DOCUMENT RESUME

ED 270 254 RC 015 759

AUTHOR Murray, Stephen L.
TITLE State Education Agency Options for Evaluating ECIA Chapter 1 Migrant Education Programs.
INSTITUTION Northwest Regional Educational Lab., Portland, Oreg.
SPONS AGENCY Department of Education, Washington, DC. Office of Planning, Budget, and Evaluation.
PUB DATE 2 May 86
CONTRACT 300-82-0377; 300-85-0198.
NOTE 77p.
PUB TYPE Guides - Non-Classroom Use (055) -- Reports - General (140)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Achievement Gains; *Basic Skills; *Educational Assessment; *Evaluation Methods; Evaluation Needs; Migrant Education; *Migrant Programs; Minimum Competencies; Norm Referenced Tests; Pretests Posttests; Program Development; Program Effectiveness; *Program Evaluation; State Programs; Testing Programs
IDENTIFIERS *Education Consolidation Improvement Act Chapter 1

ABSTRACT

Practical and technical advice is provided for staff responsible for planning evaluations of Educational Consolidation and Improvement Act (ECIA) Chapter 1 migrant programs at the state level. The report is limited to evaluating achievement gains in basic skills; this is, however, acknowledged as only one facet of a comprehensive migrant program evaluation. Existing evaluation approaches found in annual migrant evaluation reports for the 1981-82 and 1982-83 schools years are discussesd. Guidelines for developing a state plan include building a program profile of instructional services offered to migrant students, setting priorities for evaluation, and selecting or developing evaluation strategies in relation to the priorities. The report outlines elements of a program profile and discusses four evaluation approaches--norm-referenced evaluation, pre-post-matched scores, post-test only, and state assessment programs. Purpose, design features, testing features, aggregation, strengths, limitations, conditions supporting use, and recent use are presented for each evaluation type and information is summarized in a table. An appendix provides instructions for using pretest scores to select Chapter 1 participants in a norm-referenced evaluation model. A sequence of eight steps for implementation is laid out in detail. (LFL)

The analysis reported herein was performed pursuant to Contract Numbers 300-82-0377 and 300-85-0198 with the U.S. Department of Education. Contractors undertaking such projects under government sponsorship are encouraged to express freely their professional judgment in the conduct of this project. Points of view or opinions stated do not, therefore, necessarily represent official U.S. Department of Education position or policy.

ACKNOWLEDGEMENTS

This monograph benefited substantially from reviews by a number of readers representing the United State Department of Education (ED), the National Association of State Directors of Migrant Education (NASDME), and the four Regional Chapter 1 Technical Assistance Centers (TACs).

The migrant evaluation workgroup, assembled by ED, met five times to plan and review three major products dealing with Chapter 1 migrant program evaluation. Comments and suggestions by these colleagues contributed to shaping this paper.

Representing ED:

James English
Howard Essel
William Stormer
Dustin Wilson

Representing NASDME:

Sarah Moore
Kathleen Plato
Gerald Richardson

Representing TACs:

JoAnn Canales
Susan Duron
Margaret Hoppe
Jerry Jenkins
Richard Naccarato

I would like to thank these fellow committee members for their contributions.

I would also like to thank Dennis Deck, Gary Estes and Bill Savard of the Northwest Regional Educational Laboratory and Rocky Maynes of the Arizona State Department of Education for their suggestions.

TABLE OF CONTENTS

TABLE OF CONTENTS
(continued)

I. INTRODUCTION

Compensatory education programs for migrant children are intended to help them overcome disadvantages they face in attaining an education. Conditions related to migrancy create educational barriers not facing the typical student. High mobility, the most obvious and common condition, interferes with maintaining instructional continuity. Problems stemming from high mobility often are combined with limited English language skills which further complicate meeting the needs of these students.

Chapter 1 of the Educational Consolidation and Improvement Act (ECIA) of 1981 and the 1983 Technical Amendments authorize the U.S. Department of Education (ED) to allocate funds for states to operate programs for students eligible for migrant services. Previously, states operated migrant programs under the authority of Title I of the Elementary and Secondary Education Act of 1965. In recent years, approximately 8 percent of the over 3 billion dollar allocation for all Chapter 1 programs has been targeted for services to migrant students. It has been estimated that the Chapter 1 migrant education program involves over 600,000 children a year (Plato, 1984) with about 60 percent of these children directly benefiting from instructional or support services funded by Chapter 1 (c.f. Naccarato, 1986).

The ED Office of General Counsel has concluded that the evaluation requirements in the 1983 Technical Amendments to ECIA apply to all Chapter 1 programs, including state operated Chapter 1 migrant programs. Thus, Chapter 1 migrant programs have the same legal requirement to evaluate as Chapter 1 regular programs operated by local educational agencies.

The following passages, taken directly from the legislation authorizing Chapter 1 programs, specify the broad requirements for evaluation.

Applicant agencies are to assure that their programs and project :

> Will be evaluated in terms of their effectiveness in achieving the goals set for them, and that such evaluations shall include objective measurements of educational achievement in basic skills and a determination of whether improved performance is sustained over a period of more than one year, and that the results of such evaluaton will be considered by such agency in the improvement of the programs and projects assisted under this Chapter....(ECIA Chapter 1, Section 556.b.4, (as amended in 1983)

The law further specifies that:

> Each state education agency shall conduct an evaluation of the programs assisted under this chapter at least every two years and shall make public the results of that evaluation. ECIA Chapter 1, Section 555.e.1)

Satisfying the evaluation requirements for Chapter 1 regular programs has been relatively easy compared to satisfying the requirements for evaluating Chapter 1 migrant programs. One reason is that a uniform evaluation system was mandated and established for Title I regular. Most states and districts have simply continued to use the Title I Evaluation and Reporting System (TIERS) to meet the more flexible, less specific requirements to evaluate Chapter 1 regular.

While the migrant program has profited from a national computer network and information exchange -- the Migrant Student Record Transfer System (MSRTS) -- to facilitate the transfer of educational and health records among school districts, the MSRTS was not designed to support program evaluation needs. Migrant programs in general have not benefited from developmental evaluation planning. Consequently, migrant program evaluations vary between states, between projects within a state, and often between years in the same state.

A national approach to uniformly evaluating migrant programs has not been mandated, nor has the ED developed or sanctioned evaluation procedures specifically for migrant programs. Nevertheless, states are responsible for evaluating at least once every two years. These evaluations must include objective measures of student basic skills achievement, and they are to be considered in improving programs. Ed also requires evaluation information for its periodic reports to Congress. Often, the information is pulled from state reports to ED even though the data are not readily summarized.

Beyond meeting legal requirements, however, the need for evaluation information about migrant programs is longstanding and problematic for many of the same reasons that migrant students are the target of special services. Migrant mobility, for example, makes it difficult to obtain pre and posttest data that are representative of project efforts. There is clearly a need to formulate practical and technical advice on how to improve the evaluation of Chapter 1 migrant programs at the state and local level.

In the summer of 1984, ED established a migrant evaluation workgroup which included staff from the ED, the National Association of State Directors of Migrant Education (NASDME), and each of the four Regional Chapter 1 Technical Assistance Centers (TACs). The workgroup focused on two related tasks. The first was to develop a national summary of participation and achievement information drawing from annual state migrant evaluation reports to ED for the 1981-82 and 1982-83 school years (Jenkins, 1986; Naccarato, 1986). The second, which is the subject of the present paper, was to formulate advice for evaluating migrant education programs.

In carrying out its tasks, the workgroup considered five related factors as follows:

1. History of migrant program evaluation on the national level

2. Legal requirements to evaluate Chapter 1 migrant programs

3. Evaluation practices reported in the 1981-82 and 1982-83 annual evaluation reports by the states

4. Evaluation experiences of NASDME representatives and their associates

5. Experiences of TAC staff as they assisted states with Chapter 1 migrant evaluation plans in their respective regions

It quickly became apparent that programs to address the educational needs of migrant students are more diverse than those for regular Chapter 1 students. Older migrant students are likely to need services to help them gain course credit necessary to graduate from high school. Some students, certainly a larger proportion than those in Chapter 1 regular programs, require instruction to develop their English language skills. Many migrant students require significant medical, dental and social services in addition to instructional services which are the principal services for Chapter 1 regular. Finally, Chapter 1 migrant programs must devote substantial resources to recruitment in order to actively encourage these students to stay in school.

Programs to meet the various needs of migrant youth are as diverse as the conditions that constrain program design from state to state and site to site. Programmatic diversity is reflected in a greater variety of goals for migrant programs and in the systems for delivering program services. Ideally, a state plan for evaluating Chapter 1 migrant programs will reflect the diversity of programs in that state by giving comprehensive and balanced coverage of the programs offered. Comprehensive coverage implies that all significant aspects of the programs in the state would be evaluated. Balanced coverage means that the evaluation effort would be in proportion to the resources devoted to each service area (e.g. Reading, English for those with

limited English backgrounds), term (i.e. regular term or summer term), and grade level.

Evaluation of some program areas will be more technically feasible and economical than others. Each state should carefully consider the ideals of comprehensiveness and balance against the practical constraints of feasibility and cost. This document will give priority to evaluating instructional programs whose objectives promote achievement growth in the basic skills, as all states are required to evaluate these programs. Even restricting ourselves to programs that focus on basic skills, program design constraints vary from those in states which serve as "homebase" for migrant families and include over 150,000 eligible migrant students to those which serve migrant populations of a few hundred students only in the summer.

By limiting the present advice to evaluation which involves measuring achievement in the basic skills, we are not implying that these other program activities can not be evaluated nor that they are not important enough to evaluate. Rather, we acknowledge the complexity of migrant programs and believe that comprehensive evaluation ideally calls for many approaches. Evaluating achievement gains is only one facet of a comprehensive evaluation of migrant education programs.

The general approach taken in this paper resulted from careful analysis of the problem. We recommend that each state develop a Chapter 1 migrant evaluation plan that will give representative coverage of the state's program. The first step in developing the state plan is to build a profile of instructional services offered to students. Based on this profile, the state would set priorities for evaluation and then select or develop evaluation strategies accordingly. To assist the states, this paper outlines the elements of a state Chapter 1 migrant education program profile and presents an overview of four approaches to evaluating migrant programs. These

evaluation approaches have been used by states in the past, and their inclusion reflects our underlying philosophy of building upon and strengthening existing practices. This involves disseminating information describing these practices, including examining their strengths and limitations and recommending ways they can be improved or made more useful. We advocate sharing and analyzing existing practice coupled with technical assistance from TACs and other qualified evaluation consultants as the best way to improve local and state evaluation practice. The variety of circumstances in states with Chapter 1 migrant programs precludes developing a single system of evaluation to meet all evaluation needs.

The following section of this report presents a brief overview of findings from the workgroup's review of the annual state migrant evaluation reports from the 1981-82 and 1982-83 school years. The third section presents a planning strategy that recommends each state develop an evaluation plan aligned with the types of programs and services in the state. The fourth section presents four general evaluation approaches used by states. The fifth section summarizes general recommendations for state evaluation of Chapter 1 migrant program evaluation.

This report is intended to be used by staff responsible for planning migrant evaluations at the state level. It is not a detailed implementation guide because, in the absence of specific federal requirements, such guidance must follow from state level policy decisions about the direction of migrant evaluation at the local level. For the same reason this paper, while it may be of interest to migrant program staff at the local level, is not a guide for local program evaluation.

II. EXISTING PRACTICE

Classification of Evaluation Approaches Used

The workgroup gathered and reviewed the 1981-82 and 1982-8 annual m grant evaluation reports submitted to ED by all the states. NASDM had pr. iously sponsored an analysis of th 1980-81 annual reports; this an ysi is presented in Plato (1985). Initially, the workgroup clas ed each approach to evaluating student achievement using a system ade Plato's (1985) national profile of migrant program participation and achievement testing practices.

The four types of evaluation identified by Plato were:

1. Norm-referenced eval tion model from t e I Evaluation Reporting System (TIERS)
2. Pre-post matched scores using scaled scores from a standardized test
3. Criterion-referenced testing
4. State assessment programs

A fifth approach to evaluation was found in th 1981-82 and 1982-83 reports. This approach, which was called "point-in-time" assessment by its developers, is similar to the state assessment approach to migrant program evaluation.

On further review of the evaluation results presented in the state reports, the workgroup modified the initial classification scheme somewhat. The separate classification for criterion referenced testing approaches was eliminated and a category for the posttest-only design was added. The major reason for this change was to acknowledge that criterion referenced tests could be used in any of the other evaluation approaches identified. A feature

common to states originally classified as using criterion referenced testing was that they all tested students after they participated in the program, but not before. Thus, the decision was to identify the evaluation approach as the posttest-only design.

