DOCUMENT RESUME

ED 142 564 95 TM 006 184

AUTHOR Kaskowitz, David H.; Norwood, Charles R.
TITLE A Study of the Norm-Referenced Procedure for Evaluating Project Effectiveness as Applied in the Evaluation of Project Information Packages. Research Memorandum.
INSTITUTION Stanford Research Inst., Menlo Park, Calif.
SPONS AGENCY Office of Education (DHEW), Washington, D.C. Office of Planning, Budgeting, and Evaluation.
REPORT NO SRI-URU-3556
PUB DATE Jan 77
CONTRACT OEC-0-74-9256
NOTE 120p.; Paper presented at the Annual Meeting of the American Educaticnal Research Association (61st, New York, New York, April 4-8, 1977); for related documents, see TM 006 366, 367, ED 122 373-375, and 460

EDRS PRICE MF-$0.83 HC-$6.01 Plus Postage.
DESCRIPTORS Achievement Gains; Black Students; Caucasian Students; *Demonstration Projects; Diffusion; Elementary Secondary Education; Field Studies; *Measurement Techniques; Minority Group Children; *Norm Referenced Tests; Program Validation; Racial Differences; Remedial Programs; Scores; *Scoring Formulas; Spanish Speaking; Standardized Tests; *Statistical Analysis; *Summative Evaluation
IDENTIFIERS Equipercentile Growth Assumption; Metropolitan Achievement Tests; *Project Information Packages

ABSTRACT

Project Information Packages (PIPs) are informative kits that describe remedial educational programs and contain instructions for installing the projects in a new site. Six such PIPs were evaluated using a norm-referenced procedure applied to standardized test scores. Pretest scores were compared to posttest scores which were calculated according to the equipercentile growth assumption. The statistical analysis is closely examined; the empirical validity of the equipercentile growth assumption, the stringency of the educationally significant growth criterion, and the statistical properties of the norm-referenced procedure are investigated. The basis for use of standardized tests, and criticisms of the standardization procedures are discussed. (Author/GDC)

Research Memorandum

January 1977

Session 15.11 at AERA

A STUDY OF THE NORM-REFERENCED PROCEDURE FOR EVALUATING PROJECT EFFECTIVENESS AS APPLIED IN THE EVALUATION OF PROJECT INFORMATION PACKAGES

By: David H. Kaskowitz
Charles R. Norwood

Prepared for:

U.S. Office of Education
Office of Planning, Budgeting and Evaluation
Washington, D.C. 20202

Contract OEC-0-74-9256

SRI Project URU-3556

FOREWORD

This study of norm-referenced procedures and criteria for determining project effectiveness was done as part of an evaluation of six original Project Information Packages (PIPs) in their first field tests.

Each PIP is an information kit describing an exemplary educational project and the steps necessary to install it in a new site. The RMC Research Corporation was contracted in 1973 to select exemplary projects suitable for packaging, to analyze them, and to design information packages that would allow school districts to replicate them.

The six projects selected were compensatory programs funded under Title I of ESEA (except for the R-3 program). They were:

- Project Catch-Up--A reading and mathematics lab program for students in the first through sixth grades.
- Project Conquest--A lab program for first through sixth graders limited to reading instruction and not requiring certificated reading specialists.
- High Intensity Tutoring (HIT)--A middle school reading and math laboratory-type program where peers tutor lower-achieving students.
- Intensive Reading Instruction Teams (IRIT)--A reading lab program for third, and some fourth graders, focusing on three content areas (decoding, vocabulary, and comprehension and individualized reading) each with its own teacher.
- Programmed Tutorial Reading (PTR)--A reading program for first graders. Paraprofessional tutors are guided in the presentation of the curriculum by programmed text.
- Project R-3--A reading, mathematics, and social studies program for junior high school students at a selected grade. This requires teachers to coordinate the curriculum to emphasize the relevance of school subject matter to life experiences and to motivate learning.

The original PIPs, based on these selected projects, were evaluated in their first field test sponsored by USOE in 19 sites. Based on first year findings, a revised set of PIPs was developed and are currently being disseminated.

The investigations reported in this Research Memorandum were a part of a second year evaluation of the original PIPs, which focused on effectiveness of the field test projects in terms of student achievement.

Because there were no control or comparison groups with which to compare participating students, a norm-referenced procedure, based on the procedures described by RMC in their selection of the original projects, was used to determine PIP project effects. This study explored our use of the norm-referenced procedures and associated criteria as applied in the PIP evaluation where the Metropolitan Achievement Test was administered to participants.

Marian S. Stearns, Director
Project Information Packages
Evaluation

CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

I INTRODUCTION

As part of SRI's two-year evaluation of the Project Information Package (PIP) field test, program impacts on student achievement have been assessed by means of a norm-referenced procedure applied to standard achievement test scores. The procedure, in general, consists of comparing observed posttest scores to expected posttest scores, where the expected posttest scores are derived from pretest scores under certain assumptions regarding growth in achievement if children had not participated in the PIP program.

The use of the norm-referenced procedure in the PIP evaluation is based on the way in which the original exemplary programs were selected by RMC for packaging. Because the exemplary programs were supposedly identified by a norm-referenced approach, a similar approach in the evaluation of the PIP field test is considered by some to be desirable for the sake of consistency.

The norm-referenced approach to evaluation has a tradition that precedes its implementation in the PIP evaluation. For example, the developers of the Metropolitan Achievement Tests (MAT) suggest a norm-referenced procedure to evaluate student growth (Prescott, 1973). Their procedure entails the use of the standard score gains at the 50th norm percentile between pre- and posttest as a criterion for normal growth. Horst et al.(1975) presented their own, more detailed norm-referenced procedure in their monograph, A Practical Guide to Measuring Project Impact on Student Achievement. This approach uses the standard score gain necessary to maintain the pretest percentile rank as a criterion for normal growth.

The norm-referenced approach is continuing to be sold as an important evaluation tool. For example, Gamel et al. (1975) expect the norm-referenced model to be the most widely adoptable of three models they have proposed in an evaluation and reporting system to be used by state and local education agencies for projects funded by Title I.

