

1. During the first year that children in the Rock Point bilingual program started learning to read in English, their average reading scores were below that of children in the comparison group who had been reading in English since first grade.

2. By grades three and four, children from the bilingual program had a small superiority in English reading over children from the comparison schools, but the difference in scores was not enough to be statistically significant.

3. By grades five and six, the children in the bilingual program showed a marked superiority over children from the comparison schools, and the difference in scores was large enough to be statistically significant.

children in the bilingual program were reading at a grade level nearly two years above that of their counterparts in other reservation schools who had been taught reading only in English.

If the evaluation of this program had depended solely on evaluation within each school year, the program would probably have been considered a failure. The evaluation would have reported that in a majority of the classrooms there were no significant differences between children in the bilingual program and the comparison group. The pattern of long term benefits would have been overlooked.

There are other reasons why multi-year evaluation is especially critical for bilingual education program. Increased English usage is usually phased in over a number of years. This means that at some point children will be switched to testing in English. Initially, unfamiliarity with the specific English vocabulary of an academic area the child has studied in another language may produce test scores below the level the child should be earning based on his or her understanding of the basic concepts. However, this disadvantage should wear off as the child becomes more familiar with English. Observing test scores over more than one year will help to see if the results in a particular transition year were artificially low. Single year evaluations will not pick up these patterns.

The second recommendation regarding evaluation design for bilingual programs is to use an evaluation design that will permit project children to serve as their own controls. Given the great difficulty of finding comparison groups matched on ethnic, socio-economic and language variables, this type of design eliminates many of the problems of comparability and selection bias found with other evaluation approaches.

In evaluation manuals these types of evaluation designs are usually described as "baseline," "historical" or "time series" models.

An example of this type of evaluation would be to obtain test scores from children in a time period before a particular program is put in place, and then use these as the "baseline" for comparison to the test scores of the same or other children of comparable age or grade level who have been tested at some time after the new program was initiated. Another variation that is used in educational experimentation is to phase in a treatment and to use the score of part of the children who have not yet had the treatment as the comparison to children who may be in various phases of completing the treatment.

The IBI program has used an adaptation of a baseline and time series evaluation model. It allows stratified analysis by language group, and provides a multi-year perspective on changes occurring through bilingual instruction. Some of the features of this evaluation design which might be used by other bilingual programs are described below.

Children are pre-tested when they first enroll in the IBI program, using language tests and a number of academic achievement tests. These scores go into a data bank which is being added to continuously as new children enroll in the program. Each individual child's pre-test score represents the progress that child has made to that age level without benefit of bilingual education. Collectively, it provides a standard against which to measure any special effect that can be observed through participation in the bilingual education program.

Post-testing in the IBI program is done at intervals based on children's actual accumulated attendance. This i because the IBI program enrolls migrant children whose attendance is very irregular. For other programs it might be enough to schedule yearly testing and simply note for analysis how many years that child has been receiving

bilingual instruction. The analysis then compares the scores of children who have attended for one-half year, one year, two years, etc., with a "no attendance" group of children from the pre-test data bank who are matched to children in the post-test group on language classification, age, etc.

IBI serves children who come and leave irregularly throughout the year; therefore, it has selected tests which do not need to be administered during particular months in order to use the test norms. However, if a test was used with Fall norms (Fall being the time when most pre-tests would be given), the subsequent post-tests should also be given in the Fall, and the evaluation should be based on year-to-year analysis, adding the dimension of children enrolled two years, or three years or more with each subsequent year's evaluation as the program matures.

Because IBI enrolls some new children in every grade level each year, it has been possible to accumulate tests for comparative analysis for every age or grade level. However, by using standard scores, which would allow analysis across age levels, a program could still use this model even if a sufficient number of tests were not available in the pre-test data bank to do a separate analysis for every grade level.

By using internal comparison, project children can also serve as their own controls for many kinds of evaluation. Examples of this would be 1) comparing progress of children receiving one type; and 2) comparing children in classrooms with fully bilingual teachers and those without. In the end, this type of evaluation may be the most fruitful. Court cases have already decided the question of whether children who do not speak English deserve some type of special educational program. Internal comparisons will provide answers to what may be the most relevant question, not "Does bilingual education work?", but, "What types of bilingual education work best?".

Conclusion: Solutions and Non-Solutions

This is the "age of accountability" in education. We cannot ignore the echoing cry from nearly every quarter that more definitive evaluations of bilingual education programs are necessary. Those of us who are involved in bilingual education can readily agree that there are "problems" that need to be solved to make program evaluations better. We need also to be aware that there are "solutions" and "non-solutions" to these problems among the various remedies being proposed. A "non-solution" is one that seems likely to produce evaluations that will underestimate or otherwise distort the effectiveness of bilingual programs.