The second change in the classification system was to accomodate the new point-in-time assessment approach. To accomplish this, the category of state assessment programs was broadened to include assessment programs in general.

Thus, the classification scheme in this analysis is:

1. Norm-referenced evaluation model
2. Pre-post matched scores design
3. Posttest-only design
4. Assessment programs

The workgroup extracted these four general evaluation options from existing practice, rather than pulling them from a catalog of designs such as those described by Campbell and Stanley (1966) or developing an entirely new set of "models" for migrant evaluation.

Before proceeding to discuss these evaluation approaches, it should be pointed out that some evaluation approaches that states have used detect program effectiveness, while others take a census of migrant student achievement without attributing that achievement to participation in the migrant program.

Measures of program effectiveness attempt to isolate results of participation in a program. By analogy, the effectiveness of a weight loss program might be evaluated by determining the average number of pounds a group of participants lose over a six week period. If a large number of people in a weight loss program were to lose an average of 10 pounds in the first six

weeks, the implication is that a person who is similar to those in the program can expect to lose 10 pounds in the first six weeks and that the program is effective to this degree.

A more rigorous evaluation would compare the weight loss of a randomly selected experimental group and control group. The experimental group would participate in the trial weight loss program and the control group would maintain their regular routine of exercise and diet during the same period. Roughly speaking, the difference in the average weight loss of the two groups at the end of the six week period is a measure of program effectiveness.

In contrast, some approaches to evaluation which result in an accurate measure of group status at some point in time, do not readily reveal whether changes have taken place or what may have caused changes. For example, determination of the weight on January 1, 1967 of a random sample of all U.S. males who are between the ages of 40 and 45 would not tell anything about the effectiveness of a national advertising campaign to promote weight loss among 40 to 45 year old males. Another sample of 40 to 45 year old males could be drawn two years later and their weight measured on January 1, 1989. Even if the average weight of those in the 1989 sample were less than that of those in the 1987 sample, we cannot conclude that the weight loss program has been effective at reducing the weight of 40 to 45 year old males. Other factors could account for changes in status. For example, a societal trend having nothing to do with the national program could explain the lower weight of the second sample.

Similarly, an evaluation that takes a census of the population at some point in time, although it uses objective measures of achievement, does not necessarily inform us about the effectiveness of a program in raising the achievement of a group. The information may be useful for other purposes, however.

These four data collection approaches are not alternative means to the same end. The first three are quasi-experimental designs intended to produce an assessment of a program's treatment effect. The fourth approach includes cases where a state assessment program is used to collect data from migrant students. The point-in-time assessment is similar to state assessment except that it is an assessment program specific to migrant students.

The norm-referenced evaluation model, also known as Model A, estimates the amount of achievement gain that a group of students experiences over what would be expected as a result of regular schooling alone. Use of the norm-referenced model's normal growth expectation excludes growth due to migrant student participation in a locally operated Chapter 1 program, which confounds the effectiveness of the migrant program with that of the regular Chapter 1 program.

The pre-post matched scores approach implemented without a comparison group, measures the amount of gain a group of students experiences between a pretest and a posttest, but does not isolate the cause of the gain. In other words, the effects of the regular school program, the Chapter 1 regular program (when there is one) and the Chapter 1 migrant program all contribute to the gain measured by the pre-post matched scores approach. The result is a measure of the migrant student achievement growth resulting from their total educational experience. State migrant program evaluations using the pre-post matched scores approach have employed both normed and non-normed tests.

The posttest-only design has been used primarily with criterion referenced tests. Its key feature is that the test scores express results in such terms as the percentage of students achieving mastery of a particular skill after they have completed the program. Because there is no built-in control for pre-program status, it does not measure program effectiveness.

A few states have taken advantage of their existing state assessment programs to gather and report achievement data on migrant students. State assessment programs take a number of forms depending on:

1. Grade level(s) assessed
2. Use of sampling or testing the entire population of students
3. Subject matter covered
4. Type of assessment instrument used
5. Time of year testing is accomplished
6. Frequency of testing (e.g., every year, every three years)

A state assessment is designed to describe the achievement level of all or a representative sample of all students in a defined class (e.g., fourth grade students). A state assessment approach, when applied to evaluating the Chapter 1 migrant program in a state, produces a census of the achievement level of the migrant students. When the assessment is repeated over a number of years, the trend of migrant student performances may be displayed.

A recent variation of the state assessment approach is, point-in-time, which uses a test, or test battery, and a testing schedule established specifically for the migrant student population in the state. While it also produces descriptive performance information that does not directly measure program effectiveness, it more readily allows for a test that can be matched to the curriculum of the migrant program and a testing schedule that can be set taking into account the migrant student attendance cycle.

Before moving on to the fourth section of this paper, which analyzes each of these approaches to evaluating Chapter 1 migrant programs, the following observations are presented based on the evaluation results presented in the annual reports for 1981-82 and 1982-83.

Observations About 1981-82 and 1982-83 Evaluation Methods

Roughly 50 percent of the states reporting indicated that they used some form of evaluation utilizing achievement data (c.f. Jenkins, 1986). Not all of these states, however, actually presented achievement data in their annual report. Data were withheld for a number of reasons including:

1. Problems implementing the evaluation approach
2. Poor data quality
3. Reporting requirements from ED did not call for submitting the data
4. Evaluation focused on local program management and improvement rather than generating state-level aggregates

Not all states used the same evaluation approach(es) from year to year. Why this happens is not clear, but lack of continuity contributes to data quality problems and makes implementation confusing to local staff. A long term pattern of inconsistency suggests the need for a state plan and technical follow-through to guide evaluation of the migrant program.

Thirteen states reported evaluating their migrant programs using the norm-referenced evaluation model in 1981-82. Nine of these thirteen states reported evaluation results based exclusively on the norm-referenced model. The remaining four states using the norm-referenced model reported evaluation results based on at least one additional approach. Ten of these thirteen states produced data that could be aggregated with that from other states for 1981-82.

Ten states reported results from the norm-rereferenced model in 1982-83. Seven of the ten reported results were based exclusively on the norm-referenced model.

Only five of the the states producing data that could be aggregated and reported for the 1981-82 program year also produced aggregatable data in 1982-83. States reporting in 1981-82, but not in 1982-83, cited data quality problems as the primary reason. It is not clear whether these problems reflect actual reductions in data quality or less tolerance for poor quality data. But the fact remains that much less data were available in 1982-83 than in 1981-82.

Only some of the evaluations allow the reader to relate the evaluation results to instructional services offered through the migrant program. It would be useful to know, for instance, what percentage of the students receiving instructional services in a specific subject area (e.g., reading) are tested in that same subject area. State assessment approaches, in particular, do not tie the evaluation to the instruction, as students may even be tested in areas where no supplementary instruction was provided.

Some states reported that their evaluation focused on individual project improvement and was not designed to produce state level summaries of migrant student achievement. It remains to be seen how many of these states will voluntarily adopt evaluation approaches that result in meaningful state level achievement summaries. Consultations by the TACs suggests that many of these states are planning to produce interepretable state level evaluation results. The earliest such results (for the 1984-85 school year) will be available is early 1986. Some will not be available until late 1986 or early 1987.

There is reason to speculate that state evaluation approaches are related to characteristics of the migrant student population served in that state. For example, states with a high percentage (and number) of former migrant students may find evaluation models that require testing the same students before and after participation on the program are feasible and, hence, may choose to use the norm-referenced model. States with a high percentage of

interstate migrant students will find that such evaluation methods as the norm-referenced model are not feasible in their state and may use an assessment approach or a posttest-only design. Unfortunately, very little data exist on the number of migrant students actually participating in educational programs (Naccarato, 1986). Florida, which has used the norm-referenced evaluation model for some time, served a population of about 60 percent interstate migrants, 25 percent former migrants, and 15 percent intrastate migrants. Their evaluation approach is consistent with the hypothesis that states with a large number of former and intrastate migrant students would be able to use the norm-referenced evaluation model.

Kentucky, another norm-referenced evaluation model user, had about 70 percent former migrants and 19 percent intrastate migrants. Over 50 percent of the migrant students in Georgia were former migrants but there was a higher percentage of intrastate migrants (27 percent) than in Florida or Kentucky. Georgia reported its achievement test results based on the norm-referenced evaluation model for 1981-82 but not for 1982-83 when data quality problems and small sample sizes were cited as reasons for not reporting data.

The sparce data set on 1981-82 and 1982-83 participation revealed that the percentage of intrastate migrant students reported was fairly low. Some very large states were missing data, but in no case was the percentage of intrastate migrants greater than about 30 percent and in most cases it was much less. North Dakota, which operates a summer program but not a regular term program, reports that about 94 percent of its migrant students are active interstate migrants and less than 1 percent are in the former migrant category.

One of the most frequently mentioned barriers to evaluating the effectiveness of the migrant program is the high mobility rate of the students. The data that the states reported in 1981-82 and 1982-83 suggest that mobility between states is more pronounced than that within states. For

states that are the exception to this rule, the problems of tracking students may not be as much of a barrier. We must, however, be somewhat skeptical about the accuracy of the participation data for the 1981-82 and 1982-83 project years.

In the next section of this paper, a process is outlined for developing a profile of the Chapter 1 migrant program in a state. The profile is to help set state priorities for evaluating the Chapter 1 migrant program.

III. MATCHING EVALUATION TO THE STATE PROGRAM

Steps in Developing a State Plan

A state should develop a comprehensive and balanced plan for evaluating its Chapter 1 migrant instructional programs. Such a plan allows one to document the relationship between evaluation information and instructional emphasis. For example, a state that reports evaluations based on achievement test results in reading or mathematics can objectively demonstrate how those results reflect the efforts of their Chapter 1 migrant program. A state can develop and monitor a state Chapter 1 migrant evaluation plan by following these seven steps.

Step 1. Develop a state profile of students served

Step 2. Establish evaluation priorities

Step 3. Select evaluation options in relation to priorities

Step 4. Develop a long-range plan for areas to evaluate

Step 5. Implement the plan

Step 6. Periodically make the results of the evaluations public

Step 7. Periodically review the utility of the evaluation results and revise the procedures as needed.

The present paper focuses on the first three of these steps, developing a state profile of the Chapter 1 migrant program, establishing evaluation priorities, and selecting evaluation options in relation to priorities. Although emphasis is given to long range planning, we do not recommend that

states drop current evaluation approaches until they systematically examine their priorities. Rather, a state should renew its ongoing evaluation activities to better reflect priority program areas and shape future efforts in relation to these priorities.

State Chapter 1 Migrant Instructional Program Profile

As a first step in setting priorities for a state Chapter 1 migrant evaluation plan, available data should be used to develop a state Chapter 1 migrant program profile. The profile will contain descriptive information on the migrant programs in the state. Each profile, which will be unique to a state, will be used to make an objective estimate of the relative level of effort given to different programmatic areas as defined in terms of the following:

I. Program Characteristics
   - A. Term of Instruction
   - B. Grade Level
   - C. Subject Area

II. Student Characteristics
   - A. Migrant Status
   - B. English Language Proficiency

A state can derive estimates of effort from data on the number of students receiving instructional services. Such data should be generally available starting with the 1984-85 school year and, for many states, are available for

earlier years. Given the state profile, each state will be prepared to establish an evaluation plan that will be in concert with its programs. For instance, if a large proportion of students served receive reading instruction during the regular term, those planning the migrant evaluation would place high priority on evaluating the regular term reading program.

The evaluation methods most appropriate would depend on other characteristics of the state's migrant programs. Evaluation methods of choice, for instance, will also depend on information included in a state profile, such as grade level, language proficiency, and mobility of the students served.

More specific guidance on profiling a state Chapter 1 migrant program follows. Sample data displays show how programs may differ from state to state.

Instructional Program Characteristics

Term of Instruction States and local agencies offer migrant programs in either the regular term or the summer term. Regular term programs may run as long as the full regular school term, which generally lasts about 36 weeks. Summer programs usually run 8 to 14 weeks. However, they may be more intensive (i.e. more hours per day) than regular term programs because they do not conflict with regular school term classes.

In recent years, nearly a dozen states have operated regular term programs exclusively (Plato, 1984; Jenkins, 1986). As many as seven states have operated summer programs exclusively. The remaining states have operated a mix of regular term programs and summer programs. This will have an important bearing on choosing the most appropriate mix of evaluation methods to provide

a comprehensive and balanced evaluation of the state's Chapter 1 migrant education program.

A state should include information on the number of students served in the regular term and the summer term as part of its state migrant education program profile. This information will generally be available by grade level and by instructional service area, allowing for meaningful displays of data.