The Norm-Referenced Procedure

The norm-referenced procedure, as described by Horst et al.(1975) is constructed from a model of normal growth and educationally significant growth in achievement. The normal growth model sets a standard of performance for children who receive no treatment (i.e., children who would be considered members of comparison or control classrooms). In the absence of a control group, this approach relies on the norms of standardized tests to estimate how a group of children would have performed if no treatment had been present. A model of educationally significant growth sets the standard of performance of children in the treatment group relative to the performance of children who receive no treatment.

The normal growth model adopted by RMC and used in a slightly modified form in the PIP evaluation assumes that the expected normal growth for children who are not in a special education program is such that, on the average, children maintain the same percentile from pretest to posttest with respect to the norm population. Under this equal percentile assumption, for example, a child or group of children that scores in the tenth percentile on the pretest is expected to score in the tenth percentile on the posttest, all other things being equal.

No model for specifying educationally significant growth has yet been formulated with respect to standardized tests. The problem is one of linking the content and focus of the educational program with the content of the standardized tests by means of a theory of instruction. Nevertheless, as a rule of thumb Horst et al. (1975) proposed to use the standard deviation of the standard scores for the norm group to establish the criterion for educational significance. The criterion they proposed for a program to achieve educationally significant growth was that the difference between the mean posttest standard score and the expected mean posttest standard score under the normal growth assumption be at least one-third of the norm group standard deviation. For example, assume a third grade project had a mean standard score of 44 on the MAT Total Reading in the fall; this corresponds to a percentile rank of 10. Then, based on the normal-growth model and the MAT norm tables, the expected mean standard score would be 47 in the spring. One-third of the norm-group standard deviation was about 4 standard score points for Total Reading. Therefore, to achieve educationally significant growth, the data would need to indicate that the mean standard score for the project in the spring was 51 or more.

Analytically, the criteria of normal growth and educationally significant growth can be expressed as follows, disregarding for the moment variation caused by sampling and measurement:

Let $\bar{X}$ = mean standard score on the pretest.

Let $\bar{Y}$ = mean standard score on the posttest.

$g(\cdot)$ = transformation from standard score to percentile score on the pretest.

$f(\cdot)$ = transformation from standard score to percentile score on the posttest.

$f^{-1}(\cdot)$ = transformation from percentile score to standard score on the posttest.

σ = standard deviation of the norm-group standard scores.

The criterion for normal growth is satisfied if

$f(\bar{Y}) \geq g(\bar{X})$ in terms of percentiles, or equivalently

$\bar{Y} \geq f^{-1}[g(\bar{X})]$ in terms of standard scores.

The criterion of educationally significant growth is satisfied if $f(\bar{Y}) \geq f(f^{-1}[g(\bar{X})] + \sigma/3)$ in terms of percentiles, or equivalently, $\bar{Y} \geq f^{-1}[g(\bar{X})] + \sigma/3$ in terms of standard scores.

In the presence of sampling variability, Horst et al. (1975) recommended the use of the following statistic to test for normal growth:*

$$T = \frac{\bar{Y} - f^{-1}[g(\bar{X})]}{\sqrt{\frac{s_x^2 + s_y^2 - 2r_{xy}s_x s_y}{N}}} \qquad (1)$$

where:

s_x = estimated pretest standard deviation

s_y = estimated posttest standard deviation

r_{xy} = estimated correlation between the pretest and posttest

N = number of children who have both pre- and posttest scores.

Horst et al. implicitly assume that T has a Student's t distribution with N-1 degrees of freedom under the null hypothesis of less than normal growth. They recommend that the null hypothesis of less than normal growth be tested by a one-tail t test at the $\alpha = .05$ level of significance.

Horst et al. do not suggest a statistical test of whether educational significance has been attained. Apparently, the criterion was intended as a supplement to the statistical test rather than as a totally separate criterion. That is, a gain might be statistically but not educationally significant, depending upon the number of children included in the analysis. For example, the gain necessary to display statistical significance decreases as the sample size increases, other things being equal. A gain that was a bit over that expected could be statistically significant, although it might not be judged to be educationally significant.

It is also possible under Horst's scheme that a gain could be educationally significant without being statistically significant if, for example, only a few students were in the evaluation. However, we feel it is logical to assume that a gain must be statistically significant in order to consider its educational significance.

Our approach was to test for the statistical significance of educationally significant growth in a similar fashion to the test con-

*Horst et al. (1975) had N-1 in Equation (1) rather than N. This appeared to us to be a typographical error because, under the usual assumption for deriving the t statistic, N was the appropriate term.

ducted for normal growth. A 95% confidence interval was constructed for the difference between the mean standard score in the spring and the expected spring standard score:

$$\bar{Y} - f^{-1}[g(\bar{X})] \pm t_{.025,N-1}\sqrt{\frac{s_x^2 + s_y^2 - 2r_{xy}s_x s_y}{N}}$$

where $t_{.025,N-1}$ is the upper .025 point of the t distribution whith N-1 degree of freedom.

The procedures for testing whether the criteria were met were based on the position of this confidence interval, or equivalently on the value of the statistic T as defined in Equation (1). For normal growth:

if $T > t_{.025,N-1}$, the normal growth criterion is satisfied;

if $T < t_{.025,N-1}$, the normal growth criterion is not satisfied;

otherwise, it is considered to be unknown whether the normal growth criterion is satisfied. For educationally significant growth:

$$\text{Let } T' = \frac{\bar{Y} - f^{-1}[g(\bar{X})] - 1/3\sigma}{\sqrt{\frac{s_x^2 + s_y^2 - 2r_{xy}s_x s_y}{N}}}$$

if $T' > t_{.025,N-1}$, the educationally significant growth criterion is satisfied;

if $T' < t_{.025,N-1}$, the educationally significant growth criterion is not satisfied;

otherwise, it is considered to be unknown whether the normal growth criterion is satisfied.

The data collected in 1974-75, for example, showed that 44 Catch-Up second graders gained an average of 7.6 standard score points between fall and spring for Total Reading. This was 2 points less than that expected based on the normal growth model. The value of T was -3.61 and T' was -9.03, indicating that neither criterion was achieved. Nineteen Catch-Up fifth graders had an average gain of 6.67 standard score points between fall and spring for Total Reading, 1.47 points more than the expected spring score. The value of T was 7.82 and T' was -13.5. Thus, the normal growth criterion was achieved, and the educationally significant growth criterion was not.