One of the "non-solutions" regularly proposed is to lock bilingual programs into using a limited number of standardized tests. At this point, test development for linguistically and culturally different students is still in its infancy, and we are not at a point where we can afford to freeze test development by enshrining certain tests in legislation or state and Federal regulations.

Another "non-solution" is to adopt mandated evaluation models such as those developed for Title I Programs. With the Title I evaluation models currently identified, there is the very real possibility that districts will end up conducting evaluations using inappropriate comparison groups or standardized test norms such that they would produce an invalid result with linguistically different children. Grinding out evaluations in quantity that distort the gains children are making in bilingual programs will not serve any useful purpose.

On the other hand, we have every reason to be encouraged that the potential of bilingual education is being documented by "hard data" coming out of a growing number of programs. The Rock Point study cited in this paper has shown that children receiving bilingual instruction make significantly greater progress in learning to read in English

after they have first learned to read in Navajo than their counterparts in other reservation schools taught only in English. This study has been replicated with several successive classes of students, which lends even greater confidence to the result. The IBI program with which I am associated now has ten years of data on participating children. As shown by the data cited in this report, children who entered with English reading scores below the 10th percentile compared to national norms, after three years of bilingual instruction are able to score at or above national norms established for children who are native speakers of English (McConnell, 1981). The Colorado data represents bilingual programs in 38 school districts with two thirds of the classrooms reporting educationally significant gains for linguistically different children (Egan & Goldsmith, 1980). With this type of data, perhaps we can sustain the commitment to bilingual education that will be necessary to explore the promise it may offer for children who have clearly suffered under the historic alternatives our educational system has tried.

References

American Institutes for Research in the Behavioral Sciences, Palo Alto, California. Evaluation of the Impact of ESEA Title VII Spanish/English Bilingual Education Programs. Vol.I: Study Design and Interim Finding. Feb. 1977, 565 pp. (ERIC Document Reproduction Service ED 138 090)

Applewhite, S. R. "The Legal Dialect of Bilingual Education." in Padilla, R.V., (Ed.). Ethnoperspectives in Bilingual Education Research, Vol. 1: Bilingual Education and Public Policy in the United States. Ypsilanti, Michigan: Department of Foreign Languages and Bilingual Studies, 1979.

Coeman, J. S., et al. Equality of educational opportunity. Washington, D.C.: U.S. Government Printing Office, 1966.

Danoff, M. N. Evaluations of the Impact of ESEA Title VII Spanish/English Bilingual Education Program. Overview of Study and Findings. Palo Alto, California: American Institutes for Research in the Behavioral Sciences, 1978, 32 pp. (ERIC Document Reproduction Service Number ED 154 634.)

Dunn, L. M. Peabody Picture Vocabulary Test. Circle Pines, Minnesota: American Guidance Service, Inc., 1965

Egan, L. A. and Goldsmith, R. Bilingual Bicultural Education: The Colorado Success Story. Colorado Dept. Of Education. Duplicated. 1980.

Epstein, N. Language, Ethnicity, and the Schools. Washington, D.C.: Institute for Educational Leadership, George Washington University, 1978.

General Accounting Office. Bilingual Education, An Unmet Need. Washington, D.C.: Government Printing Office, May 1976.

Jastak, J. F., & Jastak, S. R. The Wide Range Achievement Test. Wilmington, Deleware: Guidance Associates, 1965.

Linn, R.L. Validity of inferences based on the proposed Title I evaluation models. Educational Evaluation and Policy Analysis, Vol. 1, No. 2, 1979, pp. 23-32.

McConnell, B. Long Term Effects of Bilingual Instruction. Pullman, Washington: Bilingual Mini Schools, 1981a.

McConnel, B. "Plenty of Bilingual Teachers." In Padilla, R.V., (Ed.). Ethnoperspectives in Bilingual Education Research, Vol. III: Bilingual Education Technology. Ypsilanti, Michigan: Department of Foreign Languages and Bilingual Studies, 1981b. (ERIC Document Reproduction Service Number ED 206 203)

Mercer, J.P. Labeling the Mentally Retarded. Berkeley, California: University of California Press, 1973.

Rosier, P. and Holm, W. "The Rock Point Experience: A Longitudinal Study of a Navajo School Program (Saad Naaki Bee Na'nitin)." Bilingual Education Series No. 8. Arlington, Virginia: Center for Applied Linguistics, 1980.

Stauffer, R.G. (Ed). The whole of Issue No. 8. The Reading Teacher, Newark, Delaware: International Reading Association, Vol. 19, May, 1966.