Evaluation strategies appropriate for summer programs are more limited than those for regular term programs. Other things being comparable (e.g. subject areas taught, grade levels and ages of students in the program), using a norm referenced test with a pre-post matched scores design is an unsound approach for evaluating summer programs. Norm referenced tests, as global measures of student achievement, are unlikely to be sensitive to instruction offered in a short term program whether it is in the summer or the regular term.

Grade Level Migrant program services by grade level range from pre-K to grade 12. Beginning with the 1984-85 program reporting year, all states will have data on the number of students served by grade level, which will be reported for both the regular term and the summer term. The major evaluation implication of the distribution of students served by grade level is that methods using standardized, norm referenced tests are more stable and reliable with students who are at least at the second grade level. Below the second grade level, the quality of these test data are suspect, and evaluation approaches that do not rely on these tests may be necessary.

Plato (1984) has reported that, nationally, about 26 percent of the students enrolled on MSRTS during the 1980-81 school year were at or below the first grade, 35 percent were at or below the second grade, and 44 percent were at or below the third grade. More recent data on participation suggests that

in some states over 50 percent of the students receiving services during the regular term are at or below the third grade and that as many as 70 percent of those in summer programs are at or below the third grade (Naccarato, 1986). These participation data are not based on a representative sample of states with migrant programs, so one should not conclude that the national rate of participation is higher relative to enrollment at the earlier grade levels. However, the data does show that for some states the participation rate at lower grade levels is higher relative to enrollment than at the upper grade levels. This supports the need for individual state profiles.

Because of variability between states, each state should include a distribution of the number of students served by grade level for both the regular term and the summer term program in its state profile. Having this information available will be important to a state as it determines its priority areas for evaluation. It also will reveal the extent of need for special evaluation strategies for early childhood programs.

Models for evaluating early childhood programs are discussed in the Handbook for Measurement and Evaluation in Early Childhood Education (Goodwin and Driscoll, 1980) and in a series of publications developed by the Huron Institute specifically for Title I and Chapter 1 programs (Haney, 1978; Haney, 1980; Kennedy, 1980; and Yurchak, 1980).

Figure 1A and Figure 1B depict the grade level distribution for students in the regular term and summer term programs for two hypothetical states (State A and State B). The evaluation needs for State A, represented in Figure 1A, are more complex than those for State B, represented in Figure 1B. State A operates programs for a substantial number of students in both the regular term and the summer term. A comprehensive and balanced evaluation

scheme should cover programs in both terms. State B, on the other hand, operates only a regular term program and, therefore, does not face as great a need for multiple evaluation strategies as State A.

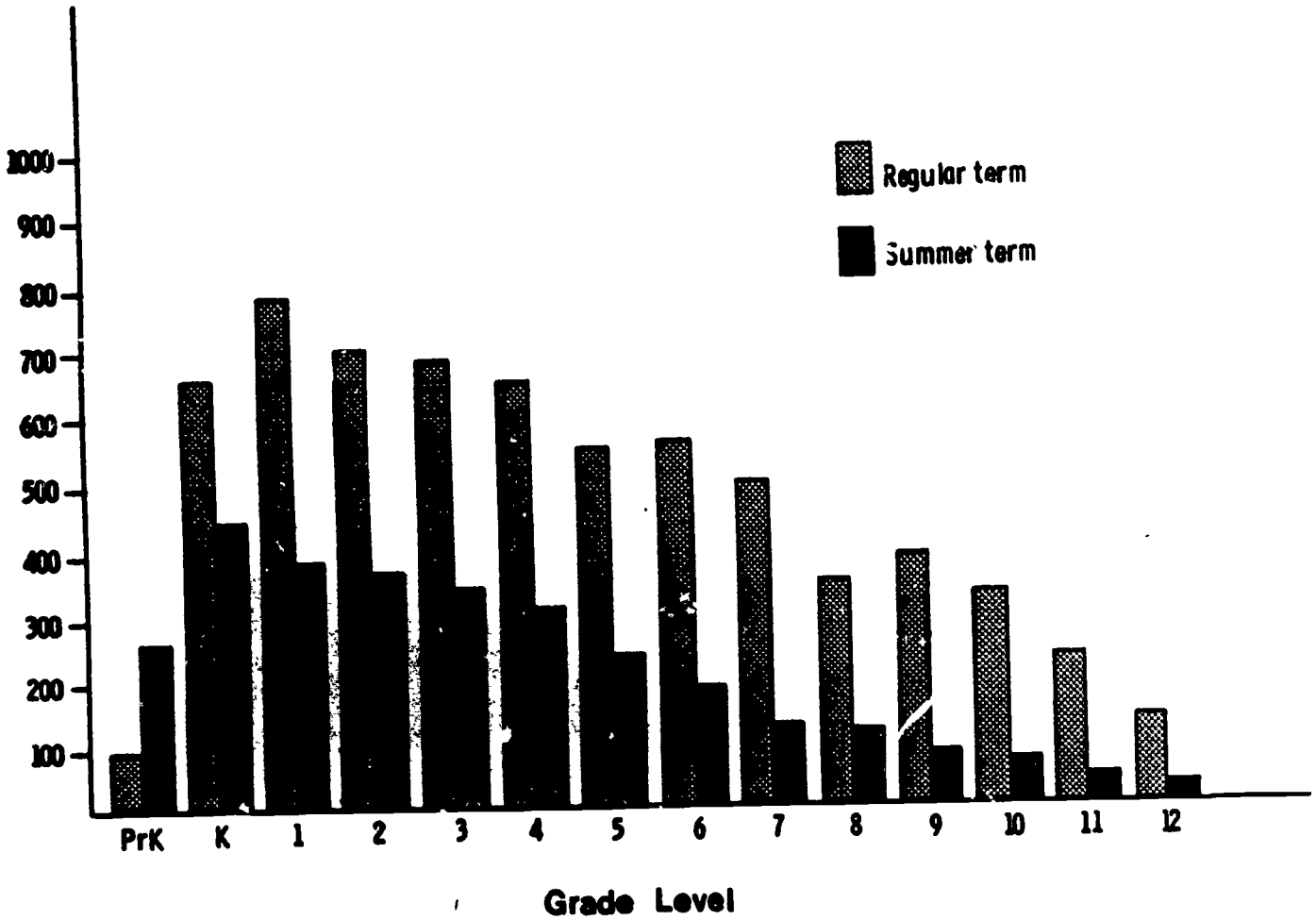


Figure 1A

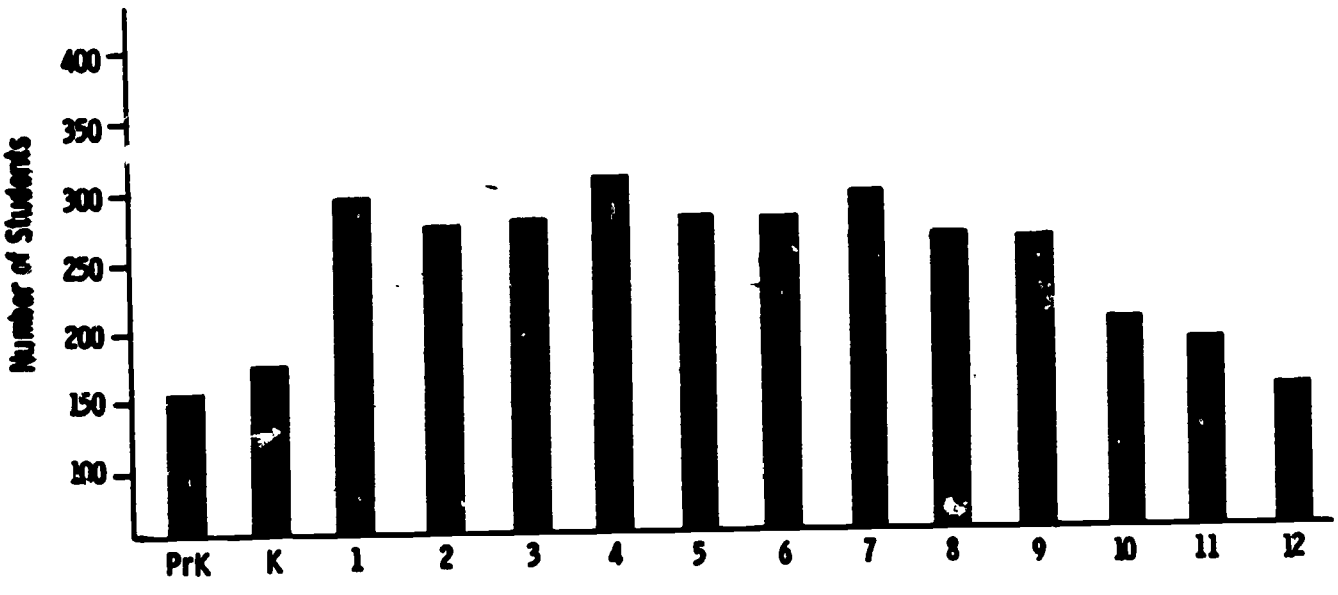


Figure 1B

Figures 1A and 1B also reveal differences in the profile shapes between the regular term programs offered by both states. State A serves a higher proportion of students at the lower grade levels than State B, with the exception of the greater proportion of children in programs at the pre-K level in State B. State B serves a greater proportion of students at higher grade levels. State A, therefore, has a greater need for evaluation methods appropriate for kindergarten and the first grade than does State B, but State B needs to consider evaluation of the programs serving the pre-K students.

Instructional Area Migrant programs may operate in a number of different instructional areas including:

1. English to those with Limited English Background
2. Reading
3. Language Arts
4. Mathematics
5. Vocational/Career
6. Other

Programs in any of these instructional areas may serve students at any grade level during the regular or summer term, although vocational/career programs are primarily at the upper grade levels.

In reviewing its evaluation plan, each state should examine a distribution of the number of students served in each instructional area in its state profile. A separate display for summer and regular term programs will highlight differences in subject matter emphasis by term. Figure 2A displays instructional service information for a state with regular term and summer term programs. Programs in the regular term serve the largest number of students in reading, while programs in the summer term serve the largest number of students in mathematics. A substantial number of students, however, is served in each of four instructional areas.

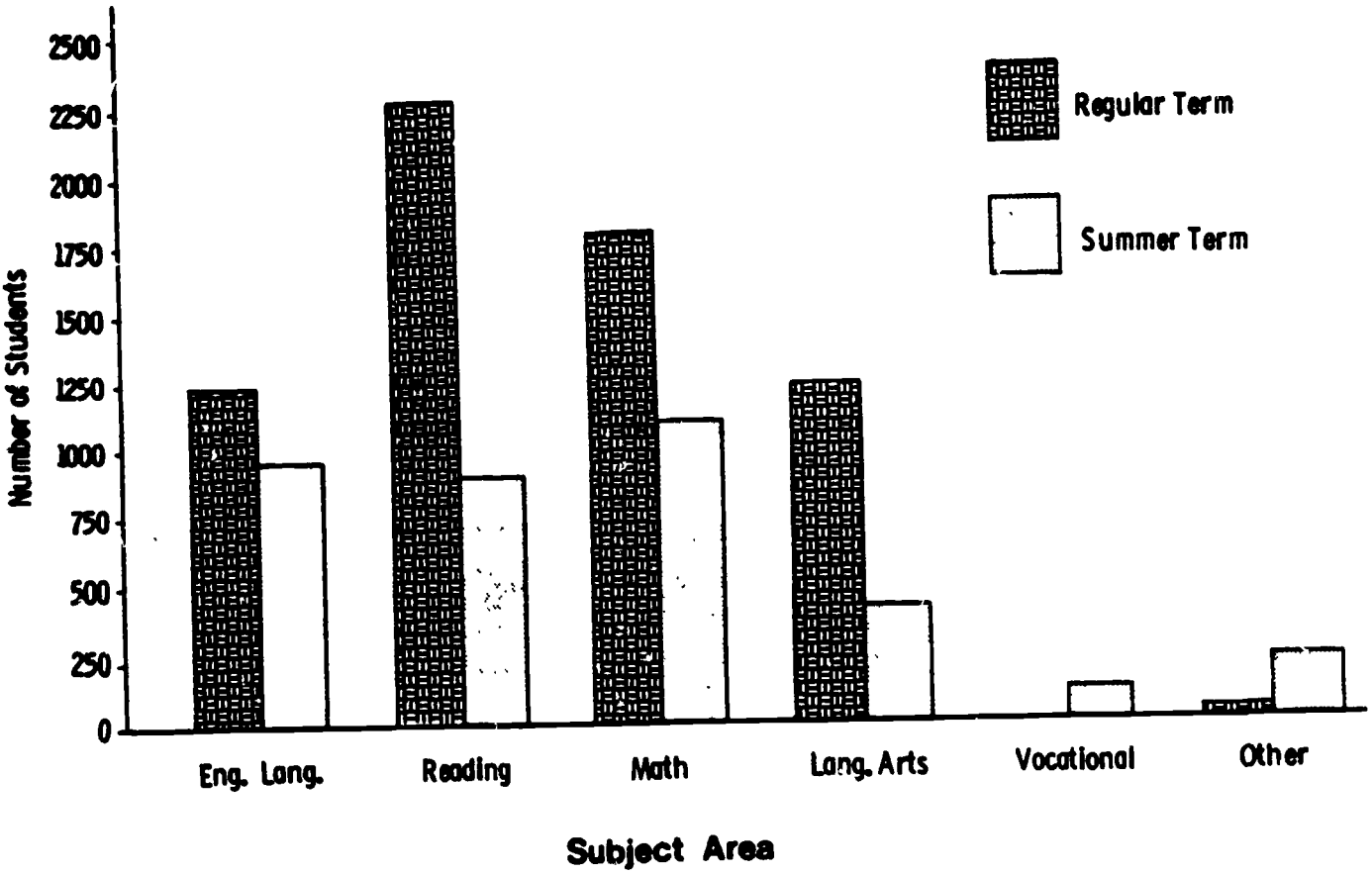


Figure 2A

It appears from Figure 2A that programs serve about twice as many students in the regular term as in the summer term. Examining Figure 2B, however, reveals that nearly all students participating in the summer term receive services in mathematics (92 percent), 79 percent of them receive services in language development and 74 percent receive services in reading.