Hazards and Criticisms of the Norm-Referenced Procedure

As pointed out by Horst et al. (1975) the strength of the norm-referenced procedure is that it eliminates the need of a control group.

As a result, the norm-referenced procedure can be implemented at much less cost and effort than the conventional design that requires a control group. Furthermore, in situations where a control group is not feasible because of practical or political circumstances, the norm-referenced procedure can still be employed.

The potential weaknesses of the procedure lie in assumptions underlying the use of standardized tests, the validity of the equal percentile assumption, the arbitrariness of the criterion for educationally significant growth, and some of the details of the statistical computations.

Assumptions Underlying the Use of Standardized Tests

Criticisms have been made of the use of standardized achievement tests in the evaluation of program impacts. These criticisms have focused on the divergence between the principles under which the standardized tests are constructed and the objectives of evaluation. According to this point of view, the assumptions and motivations under which standardized tests are constructed are not compatible with assessment of program impacts because test items tend to be irrelevant to program curriculum and objectives. The common counterargument is that above and beyond the specific objectives of an innovative educational program, a positive effect on student performance on standardized tests should be required. Improvement in standardized test scores, then, becomes an added objective for programs to be shown to be effective irrespective of the correspondence between program objectives and curriculum and test content. The test content may be ignored, according to this point of view, because all standardized achievement tests produce measures of "achievement."

Test developers and evaluators advocating the use of standardized tests do acknowledge, however, the relevance of test content. For example, the developers of the MAT, in discussing the validity of their test, state that the question of greatest importance to the potential test user is whether the course content and objectives correspond with those measured by the text (Prescott, 1973). Horst et al. (1975, p. 5) point out that "in order to get a good measure of how students performed, the evaluator must select an appropriate test and ensure that it is administered and scored correctly." Nevertheless, in program evaluation, all too often the test content is ignored under the assumption that the test is measuring some underlying trait of achievement that is independent of educational program objectives and content.

Validity of the Equal Percentile Assumption for Normal Growth

The major assumption of the normal growth model is that in the absence of an intervention program, the normal percentile rank of a child, class, or site will not change between the pretest and the post-test. The norm sample of most standardized tests is selected to be representative of the entire population of children in school. However, in the evaluation of special educational programs such as the PIPs, the children in the evaluation are commonly atypical of the entire school

population. For example, children in Title I compensatory education programs have a much lower socioeconomic status than the entire population and are more likely to be members of minority, racial, or ethnic groups. By the very nature of these programs, the children tend to have percentile ranks below that of the norm group on standardized tests. Whether these children maintain their standing relative to the norm group in the absence of a particular educational program is an empirical question that, up to now, had not been examined in any depth.

Other issues regarding the equal percentile assumption are broader in scope, relating to the way in which the standardized tests were normed and the articulation between levels of the test. Standardized test batteries commonly consist of a series of test levels where each level is targeted for children in a particular age range or grade range. For example, the MAT consists of six battery levels: Primer, Primary I, Primary II, Elementary, Intermediate, and Advanced. The grade levels for which each level of the test was primarily intended are given below:

Battery Level	Grades Primarily Intended
Primer	K.7 - 1.4
Primary I	1.5 - 2.4
Primary II	2.5 - 3.4
Elementary	3.5 - 4.9
Intermediate	5.0 - 6.9
Advanced	7.0 - 9.5

The battery levels and test forms are commonly linked by means of a standard score metric. This metric may be used, supposedly, to interpret student achievement independent of the level of the test battery or the form of the test administered. The validity of the equal percentile assumption relies in many cases on the validity of the standardization metric and the linkage of the test levels, because the level of the test administered in an evaluation often does not correspond to the level prescribed by the test developers. For example, children targeted for innovative educational programs are often performing at a level that is below their peers, so they are likely to "bottom out" (i.e., get extremely low raw scores) on the prescribed level of the test. Bottoming out would result in test scores that would indicate only that the test was too difficult and would be of limited use for evaluation. Therefore, testing may need to take place below the prescribed level. In such cases, the in-level conversions of standard score to percentile are still used and the validity of their use is linked to the validity of the standardization and scaling programs.

Educationally Significant Growth Criterion

For testing for educationally significant growth, Horst et al. (1975, p. 74) advise:

> There is no generally accepted criterion for deciding whether the size of the gain is large enough to be considered educationally significant. Since standardized tests are used, the standard deviation of the national norm group (σ) provides a useful reference. As a rule of thumb, the authors suggest that if the observed posttest scores exceed the no-treatment expectation by one-third of a standard deviation, the treatment effect be considered educationally significant.

This advice raises a number of questions. What are the properties of the standard deviation of the national norm group that makes it a useful reference? Should the standard deviation from the pretest period or the posttest period be used? Why not use an estimate of the standard deviation of the difference between fall and spring? Why use one-third standard deviation? Why not take the sampling and measurement variability into account?

Rules of thumb such as that proposed by Horst et al. (1975) have a way of becoming rules on which policy decisions are based. In the absence of other criteria for determining educational significance, this is even more likely.

It is beyond the scope of this study to establish criteria for educational significance, but the statistical properties of the proposed criterion was investigated. Nevertheless, it cannot be stressed too strongly that the criterion for educational significance must be taken in the spirit that it was proposed, as a rule of thumb and no more. Policy decisions regarding program effectivenss should not be made solely on the basis of this criterion.

One property that we feel the criterion for educational significance should have is uniformity in stringency across grade levels and pretest scores. That a program shows a favorable impact should not depend solely on some artifact of the evaluation procedure. Our analysis in the 1975 PIP evaluation (Stearns, 1975) indicated that the stringency of the criterion of educational significance does vary with pretest score and grade for the MAT. The gain in standard score points necessary for passing the criterion of educational significance relative to the normal growth gain at the 50th percentile was shown to increase with grade level. There were also indications that the difficulty of passing the educationally significant growth criterion varies with pretest scores.

The Statistical Properties of the Procedure

Two questions were raised regarding the statistical properties of the norm-referenced procedure: 1) How senstive is the test to the unit of analysis? and 2) How good an approximation is the standard deviation

of the difference between the pretest and posttest scores to the standard deviation of the difference between the posttest and expected posttest scores?