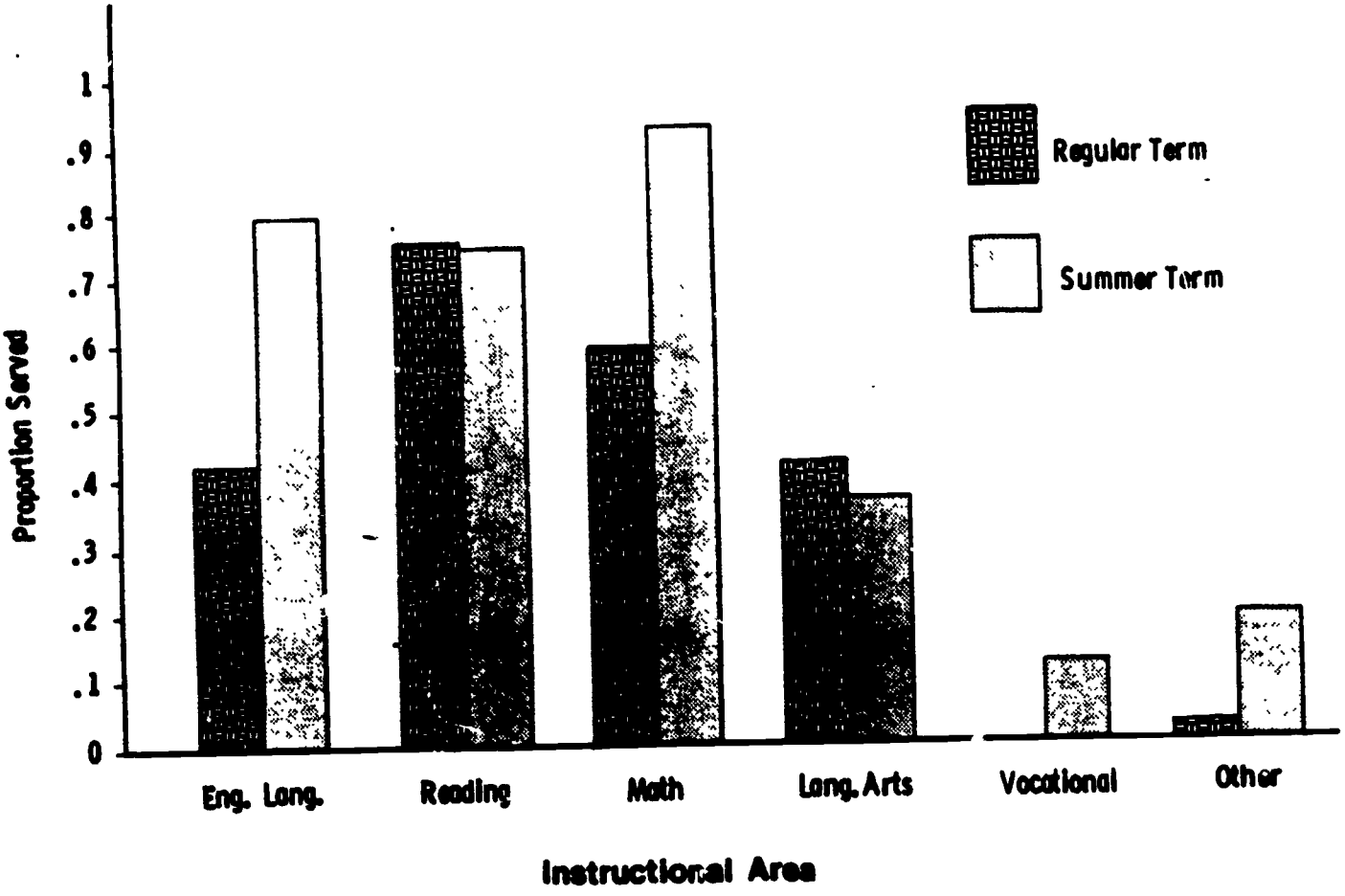


Figure 2B

In this example, the state served a total of 3,000 students in the regular term and 1,200 in the summer term. An important difference in these two terms is that more students received services in more than one instructional area in the summer than in the regular term. This should serve as a reminder that, unlike the case for grade level of students served, the categories for instructional areas are not mutually exclusive. That is, a student may receive services in more than one instructional area in a year. Thus, a total of the number of students served across the six instructional areas yields a "duplicated count" of students served. The same situation holds for a total

count across instructional terms. Some students may be in both a regular term and a summer term program. Thus, the categories are not mutually exclusive.

An analysis of the instructional service information will reveal how differentiated the state services are. Services that are not very differentiated would show almost all students receiving services across all subject areas. A more differentiated pattern of services would show students receiving services in different instructional areas. An implication for evaluation is that states with more differentiated services would need to more carefully align testing with the instructional focus. States with less differentiated service would need to guard against over testing students.

When the instructional area profile is compared to the evaluation approaches discussed later, the state can appreciate more readily where different approaches and types of instruments are more or less appropriate for evaluation. For example, states with a heavy emphasis on English language instruction for those with a limited English language background will note that:

1. Few norm-referenced tests exist for English language competency
2. Available norms may not be relevant for a local population
3. Many of the commonly used tests are designed for classifying and diagnosing students, rather than evaluating programs.

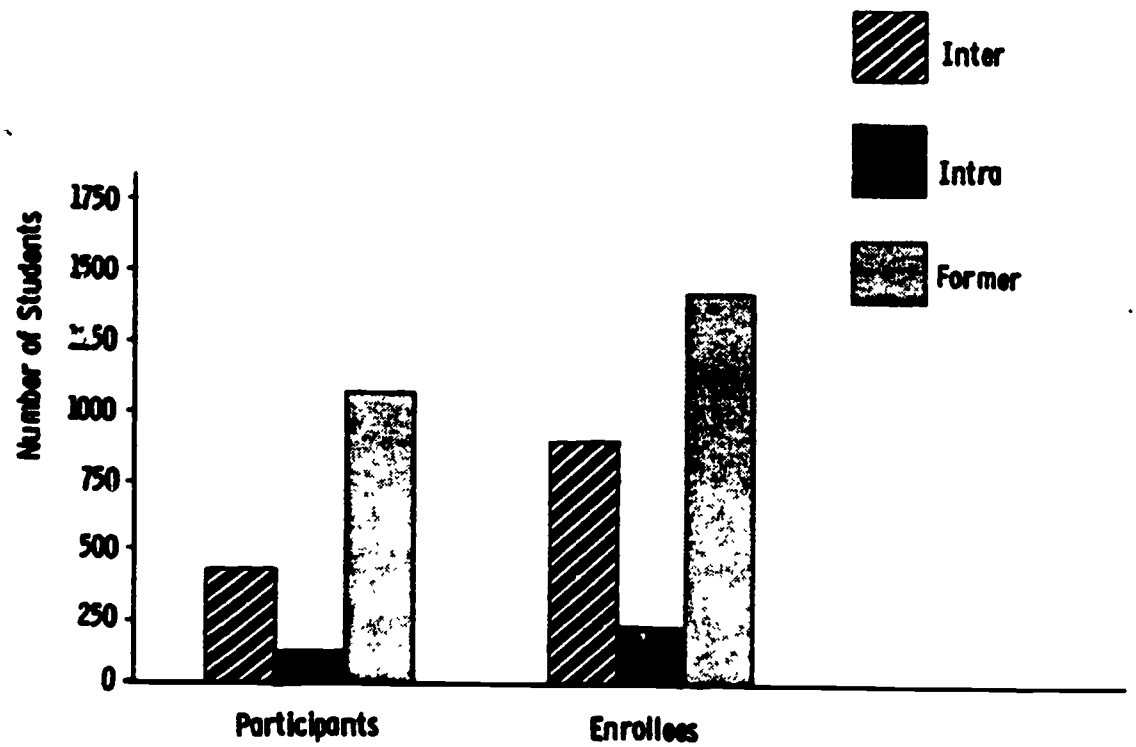
For strategies that may be useful in evaluation of these programs the reader is also referred to a status report on recommendations for evaluating bilingual education programs (Tallmadge, Lam, and Camarena, 1985) and Section 4 of this paper.

Student Characteristics

Migrant Status Rigorous, experimental evaluation requires control over:

1. Who receives program services
2. How long they receive services
3. Who is tested
4. How often they are tested
5. When they are tested

Because migrant students are expected to be more mobile than students from most segments of our society, they are not easily subject to controls for the sake of evaluation.



Migrant Status

Figure 3

Practically, evaluation approaches should take into account the rate of expected turnover. An analysis of students in terms of their migrant status (interstate, intrastate and former migrant) may be used to roughly estimate the mobility rate in a state. Ideally migrant status would be displayed for all students who are enrolled on MSRTS and the subset who receive instructional services (i.e. participants). Figure 3 (shown on the previous page) displays the number of participating and enrolled students in each of the three major categories of migrant status.

Given the emphasis on evaluating instructional programs in terms of basic skills achievement, the display of participants in Figure 3 is more important than the display of enrolled students. Figure 3 indicates that the majority of the students receiving services in either the regular term or the summer term are former migrants.

There is no certainty that families of students classified as former migrants will not migrate again. There is, nonetheless, reason to expect that a state serving a large number of former migrant students will be able to pre and posttest many of those students. When both the number and percentage of former migrant students is high, the state may place high priority on evaluating programs with designs that call for pre and posttesting. A state serving a high percentage and number of interstate migrant students in short term programs, on the other hand, may need to adopt methods of evaluation that do not require pre and posttesting the same students.

English Language Proficiency The second student characteristic of significance in considering the evaluation options available to a state is English language proficiency. English proficiency relates to two important issues in evaluating migrant programs. The first issue deals with the

invalidity of standardized achievement tests in English for students not proficient in English. A test cannot validly measure a student's competence in an area (e.g. reading) if the test is given in a language that the student does not readily comprehend.

The second issue is that the norms for standardized norm referenced achievement tests are not appropriate for students from a population that was not represented in the norming sample. This second issue is more subtle than the first. It will affect test interpretation for students who have sufficient English language skills to take a test, but who, nonetheless, are not represented in the norms for that test.

It is a good idea to include information on the language proficiency of students served by the migrant program in the state profile. Figures 4A, 4B, and 4C display pertinent information on language status.

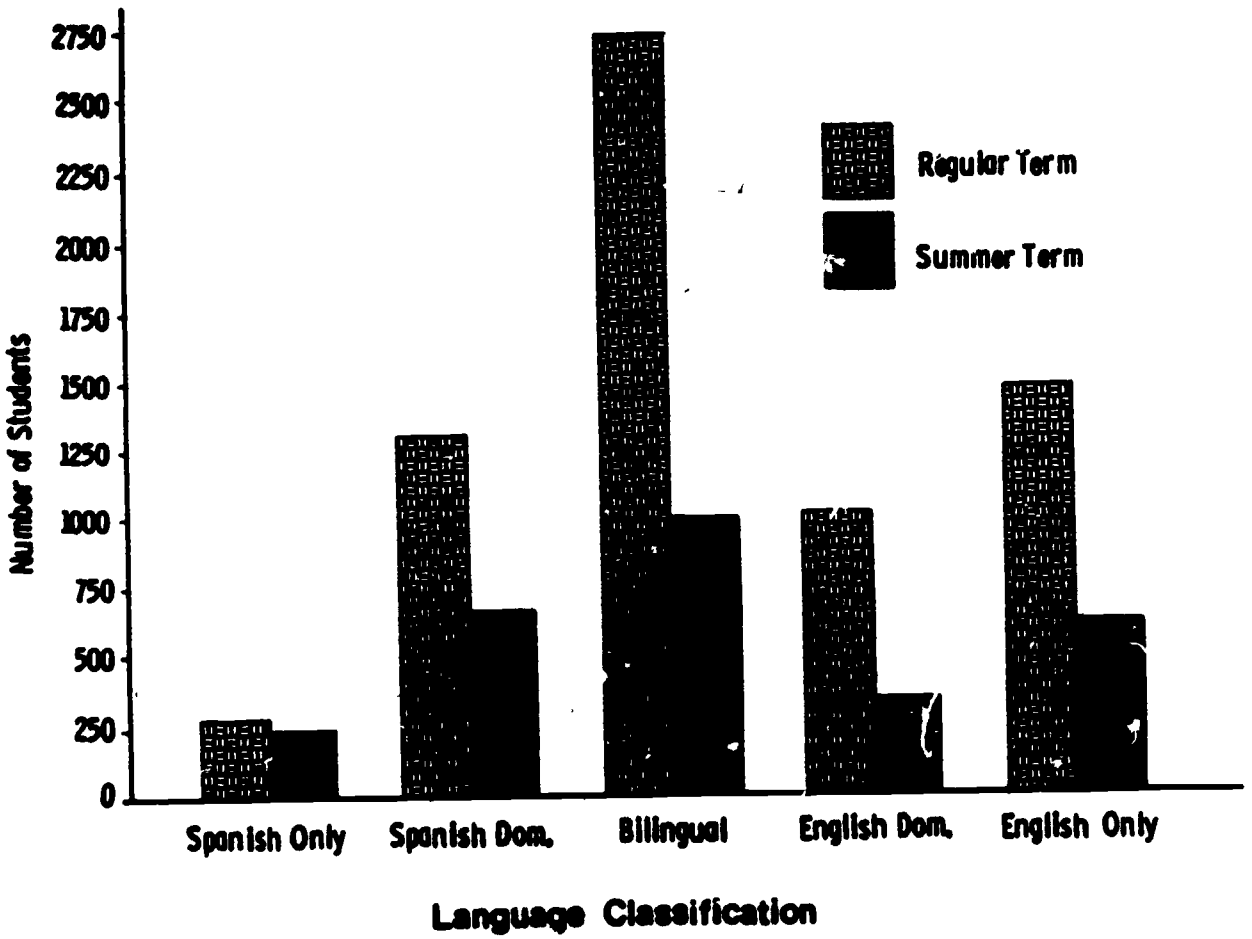
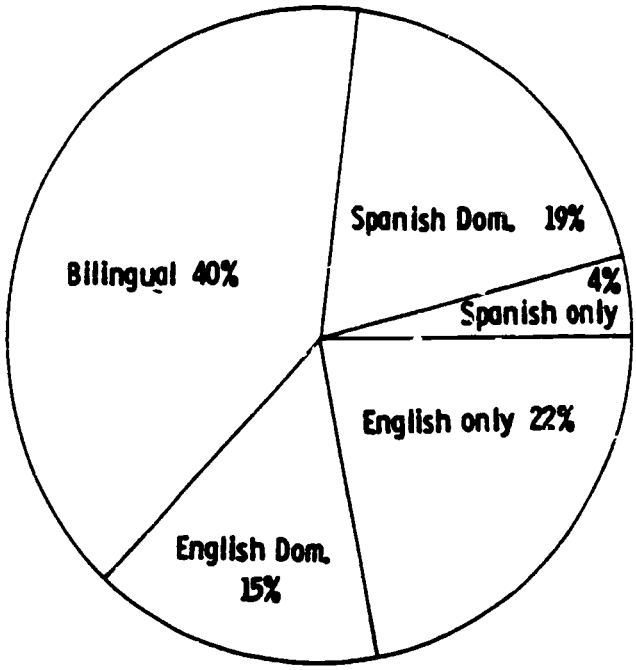
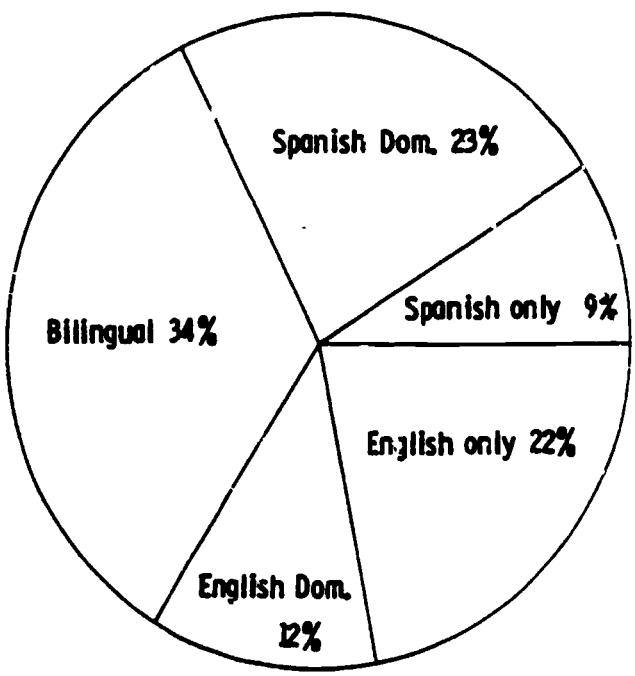


Figure 4A



Regular Term

Figure 4B



Summer Term

Figure 4C

These figures use the five Lau categories as a surrogate for language proficiency to classify students. Figure 4A, the most complete form of display, includes the number of participating students in each of the five Lau categories for the regular term and the summer term. Figure 4B, a pie chart, depicts the percentage of students in each of the five categories for the regular term and Figure 4C depicts the percentage breakdown for the summer term. In these examples about 70 percent of the students in the state are either bilingual, English dominant, or English only, indicating that a large percentage of students may be able to validly respond to achievement tests in English. Another conclusion may be drawn from the profile in another state.

Five Criteria for a State Plan

The previous section of this paper gave a number of sample data displays a state may include in its Chapter 1 migrant education program profile. Nearly all of these data should be readily available for all states reporting in the format specified by the ED. Some states may be able to prepare additional data displays, such as a breakdown of the number of students served in each instructional area by grade level. Such information would be even more helpful in setting evaluation priorities, but would require a more complex data base.

These data have inherent value for describing the service population in a state. Profiling as we advocate it, however, is to assist the state in setting evaluation priorities and plans to best reflect the efforts of programs in the state. State priorities are appropriately determined by the state as it accounts for and supports its services to migrant students.

We have assumed that each state's program is fairly consistent from year to year so that a long range plan can be based on the contents of a current profile. It is advisable to include trend data in each profile to deal with changing patterns of service.

The profile should be used to determine the evaluation plan that will give the state comprehensive and balanced coverage of its migrant program. Comprehensiveness and balance are related to the representativeness of an evaluation, which is one of five criteria recommended for a state plan. These five criteria are:

1. Is the evaluation representative of the state's programs?
2. Is the evaluation useful for program improvement?
3. Does the evaluation result in data that is easily summarized for the state?
4. Is the evaluation methodology technically feasible?
5. Is the evaluation relatively low cost?

Designing an evaluation plan that will be representative is a state specific activity. States with similar profiles may develop similar plans. States with some similar program components may have some common evaluation activities and other evaluation activities that are unique to themselves.

Reviewing and Selecting Approaches

Section IV of this paper presents an analysis of the four approaches to evaluation identified earlier. Each analysis follows a framework of eight characteristics:

1. Purpose
2. Design Features
3. Testing Features
4. Aggregation
5. Strengths
6. Limitations
7. Conditions Supporting Use
8. Recent Uses

A study of these analyses of evaluation options within the context of the state profile should help the state select a mix of evaluation techniques for evaluating the migrant program in the state.

IV. APPROACHES TO MIGRANT EVALUATION

Norm-Referenced Evaluation Model

Purpose The norm-referenced evaluation model (also called Model A) was originally developed in the mid 1970s to estimate the effectiveness of Title I projects operated by local school districts (see Tallmadge, Gamel and Wood, 1981). One of the three models developed for evaluating Title I projects, it is still used by nearly all states and local projects for evaluating Chapter 1. The norm-referenced evaluation model is a strong candidate for Chapter 1 migrant program evaluation. It has been used by a number of states and can be especially appropriate for evaluating programs serving a large number of former migrants or operating in "home-base" states. When used appropriately, it can produce evaluation results that are easily analyzed to give state level summaries of achievement gains for students in the migrant program.

The norm-referenced evaluation model was developed for an evaluation system that builds from local project evaluations. Local building results are combined to produce district summaries which in turn are aggregated to the state level. These in turn are aggregated to produce a national summary of the effectiveness of the Chapter 1 program. For example, a district with different Chapter 1 fourth grade reading projects in two buildings will have two separate evaluations which are combined to give a district measure of effectiveness for its fourth grade reading program. The district results are then combined with those from other fourth grade reading programs across the state to produce a statewide report of Chapter 1 effectiveness for fourth grade reading programs. Finally, state results are combined to produce a

national report on the effectiveness of fo ch grade Chapter 1 reading programs. This process of combining, or aggregating, to produce national measures of effectiveness has been used to evaluate Chapter 1 regular programs since the 1979-80 school year.

Project effectiveness is defined as the amount of achievement growth that a group served by the project makes over that which would be expected from their participation in the regular school program only. The growth expected from regular school participation is called the no-treatment expectation. The norm-referenced evaluation model determines the no-treatment expectation without a local comparison group.

In more operational terms, the measure of project effectiveness generated by the norm-referenced evaluation model is as follows:

Project		Posttest		Expected Posttest
Effectiveness	=	Status	-	Status

The norm-referenced evaluation model provides considerable flexibility at the local level in that individual projects choose an achievement test to match their project objectives. However, it must be possible to derive national percentiles from the test used to evaluate. The test may itself be normed, or it may be equated to a test with national norms. Cases in which norms are derived by equating are isolated because the equating process usually adds to local evaluation costs and can introduce additional error into the evaluation.

While local projects choose pretest and posttest dates that best fit their situation, these dates must be near the dates on which the publisher normed the test. For all but a few tests this requirement limits the choice of testing cycles to fall-spring, spring-spring or fall-fall.

While it was first assumed that measures of project effectiveness would be comparable regardless of the testing cycle, experience has shown that gains from an annual cycle are generally less than those from a fall-spring cycle. Consequently, results from different testing cycles are aggregated and reported separately. Evidence suggests that gains from the fall-spring cycle are inflated and that annual gains are more realistic estimates of program effects (Linn, 1982).

The no-treatment expectation for the norm-referenced evaluation model, is based on the expectation that students do no better or worse on posttests than they would have without the additional services provided by the Chapter 1 project. Thus, they would be expected to maintain their percentile rank from the pretest to the posttest. Students who, when tested in the spring of year one score at the 25th percentile are expected to score at the 25th percentile in the spring of year two if the intervening Chapter 1 program is neither effective nor harmful. This hypothesis requires that students not be selected into the Chapter 1 program on the basis of their pretest score to control for regression to the mean. To allow scores from different tests to be combined for state and national estimates of Chapter 1 effectiveness, it was necessary to use a common metric that could be averaged across tests. The Normal Curve Equivalent (NCE) was developed as th s common metric. The NCE is an equal interval scale with a range from 1-99, a mean of 50, and a standard deviation of 21.06. While the NCE is essential to implementing the norm-referenced evaluation model, the design requirements and the use of NCEs together makes the norm-referenced evaluation model a very workable procedure for estimating the effectiveness of Chapter 1 projects and programs.

Design Features The norm-referenced evaluation model has four design features:

1. Students are pre and posttested
2. Only students with a pretest and a posttest are included in the analysis
3. Tests are administered near the empirical norm dates for the test used
4. Pretest scores are not used to select students for the program unless they are the sole basis for selection and a statistical correction is applied

Points one and two simply emphasize that the norm-referenced evaluation model is a matched score design: that is, one that includes only those students who have both test scores available for analysis. Including students who have only a pretest or posttest may invalidate the model.

Testing near the empirical norm dates is necessary to produce valid normative scores. Unless it is compensated for, testing too far from the norm date will bias percentiles and NCEs. Therefore, such scores must be interpolated or extrapolated to the date the test is given. As a rule of thumb, it is appropriate to test as early as two weeks before the norm date or as late as two weeks after and still use the norm tables that correspond to the empirical norm date. Testing beyond the two week period on either side of the norm date requires interpolated norms which are often available from the test publisher in special norm tables. In general, it is inadvisable to test more than six weeks from the empirical norm date because of the distortion that may occur as a result of the interpolation process.