The current procedure converts the mean pretest score to a predicted mean posttest score. The conversion could also be made for each pupil's score individually and then the individual predicted scores could be aggregated to yield the average posttest predicted score. Because the function that converts standard scores to percentiles in intended for use at the individual level, the question arises whether this alternative might yield different results.

Horst et al. (1975) recommend using the sample standard deviation of the difference between the posttest and pretest rather than the standard deviation of the difference between the posttest and the predicted posttest scores under the normal growth assumption. The relationship between these two variances depends on the form of the function that predicts the posttest score from the pretest score and the relationship between pre- and posttest scores. Again, the question arises as to the sensitivity of the norm-referenced procedure to this approximation.

Organization of the Working Paper

The remainder of the working paper is organized around the four areas of criticism discussed above. In Section II the basis for use of standardized tests and criticism of the standardization procedures are examined; the empirical validity of the equipercentile growth assumption is examined in Section III; in Section IV the stringency of the educationally significant growth criterion is investigated; Section V deals with a study of the statistical properties of the norm-referenced procedure. Finally, Section VI describes an alternative procedure and gives the conclusions of the study.

II USE OF STANDARDIZED TESTS

A critical aspect of any evaluation study is in the specification of program goals and the formulation of criteria for determining whether the goals have been achieved. Program goals and criteria may be specified internally by program developers or may be imposed externally by policy makers or evaluators. They may be formulated with regard to specific aspects of the program or they may be in a very general form that is supposedly applicable to a wide variety of programs. Obviously, the assessment of program success depends on the criteria selected. A program may be extremely successful in accomplishing specific internally established objectives, but unsuccessful in attaining some general externally imposed goal.

The norm-referenced analysis, by its very nature, is externally imposed and is, for the most part, not related to specific program objectives. That is, few, if any, educational programs set out with the specific goal of increasing students' scores on specific standardized achievement tests. Nevertheless, policy makers, evaluators, and the public have commonly imposed gains on standardized tests as a major criterion for program success.

Such is the case because standardized tests are convenient to use, they are the conventional instruments for measuring achievement, and they are considered by many to be "sensitive to any significant cognitive growth and should usually prove adequate for assessing the impact of special treatments (Horst et al., 1975, p. 5)." Use of these tests also allows the evaluator to use the extensive data compiled for norming the tests to assess program impacts.

Ralph Hoepfner (1976) stresses that the acceptability of standardized tests is a major factor in their use over criterion-referenced tests. His argument is that for an evaluation to have an effect on policy makers it is to the evaluators' advantage to use test instruments that are considered acceptable to these policy makers and standardized tests are much more acceptable than other types of instruments.

Also underlying the use of standardized tests is the notion that test scores are related to some scale of achievement. According to this "generic-true-score test theory" (Lord and Nevick, 1968), different tests are merely parallel test forms and therefore the choice of a particular test is not important, as long as it is a member of a broad group of acceptable tests.

Criticism of the Use of Standardized Tests

By no means is there consensus regarding the use of standardized tests to establish a generic criterion for evaluation of educational programs, especially programs targeted for low-income, minority children.

Most evaluators and test developers would agree that if a standardized test is to be used in an evaluation, it must be selected to be appropriate for the students and program being evaluated. The developers of the MAT, for example, include in their criteria for selecting an appropriate test battery the coverage of desired skill and subject matter and match between test and local objectives (Prescott, 1973). Horst et al. (1975) do mention the importance of selecting an appropriate test, but they leave it to the evaluator to establish his own criteria of what is appropriate.

Many evaluation experts would carry the argument against the use of standardized tests to the point of rejecting their use in large-scale evaluations. At the U.S. Office of Education Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation held in May 1976, for example, a number of participants criticized the use of standard achievement tests or emphasized the need for extreme care in their use for evaluation purposes.

Wargo (1976) pointed out that standardized tests have been criticized on the grounds that they had been developed "by, with, and for white middle-class America." Because of underrepresentation or no representation of minorities and disadvantaged students in various stages of test development, tests may have one or more biases against these populations. Furthermore, the procedures used in developing tests may make them useless for evaluating growth in low-achieving students.

Both Bianchini (1976) and Tyler (1976) stressed the importance of agreement between the content of the curriculum and the items on a test. Bianchini used experience from the California Miller-Unruh Statewide Testing Program at Grade 1 as an example. In the first year of the program when the Stanford Reading Test was used, about 65% of the first graders had reading scores within the first quartile range of the national norms. But when California changed to the Cooperative Primary Reading Test five years later, the scores for pupils were much closer to the national norms. Bianchini points out that a major factor in explaining these diverse results may be the degree of congruence of the tests with the curriculum materials provided by the state. Only 19% of the vocabulary in the Stanford Reading Test were contained in the instructional materials provided by the State at that time, as contrasted with 55% of the vocabulary used in the Cooperative Primary Reading Test. Bianchini's (1976, p.21) point was that "in any program at the early grades it is particularly important for all children that the test content be related to instructional content. The reasons for this is that little children learn only within the bounds of the curriculum they experience in the early grades."

Tyler makes the point that a number of different curricular programs are available to achieve the same common objective in reading or math. For example, reading comprehension can be achieved by any of a number of different sequences of instruction. Because such

instruction will generally take place over a number of years, the evaluator must either allow for enough time for the entire instructional sequence to take place prior to testing or must tailor his testing to the type of instruction that is taking place.

Tyler (1976, p.1) goes further to state that "we can learn very little about the strengths and weaknesses of programs of 'compensatory education,' or those designed for children of minority groups, from the results of these [norm-reference achievement] tests. At best, they are rough and imprecise measures, and most probably they are invalid." His reasons are based on the contrast between the purpose and procedures of test developers and the purpose and procedures of evaluation. While the purpose of a standardized test is to measure individual differences, the purpose of evaluation is "to find out how many have learned what the program seeks to teach." While the test developer proceeds under the assumption that test scores should be normally distributed, no such assumption is needed for evaluation purposes. Tyler's point was that the very procedures for developing standardized tests are those that make the tests invalid for evaluation purposes. He states:

> I have examined many of the items in contemporary tests and find that there are usually no more than 5% of them which represent behavior of the lowest third of the pupils and 5% which represent behavior of the highest third. I have also noted that a considerable fraction of the items are not common to most curriculums. To produce what is usually an artificial "normal distribution," the test has lost in validity because of its sampling bias. It has also increased the proportion of items that are not taught in most schools because these are more likely to differentiate among pupils. (Tyler, 1976, p.3)

Criticism of the MAT Standardization Procedure

One major advantage of standardized tests is that they are accompanied by norms; of course, this property makes the norm-referenced procedure possible. If we accept the validity of using a standardized test in evaluation, the norm-referenced analysis can be considered valid only to the extent that the norms are considered valid. In the next section, we will examine some of the empirical results regarding the validity of the norms, especially as they apply to the evaluation of educational programs targeted for disadvantaged children. In this section, we examine the standardization procedure and properties of the resulting norms.