Unless some control is exercised over selecting students, the norm-referenced evaluation models is subject to bias from regression to the mean. Regression to the mean is a statistical artifact in which a group

selected because of its extreme standing on an imperfect measure will tend to move closer to the mean of the total group on a second measurement.

The standard version of the norm referenced model specifies that the pretest scores not be used in selection at all. When evaluators adhere to this restriction, the mean pretest score of the Chapter 1 group can be used to estimate the no treatment expectation without statistical correction.

A variation to the standard selection procedure permits selection on the pretest if a statistical correction is applied. To apply the statistical correction, however, it is necessary to base selection exclusively on the pretest. The procedure for selecting on the pretest with the norm-referenced evaluation model is described in Appendix A.

Testing Features Two testing features help distinguish the norm-referenced evaluation model:

1. The same test series is used for the pretest and the posttest
2. The test used must have national norms or be equated to one that does

It is well known that not all achievement tests in the same subject areas measure the same thing and that some tests are more difficult than others. Tests produced in different years when there are true differences in the achievement levels of students, for example, will produce different percentile performances for subsequent groups of students. Using two different tests to pretest and posttest students will therefore produce invalid gains.

Aggregation Conditions for aggregating the results from a project application of the norm-referenced evaluation model with those of other projects are:

1. Where results are aggregated and reported across different tests, the metric of choice is NCEs
2. Results are combined within, but not across, subject areas
3. Results are aggregated within comparable testing cycles
4. Data must pass through quality control screens
   - Testing on or near the empirical norm dates
   - The procedure for estimating the no treatment expectation is consistent with the selection process
   - Test floor and ceiling effects are absent
   - Data have been screened for conversion errors, use of the correct norm tables, etc.
   - Same test series has been used for both the pretest and the posttest

Conventional procedures for aggregating data across multiple projects require NCEs because the scale is equal interval. Gains for one project, however, ought to be considered only roughly comparable to the gains for another project because of differences in tests and in the match between the test used and the objectives of each project.

In aggregating results upward (i.e. from school to district to state to nation), NCE gains from different tests are combined only within the same subject area. For example, a fourth grade reading project in district A may use the vocabulary subtest from Test X as its outcome measure, while district B may use the comprehension subtest from Test Y. The gains from both projects are aggregated, assuming they have used the same testing cycle, into the state gains for fourth grade reading. The aggregate gain is similar to such indices as the Gross National Product. However, this example emphasizes a limitation

that should be placed on comparing projects with one another.

As mentioned earlier, experience has shown that fall to spring gains are higher than annual gains. Therefore, it is inadvisable to combine fall to spring gains with annual gains, especially if there are plans to track gains across years. Differences in the proportion of annual and fall to spring gains in the data set will affect the results. Projects wishing to compare their results to some state average gain will need to refer to gains separated by testing cycle.

Before data are included in an aggregate, they should be passed through a quality control screen. Those listed above are consistent with standard screens used for gains produced with the norm-referenced evaluation model.

Strengths Among the strengths of the norm-referenced evaluation model are:

1. Local project effectiveness is measured without a local control group
2. Results can be readily combined across a number of projects to give district, state and national effectiveness
3. Well established routine for evaluating Chapter 1 projects across the country

Generating a measure of effectiveness without a local control group makes the norm-referenced evaluation model an attractive option, as it will rarely be feasible to find a local comparison group. Even if one can be found, serious ethical and legal issues exist related to withholding treatment from a qualified group.

The norm-referenced evaluation model was originally developed to facilitate aggregating results from individually conducted evaluations to produce gain estimates at the state and national level. When the projects

reporting are representative of the population of projects, the aggregate gains can be interpreted as state and national level effects of the program.

Extensive experience with the norm-referenced evaluation model is a practical advantage in that its assumptions and technical requirements are probably more widely known than would be those of a new evaluation model.

Limitations Some of the limitations of the norm-referenced evaluation model, as it may be applied to Chapter 1 migrant programs, are:

1. Not appropriate for measuring projects of short duration
2. Limited to outcomes measured by tests with national norms or equated to nationally normed tests
3. Of limited value for evaluating projects with high student turnover between the pretest and the posttest
4. Results confound the effects of Chapter 1 migrant and Chapter 1 regular programs for students participating in both
5. May be misleading if migrant students are not represented in national norms

At least three related reasons exist why the norm-referenced evaluation model is not appropriate for measuring the effects of short term projects such as summer programs. First, the norm dates for most standardized tests are in the fall and in the spring, which limits the times when a test can be given as a pre and posttest. Second, a norm referenced test is generally a global measure of achievement. Short term projects, on the other hand, are likely to have fairly specific and limited objectives that are only a small part of what is measured by a nationally norm referenced test. Only a small percentage of the items on a norm referenced test probably will relate to the objectives of a short term program.

The norm-referenced evaluation model requires a nationally norm referenced test or a test that has been equated to a nationally norm referenced test. This requirement limits the range of objectives that can be measured to those that are measured by, or equivalent to those measured by, a nationally norm referenced test. Many Chapter 1 migrant programs focus on language development, an area for which very few, if any, norm referenced tests are available.

To validly assess the effectiveness of a project, it is important that the sample of students tested in the evaluation are representative of those served by the project. A high rate of student turnover threatens the validity of the norm-referenced evaluation model when the characteristics (e.g test performance) of the students who are only present for one testing differ from those who are present for both tests.

The norm-referenced evaluation model's no treatment expectation is supposed to reflect the growth due to the regular school program only. For students who receive reading instruction through both Chapter 1 migrant and Chapter 1 regular, gain will reflect both programs rather than just the Chapter 1 migrant program. It is unclear how common it is for students to be in more than one compensatory education program. Therefore, it is not clear to what extent the confounding of effects presents a real problem or just a hypothetical one.

The norm-referenced evaluation model uses a national norm group as a surrogate comparison group to support the no-treatment expectation. Therefore, it assumes students like those in the local program are represented in the national norms for the test being used to evaluate. It is apparent, for instance, that migrant students with very limited English language skills will be omitted from norming studies. A local migrant program serving many students with such language deficiencies may find the norm-referenced model inappropriate.

Conditions Supporting Use The norm-referenced evaluation model is most useful for evaluating basic skills projects with limited student turnover between the time the pretest and the posttest are scheduled to be given. States with relatively large numbers of former migrant students or "home base" states may find that the norm-referenced evaluation model is feasible for much of its population of migrant students.

Consideration might be given to using the Norm-Referenced Model for students who move within a state between the pretest and posttest. Such students may have participated in more than one Chapter 1 migrant program, which means that the resultant gains would not be isolated to a specific program. On the other hand, these students may have missed some school or have been in a Chapter 1 program for only a short period. Using the norm-referenced model in this way would call for information on the instructional history of the students during the time between the pretest and the posttest.

Recent Use Thirteen states used the norm-referenced evaluation model to report on the effectiveness of their Chapter 1 migrant programs for the 1981-82 or the 1982-83 school years. Ten of the thirteen states using the norm-referenced evaluation model produced 1981-82 data suitable for aggregation. In 1982-83, only 5 states produced data suitable for aggregation. The 5 states reporting in 1981-82 but not in 1982-83 cited various reasons. Two states indicated that their data were not of sufficient quality to be reported.

There were no new norm-referenced evaluation model users in 1982-83 and one half of those reporting results in 1981-82 did not report them in 1982-83. There was an even greater reduction in the number of students with

data from 1981-82 to 1982-83. In 1981-82, the total number of students with pre and posttest scores in reading was 17,787; by 1982-83 the number dropped 4,171.

One of the common concerns about the norm-referenced evaluation model to evaluate migrant programs is the belief that migrant students do not achieve at the level of their grade level peers and that the model will unfairly evaluate their performance by making the programs look bad. It is true that the norm-referenced evaluation model assumes that students in the local population are represented in the publisher norms. Although migrant students may not be represented in those norms, results from states using the norm-referenced evaluation model show that migrant students gains differ little from those of students in regular Chapter 1 programs.

Pre-Post Matched Scores

Purpose A second way of evaluating migrant programs at the state level has been to test students before and after their participation in a migrant project, but without following the controls for the norm-referenced model. Either a normed or a non-normed test (e.g., an objectives-referenced test, a criterion-referenced test) is used to assess student performance. In a sense, the norm-referenced evaluation model is a special case of the pre-post matched scores design. When a nationally normed test is used, the pre-post matched score design superficially resembles the norm-referenced evaluation model. In its general form, however, the pre-post matched scores design does not necessarily include the controls needed to obtain a valid measure of program effectiveness. If it does include those controls, then it would be equivalent to the norm-referenced evaluation model where the NCE metric is used to

compute gain scores.

Probably the most significant control overlooked with the pre-post matched scores design involves insuring the appropriate match between the process of selecting students and the way in which the estimates of the treatment effect are calculated. The selection process is not specified clearly enough to know the appropriate method of analysis.

The purpose of the pre-post matched scores design is simply to assess change; whether the change is the result of the migrant program or some other factors is not taken into account. Stated differently, the pre-post matched scores design alone does not separate the effect of the program from other sources of change such as the regular school program, the regular Chapter 1 program or maturation. Consequently, it cannot produce a measure of program effectiveness without additional controls.

Design Features The pre-post matched scores design has but two key design characteristics:

1. Students are tested prior to and after their participation in the program
2. Only students with a pretest and a posttest score are included in the analysis

Because the pre-post matched scores approach is not restricted to norm referenced tests (or tests for which national norms can be derived through equating), there is no necessary restriction of testing near empirical norm dates.

Testing Features The major testing features of the pre-post matched scores approach are:

1. The same test is used for the pretest and the posttest
2. The test is administered according to the same procedures for the pretest and the posttest

Since the result of a pre-post matched scores approach is a gain score, use of the same measure and score metric for the pretest and the posttest is essential. In most cases this means using the same test on both occasions. While no cases were reported in the annual evaluation reports reviewed, the use of different, but equivalent tests for the pretest and posttest is possible.

Aggregation The requirements for aggregating gain scores resulting from the pre-post matched scores design are more restrictive than those for the norm referenced evaluation model. Two significant requirements are that:

1. The same test or equivalent tests and the same testing schedule must be used for all projects whose results are to be aggregated
2. An equal-interval scale (e.g. standard score, normal curve equivalent) should be used to aggregate across individual projects

Straight aggregation of results across different projects are not interpretable if different and nonequivalent tests are used. Likewise, gains based on varied lengths of time between the pretest and the posttest cannot be comparable and, therefore, cannot be meaningfully aggregated.

Choosing a score for a non-normed test presents a problem. Often the requirement for an equal interval scale is impractical and another scale must be used.

Strengths Two strengths of the pre-post matched scores model are:

1. A local control group is not required
2. Change is measured
3. A test may be chosen to measure the objectives of the program more accurately

The value that the pre-post matched scores design might have depends on the extent alternative explanations for growth can be ruled out. Assuming alternative explanations for observed gains are ruled out, the pre-post matched scores design can be implemented without a control group and is more flexible than the norm-referenced evaluation model.

Limitations Among the limitations of the pre-post matched scores design are:

1. Results are difficult to interpret and potentially misleading because the design does not isolate the program as the reason for changes that may be detected
2. The results from different tests cannot combine without technically demanding equating studies
3. Testing effects may contribute to gains when the interval between the pretest and posttest is relatively short (i.e. six weeks or less)

Conditions Supporting Use The pre-post matched scores design may be used to evaluate short term projects. Summer projects are generally short and take place at a time when the regular school program is not a competing source of learning. On the other hand, because of the short time period between the pretest and the posttest, some gain may be due to the effects of testing alone.

When all projects in a state follow a common curriculum, working toward the same instructional objectives, it should be appropriate to coordinate testing across projects in a state and aggregate scores for a statewide report.

Recent Use Two cases in which the pre-post matched scores approach has been used are in states that have only summer programs. One state used non-normed tests developed by the publisher of the curriculum materials used statewide in their program. Students were tested only in those subject areas in which they received instruction and the results were aggregated to give statewide summer gains in reading, mathematics and language arts. Because students would not otherwise be in school, regular school effects can be ruled out as contributing to observed gains. However, testing effects may still contribute to observed gains.

A second state pre and posttested with a nationally normed test that gives global scores in reading, mathematics and spelling. All students in the summer Chapter 1 migrant programs were given the complete test battery, as all students received instruction in all three subject areas.

Posttest-Only Design

Purpose A posttest-only design assesses the performance of students after they participate in an instructional program. On occasion when standards for skill attainment are established, a criterion referenced test may measure those skills after the student completes the program. Strictly speaking, however, the posttest-only design does not measure program effectiveness unless it can be safely assumed that the skills taught were not in the

student's repertoire before the instruction. If this assumption is unreasonable, then the posttest-only design is a very weak approach to evaluation. As deficient as the pre-post matched score design is, it does measure the post-program status of the students.

Design Features All or a representative sample of all students in the program are tested shortly after the completion of the program.

Testing Features The scores reported as measures of skills must adequately sample the skill domains taught and to be measured. This requires that skill domains be well defined and that multiple items be used to measure each skill.

Aggregation To aggregate the results from independent instances of the posttest-only design, that the same test or equivalent tests must be used across all programs for which the data are to be aggregated.

Strengths The posttest-only design is not generally useful as an evaluation method. Therefore, no strengths are listed.

Limitations The limitations of the posttest-only design are not so numerous as they are serious. The major problems are:

1. Lack of the controls provided by a comparison group
2. Lack of control of pre-program status of the students served
3. The students on whom data are collected may not be representative of the students receiving services

Conditions Supporting Use If the evaluation can establish that students lack the skills being measured before entering the program, then it might be possible to demonstrate that the program has been effective. If the evaluation can combine such evidence with proof that students had no other opportunity to learn the skills measured, then the posttest-only design using criterion referenced measurement may support claims that the program was effective.

Recent Use Eleven states reported using criterion-referenced testing as the primary method for evaluating their migrant programs. While it is not always clear from the state reports, it appears that the posttest-only design was used in most cases.

Assessment Programs

Purpose In general, state assessment programs produce information about the achievement status of all or some subset (e.g., fourth graders, high school seniors) of a state's student population. Closely related to state

assessment approaches to migrant program evaluation, point-in-time assessment involves selecting a test and testing schedule specifically for the migrant students in a state. The purpose of both state assessment and point-in-time assessment is descriptive; that is, they assess a group's level of performance (e.g., skills mastered, percent of objectives mastered, national percentile rank) without identifying or evaluating possible causes (e.g., participation in the migrant program) for the observed performance.

State assessment programs are frequently established by state legislation which determines the grade levels, subject areas to be assessed and the time of year for the testing. States using state assessment data for their migrant evaluation are limited to a test selected for reasons that may be unrelated to the objectives of local migrant educational programs. Migrant specific point-in-time assessment, is not constrained by these limitations.

Design Features The three major design characteristics of the two assessment approaches are:

1. States administer the same test to students (all or a random sample) in selected grade levels at the same time each year (or periodically)
2. Migrant students are identified for separate analyses
3. The test, if norm-referenced, should be given near the empirical norm dates to facilitate interpretation

Because it is not specifically tailored to assessing the migrant student population, state assessment may not yield results representative of the migrant students served by the migrant program. Because it is designed specifically for the migrant program, point-in-time assessment only samples the migrant student population in a state.

When a norm referenced test is given at a time other than the empirical norm date, it is necessary to compensate for testing away from the norm date to produce percentiles that will validly represent the group status. Compensation consists of interpolating to obtain an estimate of the percentile rank that a group would attain if the norms were referenced to the date the test was given. If the score used to report the status of the group tested is a percentile, then the further away from the norm date, the more tenuous the interpolated percentiles. On the other hand, if the scores reported are scaled scores or raw scores, it is not as critical that the test be given at the empirical norm date.

Testing Features In its most straight forward form, the major testing requirement is that the same test or test battery is given to all students across whom results are to be aggregated. All states using the assessment approach have done this. Variations of this basic approach are possible. Two variations include matrix sampling, in which not all students take the same test, and item banking approaches, in which items have been scaled within subject matter areas. There were no instances of matrix sampling or item banking approaches reported for the 1981-82 or 1982-83 program years.

Aggregation As implemented in states using an assessment approach, the same test is given to all students to be included in the aggregate. More generally, the requirement is that the same scale or equivalent scales are included in any aggregate.

Strengths The strengths of the two assessment approaches are:

1. Designed to give consistent state level data from year to year
2. Likely to produce test data on more students than approaches requiring matched pre and posttest scores
3. Allows comparisons with the general population of students in those grade levels sampled

Limitations Major limitations of the state assessment approaches are:

1. Results cannot be attributed to participation in the migrant program
2. Approaches may require additional testing at the local level
3. The test used may not be the best match to the objectives of the migrant program unless the point-in-time approach is used and the test is selected to match common objectives for the migrant program in the state

Assessment approaches are by their nature descriptive. As such, they do not result in information about the effectiveness of the migrant program in affecting changes in student achievement. This point is as much a limitation of point-in time assessment as of state assessment. Assessments are simply not designed to support the causal inferences needed to determine program effectiveness.

State assessment programs may result in additional testing at the local level when the district has its own testing program whose needs cannot be meet by the data from the state program. This limitation also applies to point-in-time assessment.

State assessment programs select tests or items for reasons that are usually independent of the migrant program curriculum. As a result, the match between the goals of the migrant program instruction and the state test is

usually not determined. However, point-in-time assessment should result in a better match between the migrant program objectives and the test used.

Conditions Supporting Use The major condition for the state assessment approach is that there is a state assessment program measuring skills taught in the migrant projects in the state. Secondarily, it is essential that there be a valid procedure to identify the migrant students tested. The point-in-time assessment is supported to the extent that there is a common set of goals or objectives associated with the migrant program in the state and that agreement on a most appropriate test can be established.

Recent Use Five states have used general state assessment to evaluate their migrant programs for either, or both, of the 1981-82 and 1982 [illegible] program years. One state has used the migrant specific form of state assessment. Their purpose was to describe the achievement level of migrant students at one point in time during the summer term. In four cases, migrant student status was reported as their national percentile standing by grade level. In one case the state reported the raw score averages for the migrant students.

Where the same test had been used for more than one year, the state reported the raw score achievement trend for migrant students.

Summary of Evaluation Approaches

This Section has outlined four approaches which states have used to evaluate Chapter 1 migrant education programs. Each approach used objective

measures of achievement in the basic skills, and each was the basis for a state summary included in an annual state Chapter 1 migrant program report to ED. An overview of each approach discussed its purpose, design features, testing features, strengths, and limitations. We also commented briefly on the conditions that would support using each approach. Table 1 summarizes the analysis of the approaches outlined.

Resources useful in planning, developing or improving Chapter 1 migrant evaluation systems include consulting services and numerous resource documents. TAC services are available for Chapter 1 program evaluation and program improvement. Each TAC has direct access to the TAC Materials Clearinghouse which houses extensive workshop materials and research documents related to Chapter 1 programs.

Some of the more useful resource documents available include:

- The Model A Owner's Manual (Demaline and Rader, undated)
- Test Information Summaries for Chapter 1 Evaluation (Strand, 1984)
- Characteristics of Selected Tests (NWREL, 1986)
- Interpretation Guide for Chapter 1 Evaluation Results, Second Edition (Davis, Deck, Demaline, in press)

TABLE 1

ANALYSIS OF EVALUATION OPTIONS
FOR CHAPTER 1 MIGRANT PROGRAMS

CHARACTERISTICS	EVALUATION APPROACHES			
	Norm-Referenced	Pre-Post Matched	Posttest Only	Assessment Programs
DESIGN FEATURES	Students are pre and posttested. Only students with a pretest and posttest are included in the analysis Tests are administered near the empirical norm dates for the test used. Pretest scores are not used to select students for the program unless they are the sole basis for selection and a statistical correction is applied.	Students are tested prior to and after their participation in the program. Only students with a pretest and a posttest score are included in the analysis.	All or a representative sample of all students in the program are tested shortly after the completion of the program.	States administer the same test to students (all or a random sample) in selected grade levels at the same time each year (or periodically). Migrant students are identified for separate analyses. The test, if norm-referenced, should be given near the empirical norm dates to facilitate interpretation.
TESTING FEATURES	The same test series is used for the pretest and the posttest. The test used must have national norms or be equated to one that does.	The same test is used for the pretest and the posttest. The test is administered according to the same procedures for the pretest and the posttest.	The scores reported as measures of skills must adequately sample the skill domains taught and to be measured. This requires that skill domains be well defined and that multiple items be used to measure each skill.	In its most straight forward form, the major testing requirement is that the same test or test battery is given to all students across whom results are to be aggregated. All states using the assessment approach have done this. Variations of this basic approach are possible. Two variations include matrix sampling, in which not all students take the test, and item banking approaches, in which items have been scaled within subject matter areas. There were no instances of matrix sampling or item banking approaches reported for 1981-82 or 1982-83 program year.
AGGREGATION	Where results are aggregated and reported across different tests, the metric of choice is NCEs. Results are combined within, but not across, subject areas. Results are aggregated within comparable testing cycles. Data must pass through quality control screens.	The same test or equivalent tests and the same testing schedule must be used for all projects whose results are to be aggregated. An equal-interval scale (e.g. standard score, normal curve equivalent) should be used to aggregate across individual projects.	To aggregate the results from independent instances of the posttest-only design, that the same test or equivalent tests must be used across all programs for which the data are to be aggregated.	As implemented in states using an assessment approach, the same test is given to all students to be included in the aggregate. More generally, the requirement is that the same scale or equivalent scales are included in any aggregate.

TABLE 1 (CONT.)

ANALYSIS OF EVALUATION OPTIONS
FOR CHAPTER 1 MIGRANT PROGRAMS

CHARACTERISTICS	EVALUATION APPROACHES			
	Norm-Referenced	Pre-Post Matched	Posttest Only	Assessment Programs
STRENGTHS	Local project effectiveness is measured without a local control group. Results can be readily combined across a number of projects to give district, state and national effectiveness. Well established routine for evaluating Chapter 1 projects across the country.	A local control group is not required. Change is measured. A test may be chosen to measure the objectives of the program more accurately.	The posttest-only design is not generally useful as an evaluation method. Therefore, no strengths are listed.	Designed to give consistant state level data from year to year. Likely to produce test data on more students than approaches requiring matched pre and post-test scores. Allows comparisons with the general population of students in those grade levels sampled.
LIMITATIONS	Not appropriate for measuring projects of short duration. Limited to outcomes measured by tests with national norms or equated to nationally normed tests. Of limited value for evaluating projects with high student turnover between the pretest and the posttest. Results confound the effects of Chapter 1 migrant and Chapter 1 regular programs for students participating in both. May be misleading if migrant students are not represented in national norms.	Results are difficult to interpret and potentially misleading because the design does not isolate the program as the reason for changes that may be detected. The results from different tests cannot combine without technically demanding equating studies. Testing effects may contribute to gains when the interval between the pretest and posttest is relatively short (i.e. six weeks or less).	Lack of the controls provided by a comparison group. Lack of control of preprogram status of the students served. The students on whom data are collected may not be representative of the students receiving services.	Results cannot be attributed to participation in the migrant program. Approaches may require additional testing at the local level. The test used may not be the best match to the objectives of the migrant program unless the point-in-time approach is used and the test is selected to match common objectives for the migrant program in the state.

V. SUMMARY AND CONCLUSION

It is unlikely that a single model will satisfy all a state's needs for Chapter 1 migrant program evaluation information. Ideally, each state agency should develop an evaluation plan that, when implemented, will result in a comprehensive and balanced evaluation of the programs that it operates. For evaluation to be comprehensive, all program areas serving a significant number or proportion of all students served should be addressed. It is undesireable for a significant program area to be omitted from evaluation because the primary evaluation approach used is not suitable for evaluating those programs. This does not mean that all programmatic areas should be evaluated by standardized achievement tests. Other non-test approaches to evaluation may be needed.

Balance refers to an ideal in which program evaluation information addresses program areas in relation to the number of students served or the costs of the services provided. A rough check of the comprehensiveness and balance of a state's evaluation is to compare the number of students who are included (or who theoretically have a chance to be included) in the evaluation results for each program area and the number of students who are included in the participant counts for each program area.

What will be balanced and comprehensive will depend on the state, as state programs vary greatly in size and complexity. Some states only operate a summer program in limited subject areas for a few hundred students. Other states, serving well over 100,000 students, have programs in both the regular term and the summer term serving students in a number of subject areas including reading, language arts, mathematics and language development. States with greater program complexity require multiple evaluation methods to

produce a comprehensive and balanced evaluation of the state's programs.