Because the MAT was selected for the PIP evaluation, we have examined in some detail the procedure by which these tests were standardized and the properties of the resulting norms. The 1970 MAT, published by Harcourt, Brace, and Jovanovitch, was made to 1970 test-industry standards. The test is in some ways superior to other standardized tests of that period, principally because the MAT's

standardization program included both fall and spring testing.

The test makers provide a standard score scale that has the function of providing scores that are continuous across levels of the test. The publishers state (MAT Guidelines #1, p.1):

> The standard score scale for the Metropolitan Achievement Tests provides two basic conveniences for the test user. The scale makes forms within a battery equivalent and provides a continuous, equal interval system for each test across all material. Once raw scores are converted to standard scores, one need not be concerned in further interpretation with either the battery or the form from which the raw scores came.

The standard score scale is important in the norm-referenced analysis because the numerator and denominator of the test statistic are expressed in standard score units. In addition, the standard score scale is essential for out-of-level testing to link the raw score of a test to the associated percentile rank. Out-of-level testing is commonly recommended when it is anticipated that students will get extremely high or extremely low raw scores on the test that is at the level recommended by the test developers. Such scores are considered unreliable and invalid and not appropriate for use in evaluation.

Testing out of level would also arise if an evaluator followed the advice given by Horst et al. (1975) to administer the same level of a test for both pre- and posttesting. For the second and third grade, the level of the MAT changes from fall to spring. Therefore, testing at the same level pretest and posttest would necessitate out-of-level testing. Also, when the period between pretest and posttest spans more than one school year, the recommended test level often changes. The recommendation by Horst et al, however, was given so that an evaluator would avoid the potential measurement errors involved in out-of-level testing. It would appear that the advice of Horst et al., should have been always to test at the test-developer's recommended level rather than at the same level pretest to posttest.

Based on our examination of the standardization program design and the properties of the resulting norms we believe that the claims made by the test developers regarding the standard score scale are too strong. Our major concern is that the norming and scaling programs were done with a cross-sectional rather than a longitudinal design. Other concerns are: (1) the procedure to derive equivalent scores across levels of the MAT was conducted without regard to grade level and (2) only fall test data were used to develop the standard score scale.

The standardization and equivalencing programs were conducted with a design that did not require that a student be tested twice either on different levels of the test at the same point in time or on the same level of the test at different points in time. For the overlapping

program used to derive equivalent raw scores across test levels, for example, several different levels of the MAT were administered at the same grade level, but to different students. Scores from the Otis-Lennon Mental Ability Test were used to match groups of students receiving different levels of the MAT. Raw scores on the levels of the test were then linked by means of an equipercentile assumption that students would remain at the same percentile rank within a specified grade combination across levels of the MAT. The standard score scales were then developed using the Thurstone Absolute Scaling Method. The basic assumption for this method is that there is an underlying scale and transformations of raw scores to this scale are such that the transformed scale scores have a normal distribution at each grade level.

The validity of the standard score scale as used in the norm-referenced analysis depends on the soundness of the assumptions underlying the standardization procedure. For example, pooling distributions across grade levels as part of the procedure to equate levels of the test is valid as long as one can assume that the equivalent raw score relationship among levels of the test does not depend on grade level. The cross-sectional approach to developing the standard score scale is valid to the extent that students' educational experiences are stable over time. For example, if children in the eighth grade MAT norm group had educational experiences in the elementary grades that were much different from the experiences of children in the lower grades, the equipercentile growth curves implicit in the MAT norms would not be the same as those actually followed by either group.

Some empirical studies and some properties of the norms give evidence of poor articulation between levels of the MAT battery. Barker and Pelavin (1976), for example, examined the articulation between the various levels of the MAT test battery by testing disadvantaged children twice within seven days on different levels of the MAT. Because no large change in the childrens' achievement could be expected in a seven-day period, on the average a child should get the same standard score both times, even if he took a different test each time. They found that there is only weak evidence that disadvantaged children get the same score both times. On this evidence, they conclude that evaluations that are predominantly concerned with students who are educationally disadvantaged should base their evaluation on something other than standardized test scores.

Other evidence regarding the possible invalidity of the MAT norms comes from plots of standard score growth under the equipercentile assumption. Figures 2.1 and 2.2, for example, show selected equipercentile standard score growth curves for the Total Reading and Total Math subscales of the MAT. In both figures, the growth curves change abruptly between the first and second and seventh and eighth grades. Presumably, the "sampling error" in these graphs is small, so we should regard the changes as real (whether the changes are real or not, the norm-referenced analysis treats them as real). Is this how achievement scores change, or are the fluctuations due to the cross-sectional design of the standardization program? Whatever the fluctuations are, they

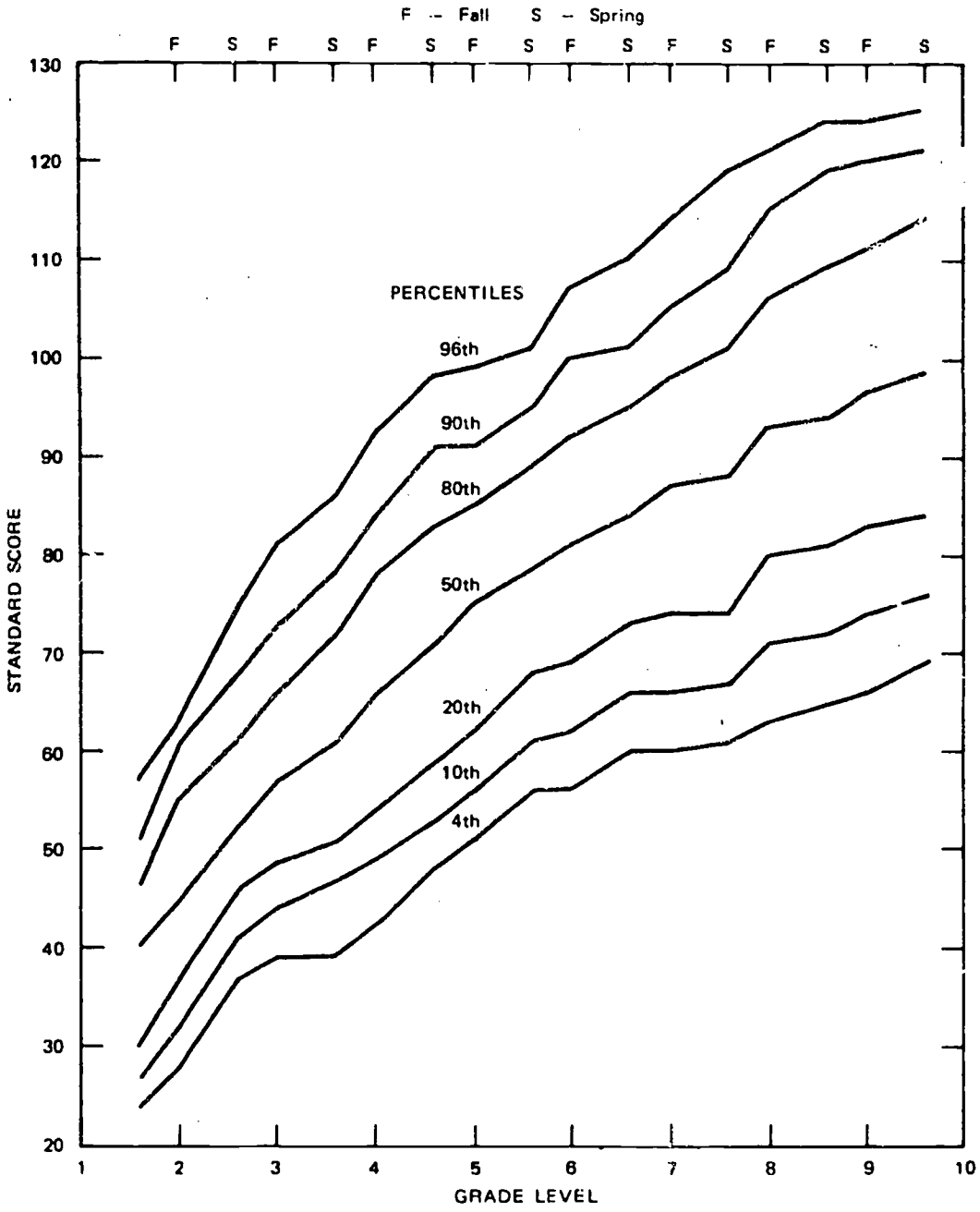


Figure 2.1 STANDARD SCORES BY GRADE LEVEL FOR SELECTED PERCENTILE RANKS: MAT TOTAL READING

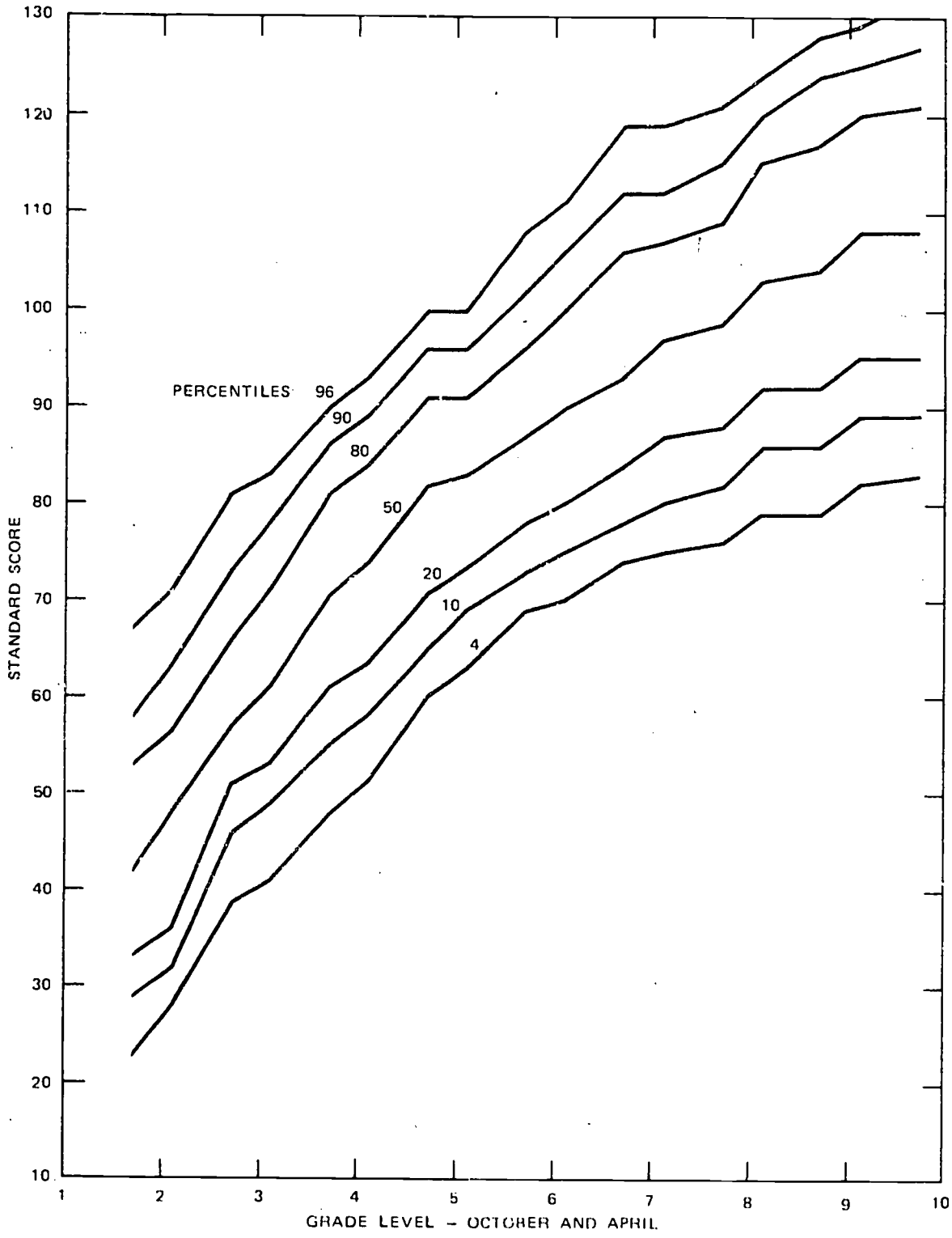


FIGURE 2.2 STANDARD SCORES BY GRADE LEVEL FOR SELECTED PERCENTILE RANKS: MAT TOTAL MATH

are not uniform across percentiles. For example, in the Total Reading curves, some interesting dips between the fourth and sixth grades for the curves greater than the 80th percentile do not appear on the lower equipercentile curves. A curious feature of the Reading score growth in the seventh grade is that for children of the 50th percentile or less no growth is expected until the summer, at which time the curves fairly shoot up on the standard score scale.

On the whole, "summer growth" for Total Reading is about as large as the "school year growth." This means that the growth during May and early June, plus the growth in September and early October, is nearly equal to the growth in the seven months of instruction between October and the following May. If there is not an abrupt change in the learning rate in May or October, we must conclude that there is significant growth in most MAT relevant skills when there is no school; or, put in another way, the MAT measures skills that grow about as fast out of school as in.

Table 2.1 shows the average of the ratios of spring-to-fall growth to the following fall-to-spring growth for Total Reading and Total Math for selected percentiles. The table shows that there is great variation in the average summer growth as a function of percentile, with the center percentiles being most subject to it. On the average, about 70% more growth is found in May, June, and September than in October through April at the 50th percentile for Total Reading.

Because the MAT norms are cross sectional, we do not know whether these findings represent facts or artifacts. The Coleman (1966) report and Jencks (1972) study have both shown that school does not influence standard reading scores very much. Such findings seem to be built into the norm tables, but we do not know whether the findings are based on the growth of skills measured by the MAT, on poor test linkages, or on the cross sectional design of the norming procedure.

In summary we have discussed two very basic issues regarding the norm-referenced analysis. One is the question of whether standardized tests are appropriate at all for evaluation purposes. At one level, this is a matter of policy. Judgments regarding the soundness of such policy and regarding its implications are related more to the goals and objectives of the evaluator rather than the program being evaluated. Use of standardized tests in evaluation is appealing because it is a conventional, widely-accepted practice. However, all too often the criterion of gains on a standardized test is adopted as the sole evidence of program impact independent of the specific objectives and goals of the educational program being evaluated. In our opinion, use of standard achievement tests in program evaluation may be considered reasonable to the extent that the evaluator's goals are clearly and deliberately defined in terms of the test's content.

Table 2.1

AVERAGE OF THE RATIOS OF SPRING-TO-FALL GROWTH TO THE FOLLOWING FALL-TO-SPRING GROWTH FOR TOTAL READING AND TOTAL MATH FOR SELECTED PERCENTILES

Percentile	Number of Ratios with Nonzero Denominators		Average Ratio for Total Math	Average Ratio for Total Reading
	Math	Reading		