A state Chapter 1 migrant evaluation plan may be developed according to the following seven steps:

1. Develop a state profile of students served
2. Establish evaluation priorities
3. Select evaluation options in relation to priorities
4. Develop a long range plan for areas to evaluate
5. Implement the plan
6. Periodically make the evaluation results public
7. Periodically review the utility of the evaluations and revise the plan as needed

The state profile consists of a series of data displays that will describe the Chapter 1 migrant program participation at the state level. The specific data displays that will be useful to a state will depend on the programs in that state. However, a basic set of displays would include distributions of the number of students participating by:

1. Grade level
2. Instructional area
3. Migrant status
4. English language proficiency

These distributions should be displayed to distinguish between regular term programs and summer term programs. States with individual student data bases may be able to create more detailed analyses to show such things as instructional area emphasis by grade level and term of instruction. A state's evaluation priorities should reflect the relative emphasis of the services its

programs provide, with the reminder that at least part of its evaluation efforts assess student achievement gains in basic skills.

Common program participation data which are needed for profiling should be available for the 1984-85 program years. A state using these data to determine its pattern of services will be in a position to set evaluation priorities based on data.

As general advice, the norm-referenced evaluation model will be useful in states having a large number of former migrant students in regular term reading or mathematics programs at or above the second grade. States using the norm-referenced model have had large numbers of students who are available for both pre and posttesting on a fall-spring or annual testing cycle. The evaluations in these states have provided useful information on the gains achieved by students being served by migrant programs. States should continue to use the norm-referenced evaluation model to evaluate reading and mathematics programs serving a large number of former migrant students. Additional evaluation, however, may be warranted where significant programs are not represented through results from the norm-referenced evaluation model.

Some state programs do not support requirements for validly implementing the norm-referenced model or norm-referenced testing for evaluation with other models. For example, a number of states operate short term summer programs. These programs are not appropriately evaluated with the norm-referenced model for three reasons.

First, testing near the grade level norm dates of a standardized achievement test is not always feasible. When these summer programs serve students who move into a state because of summer labor demands, it is unlikely they will be found in the same area in the early spring and again in the fall when most tests are normed. In theory, these students may be located and tested with a norm referenced achievement test in the fall following their

summer program participation (posttest-only design). However, the posttest-only design is useful only if the testing is highly specific to the objectives of the program being evaluated, if evidence shows standards of mastery are established, and if the skills tested were not in the students' repertoire before they entered the program. Control over what is tested will be in the hands of the agencies where the students reside at the time testing, rather than the agency operating the summer program. Because all students who move from a particular location in the summer are not likely to be in the same location in the fall, follow-up testing would be a major coordination burden.

Second, short term programs must focus on a narrower range of skills than can be taught to in a regular term program, which may last 30 to 36 weeks. Norm referenced achievement tests are measures of global achievement and, therefore, are not likely to be a good match with the objectives of a short term summer program. Therefore, a norm referenced achievement test is unlikely to be a sensitive measure of short term programs.

Third, some summer programs provide instruction in a variety of basic skills depending on the needs of individual students in the program. Some students may receive primarily reading instruction, while others receive only language instruction or only mathematics instruction. Other students may receive instruction across two or three instructional areas. Short term programs that are individualized in the sense that they provide instruction across different subject areas do not lend themselves well to using a single achievement test as an evaluation tool.

Summer programs and other short term programs need to use achievement tests that focus on the more specific skills taught in these programs. A good quality curriculum embedded test, a test constructed especially for a given curriculum, or a criterion referenced test are good choices for these programs. One may also consider using an item bank approach to evaluating

short term programs.

While it means giving up the control offered by a comparison group, summer programs may follow the pre-post matched scores design. An occasional comparison group may help determine if testing effects, regression effects, or other factors account for gains observed. There should be no contamination from regular school effects because summer programs are offered when regular term programs are not offered.

Data from tailored tests that are not norm referenced are not easily aggregated into an overall state report unless the same test can be given to students receiving instruction for the same objectives. If there is not a common curriculum, the state must consider the trade-offs between different purposes of evaluation.

Evaluation for program improvement, which is one major purpose for evaluation, calls for outcome measures that are highly congruent with the objectives of individual programs. This means that a state level evaluation would contain individual program evaluations summarized in some fashion. Such reports have been produced in the past; they tend to be tedious reading for those who are expecting to see a highly uniform and quantitative report on the effects of the migrant program at different grade levels.

Evaluation for publicizing state level information is a second purpose for evaluating Chapter 1 migrant programs. The law requires states to make their evaluation findings public at least every two years, but they are not required to report aggregated results. ED forms for the 1984-1985 program year asked states to provide statewide summaries of the achievement information in the appropriate subject matter areas (reading, mathematics and other) and to indicate whether the results reported are representative of the state's

program for migratory children. To the extent that similar models are used for similar programmatic areas, data will be more easily aggregated at the national level.

REFERENCES

Campbell, D.T., & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching. In N.L. Gage (Ed.), Handbook of Research on Teaching. Chicago: Rand McNally. (Also published as Experimental and Quasi-experimental Designs for Research. Chicago: Rand McNally, 1966.)

Davis, A., Deck, D., and Demal: R. (Undated). Interpretation Guide for Title I Evaluation Results. Portland, OR: Northwest Regional Educational Laboratory.

Demaline, R. and Rader, D. (Undated). Model A Owner's Manual. Portland, OR: Northwest Regional Educational Laboratory.

Goodwin, W.L. and Driscoll, L.A. (1980). Handbook for Measurement and Evaluation in Early Childhood Education. San Francisco, CA: Jossey-Bass Publishers.

Haney, W. (1978). ESEA Title I Early Childhood Education: Review of Literature on Evaluation and Instrumentation. Cambridge, MA: The Huron Institute.

Haney, W. (1980). Short Term Impact Evaluations of Early Childhood Title I Programs. Cambridge, MA: The Huron Institute.

Jenkins, J.A. (1986). A National Summary of Achievement Information as Reported by State Migrant Education Programs for Fiscal Years 1982 and 1983. Interim draft prepared for the United States Department of Education. Atlanta, GE: Educational Testing Service.

Kennedy, M. (1980). Longitudinal Information Systems in Early Childhood Title I Programs. Cambridge, MASS: The Huron Institute.

Linn, R.L. (1982). The validity of the Title I Evaluation and Reporting System. In E.R. Reisner, M.C. Alkin, R.F. Bowch, R. L. Linn, J. Millman (Eds.), Assessment of the Title I Evaluation and Reporting Sytem. Washington, D.C.: U.S. Department of Education.

Naccarato, R.W. (1986). Participation Data for the Migrant Education Program Fiscal Years 1982 and 1983: A National Summary. Draft report prepared for the United States Department of Education. Indianapolis, IN: Advanced Technology.

Northwest Regional Educational Laboratory (1986). Characteristics of Selected Tests. Portland, OR: Northwest Regional Educational Laboratory.

Plato, K.C. (1984). Program for Migrant Children's Education: A National Profile. National Association of State Directors of Migrant Education.

RMC Research Corporation (1981). Evaluator's References: Volume II of ESEA Title I Evaluation and Reporting System Documentation. Prepared for the United States Department of Education Office of Program Evaluation. Mountain View, CA: RMC Research Corporation.

Strand, T. (1984). Test Information Summaries for Chapter 1 Evaluation. Evanston, IL: Educational Testing Service.

Tallmadge, G.K., Lam, T.C.M., and Camarena, M.L. (1985). The Evaluation of Bilingual Education Programs for Language-Minority, Limited-English-Proficient Students: A Status Report With Recommendations for Future Development. Report prepared for the United States Department of Education. Mountain View, CA: SRA Technologies, Inc.

Tallmadge, G.K., Wood, C.T., and Gamel, N.N. (1981). User's Guide: ESEA Title I Evaluation and Reporting System. Washington D.C.: U.S. Department of Education.

Yurchak, M.J. (1980). ESEA Title I Early Childhood Education: A Descriptive Report. Cambridge, MA: The Huron Institute.

Appendix A

USING PRETEST SCORES TO SELECT CHAPTER 1 PARTICIPANTS IN THE NORM-REFERENCED EVALUATION MODEL 1/

Pretest scores may be used to select Chapter 1 participants only when it is possible to adjust the pretest mean for the bias that will be introduced by this procedure.

Rules for implementation. The rules presented in this section apply only to this variation of the basic model.

- o A cutoff score must be established on the pretest such that all students scoring below it will be included in the project and all those scoring above it will be excluded. This cutoff score must be strictly adhered to.

- o Times of pretesting and posttesting should follow the rules prescribed for the basic model.

- o A reliability figure (test-retest, alternate form, etc.) must be available for the test's norming sample. (This will be needed when the pretest mean is corrected for regression effect bias.)

1/ Taken from: Tallmadge, G.K., Wood, C.T. and Gamel, N.N. (1981) User's Guide: ESEA Title I Evaluation and Reporting System. Washington D.C.: U.S. Department of Education, pgs. 31-35.

Sequence of Steps for Implementation

Step 1. Select a nationally normed achievement test on the basis of its content and the time or times of year for which it has empirical norms. A reliability figure for the test's norming sample must be available from the test publisher.

Step 2. Pretest the group from which project participants will be selected. Administer the test within two weeks (or six weeks, using projected norms) of either side of the midpoint of the period during which the norm group was tested. Make-up tests should be administered not more than two weeks after the initial testing. Carefully follow the procedures outlined in the publisher's manual.

Step 3. Establish a cutoff score on this test, below which all students will be assigned to the Chapter 1 group and above which no students will participate in the project. (One way to do this is to rank order the students, from lowest to highest, on the basis of their pretest scores. After determining how many students can be served by the project, begin with the lowest scoring student and count down the list until the number of selected students equals the number of openings.)

Step 4. After the project, posttest the participants. Posttesting must be conducted within two weeks (or six weeks, using projected norms) of the midpoint of the period when the norm group was tested. Make-up posttests should be given within two weeks of the original posttest. All tests should

be administered according to the procedures that were used when the norming sample was tested. Every effort should be made to locate and posttest students who were pretested and who had a significant amount of involvement in the project.

Step 5. Score the tests and record the posttest scores.

Step 6. Identify those students who have both a pretest and posttest score. Convert their raw scores to expanded standard scores or NCEs and calculate pretest and posttest means.

Step 7. Correct the pretest mean for selection on the pretest using equation 7. (For the derivation of this equation, see Roberts, 1980).[2/]

$$\bar{X}'_p = \bar{X}_p + [(1 - r_{xx})(\bar{X}_t - \bar{X}_p)] \qquad (7)$$

where

$\bar{X}'_p$ = corrected mean pretest score of the Chapter 1 group.

$\bar{X}_p$ = the mean score of the Chapter 1 group on the selection/pretest measure.

2/ Roberts, A.O.H. Regression to the mean and the regression effect bias. Mountain View, CA.: RMC Research Corporation, October 1980.

$\bar{X}$ = the mean score of the total group (from which the Chapter 1 students were selected) on the selection/pretest measure.

r_{xx} = the test-retest reliability for the total group.

The means of the Chapter 1 group and of the total group can be calculated directly from the available test data. However, unless the test is administered a second time to the total group, the test-retest reliability cannot be exactly determined and must be estimated from the test-retest reliability for the norming sample (p_{xx}). Not all test publishers report test-retrest reliability figures. If test-retest reliability is not reported, but some other reliability coefficient is available, the test-retest reliability can be estimated by adding a correction factor to the reported figure. Table 1 shows the constant to be added to each reliability to obtain an estimate of the national norming sample's test-retest reliability. For example, a publisher reports a KR 20 reliability figure of .96 for the test level of interest. Using the correction factor in Table 1 for this type of reliability, the norming sample's test-retest reliability would be estimated to be (.96) + (-.09) or .87.

TABLE 2

The Constants to be Added to Each Type of Reported Reliability to Estimate the Norming Group's Test-Retest Reliability (p_{xx})

	Reliability Reported	Correction Factor
Norming group's test-retest reliability (p_{xx})	= Alternate form reliability (When 2 to 3 wks separated administrations)	+ -.01
	= Alternate form reliability (When 2 to 3 days separated administrations.)	+ -.06
	= KR 20 (or alpha) reliability	+ -.09
	= Corrected split half reliability.	+ -.09

A second problem with the test-retest reliability figure is that it is likely to be significantly lower when the test is used in a Chapter 1 schoold than when it is administered to a nationally representative norming sample. To obtain a better estimate of the local total group's test-retest reliability (r_{xx}) for use in equation 7, the norming group's reliability coefficient (p_{xx}) should be adjusted sing equation 8 below.

$$r_{xx} = 1 - \frac{o^2}{s^2} (1 - p_{xx}) \qquad (8)$$

where

p_{xx} = the test-retest reliability for the norming sample.

o^2 = the variance of test scores for the norming sample.

s^2 = the variance of test scores for the total group from which the Chapter 1 students were selected.

The estimated test-retest reliability for the local group can then be used in equation 7 to adjust the pretest mean.

Step 8. If the means are in standard scores, convert them to NCEs. Subtract the corrected mean pretest NCE (the no-project expectation) from the mean posttest NCE. The difference is the NCE gain.