1	9	7	.106	.250
4	8	7	1.395	.454
6	7	8	1.519	.787
10	7	8	1.068	.988
20	7	7	1.047	1.661
50	7	8	1.242	1.724
80	9	8	1.203	1.025
90	9	8	.775	1.473
96	9	8	.434	.833
99	9	4	.456	.872

In addition to the issue of the soundness or using standardized tests at all is the question of the validity of the norms, especially as they are used in the evaluation of programs that are targeted for children who are very different from the norm group. Some aspects of the standardization procedure, in particular the cross-sectional design, call into question whether the norms are adequate for predicting longitudinal gains even for the norm groups used in the standardization program. Some peculiar properties of the curves describing the predicted gains in standard score under the equipercentile assumption also need to be considered in deciding whether a norm-referenced analysis is a reasonable approach to evaluation. In the next section, the equipercentile assumption of normal growth is examined in more detail.

III THE EQUIPERCENTILE NORMAL GROWTH ASSUMPTION

To derive the expected posttest score given the pretest score, the norm-referenced procedure prescribes an equipercentile growth assumption that, all other things equal, the norm percentile score of a child, class, or site will, on the average, stay the same between the pre- and posttest.

If $g(\cdot)$ = the transformation from standard score to percentile score on the pretest,

$f(\cdot)$ the transformation from standard score to percentile score on the posttest,

then, according to the equipercentile model of normal growth, the expected posttest standard score given the pretest standard score x is $f^{-1}[g(x)]$. The functions f and g are derived on the basis of the standardization program conducted by the test developers. The function $f^{-1}(g(x))$ will be called the equipercentile growth curve.

The children included in the standardization program are generally selected to be representative of the entire school population. However, children in educational programs such as the PIP projects are quite different from the entire school population with respect to demographic characteristics, socioeconomic characteristics, and academic achievement. They are more likely to be members of a minority; they tend to be from low-income families; and they tend to have low pretest scores on standardized tests such as the MAT. More significantly for the norm-referenced procedure, evidence from the first-year PIP evaluation (Stearns, 1975) and from other evaluations (e.g., Coleman et al., 1966; Mayeske and Beaton, 1975; Armor et al., 1976) indicate that they tend to be pupils who lose ground over time, relative to the norm population. That is, as such a child progresses through school his percentile rank relative to the norm group declines rather than remaining the same.

If this is the case, then the equipercentile assumption would produce estimates of the expected posttest scores that were too high. This could lead to false conclusions that a program was not effective in raising scores on standardized tests above the expected level when in fact it had a salutary effect.

Method of Approach

The study of the equal percentile assumption consisted of examination of several large-scale data bases containing longitudinal MAT test data on children who would ordinarily qualify for educational programs exemplified by the PIP projects. Some analytic work was also done regarding the longitudinal nature of the norms.

The data bases were taken from the Follow Through Project (FT) evaluation obtained from SRI, the Compensatory Reading (CR) program

evaluation obtained from the Education Testing Service and a subset of the MAT norming data obtained from Psychological Corporation. From the Follow Through evaluation, data on a subset of children in the comparison group, called Non-Follow Through (NFT) were examined. These were children who had entered kindergarten in fall 1971 and had been tested in at least two of three subsequent spring test periods (1973, 1974, or 1975) when the MAT had been administered.

From the CR evaluation, three groups were of particular interest, those children in compensatory reading programs who were participating in the federal school lunch program (CR/SL), those children in compensatory reading programs and not participating in the federal lunch program (CR/NSL), and those children who were in schools that had no compensatory education program and who were participating in the federal school lunch program (NCR/SL). Participation in the federal school lunch program was the only available indicator of socioeconomic status. About 75% of the CR/SL children were in schools where compensatory reading programs were funded to some extent by Title I and about 58% of the CR/NSL children were in schools where the compensatory reading programs were funded to some extent by Title I. These two groups, then, would consist largely of children similar to those for whom the PIPs are targeted.

We had planned initially to focus on the NCR/SL group because it consisted of children who were not in compensatory reading programs and who appeared to be from low-income families. However, the very fact that the schools did not have a compensatory reading program would give evidence that the NCR/SL group is unique as a "disadvantaged" group. The demographic profile of this group relative to the CR/SL group also indicated that this group might be unique. Therefore, we included the CR/SL and CR/NSL groups in the study as well. Interpretation of data from these two groups needs to take into account the effects of being in a compensatory reading program. Nevertheless, we thought it would be of interest to examine how these groups operate relative to the equipercentile model. Even this group might have shown declines in percentile from pre- to posttest.

The subset of the MAT norm data consisted of longitudinal fall and spring test scores for those children tested at both times. These data were used to examine the relationship between the gains predicted from the cross-sectional standardization design and those observed from the longitudinal subgroup. These data were initially analyzed by Dr. Michael Beck of Psychological Corporation, who reported the results in a paper presented at the 1975 Convention of the National Council on Measurement in Education (Beck, 1975).

Neither the FT evaluation nor the CR evaluation were designed so that the groups selected for this study are representative of the entire U.S. school population or of any particular subpopulation such as children who would qualify for Title I programs. Table 3.1 shows the number of pupils in each group with both pre- and posttests and the percent of children by geographic region, city size, and socioeconomic indicators. Grade 4 was selected for the CR evaluation group as representative of all three grade levels.

Table 3.1

DESCRIPTION OF THE DEMOGRAPHIC AND SOCIOECONOMIC CHARACTERISTICS OF THE GROUPS OF PUPILS USED IN THE PIP CRITERION STUDY

Group	Geographic Region				City Size	Percent Minority	Socioeconomic Status
	Northeast	North Cen	South	West			
Non-Follow Through (n=2095)	49%	26%	13%	12%	Approximately 49% in cities with a population of 200,000 or more; 21% in cities between 50,000 and 200,000 and 30% in cities less than 50,000 or rural areas	62%	Median household income about $5,000 (1971)
Compensatory Reading Study (4th Grade) CR/SL (n=3038)	9%	22%	47%	21%	Approximately 13% in cities with a population of 200,000 or more; 45% in cities less than 200,000; 42% in rural areas	48%	100% in federal school lunch program
CR/NSL (n=2150)	25%	29%	25%	21%	Approximately 14% in cities with a population of 200,000 or more; 58% in cities less than 200,000; 27% in rural areas	20%	0% in federal school lunch program
NCR/SL (n=534)	19%	29%	56%	14%	Approximately 4% in cities with population of 200,000 or more; 65% in cities less than 200,000; 32% in rural areas	25%	100% in federal school lunch program
MAT Standardization Sample	23%	28%	27%	22%	Approximately 21% in cities of 250,000 or more; 48% in cities less than 250,000; 30% in rural areas	About 10%	Median family income $5,500 (1960)

The NFT and CR groups are substantially different in composition and differ substantially from the MAT standardization sample. For example, relative to the norm group, the Non-Follow Through group appears to be underrepresentative of the South and moderate size cities and over-representative of large cities and the Northeast. On the other hand, the CR/SL sample and NCR/SL sample are underrepresentative of the Northeast and large cities and overrepresentative of the South and moderate size places. Both the Non-Follow Through and Compensatory Reading groups have a much higher percentage of minority children than the MAT standardization group. The median household income for the Non-Follow Through group appears to be only moderately less than the median family income for the MAT standardization sample, but the MAT estimate was based on the 1960 U.S. Census and the NFT estimate was based on data collected in 1971.

Table 3.2 indicates the grade levels included in each data set and the levels of the MAT administered at pretest and posttest. The NFT group was tested successively in the spring in grades 1, 2, and 3. (This group was also tested on the Wide Range Achievement Test in the fall of their kindergarten year.) The groups in the CR study were tested in the fall and spring of a single school year. Three grade levels--2, 4, and 6--were included in this study. The MAT norm subpopulation includes children in grades 2 through 8 who were tested in both the fall and the spring of a single school year.

The battery level used to test the NFT group corresponded to the level specified by the MAT developers for each grade and time period. This level also corresponds to the one administered in the MAT standardization program. For the CR evaluation, children were tested one level below that used in the MAT standardization program in the spring of second grade and the spring and fall of sixth grade. For the PIP evaluation, children were tested out of level in the fall in grades 5, 6, and 7 in the 1975-76 evaluation.

The standard score metric on a standardized test such as the MAT can theoretically be used to perform the norm-referenced procedure even when testing is conducted out of level, because the conversion of standard score to percentile is supposedly independent of the level of test administered. However, a number of recent studies (Barker and Pelavin, 1975; Pelavin and Barker, 1976) have indicated that some levels of the MAT may not have been adequately articulated so that standard scores on two different levels of the MAT may not be equivalent. This may mean that the equal percentile assumption may be valid with respect to some combination of levels of the test and not valid with respect to others.

Also significant is that the time of testing and the administration of the tests were not uniform across and within data bases. For the MAT standardization group, fall testing was in October and spring testing was in April. Most testing of Non-Follow Through pupils took place between mid-April and mid-May of each year with some testing occurring as late as June. For the CR evaluation, testing was scheduled for the third full week after the opening of school in the fall and for the fifth week prior to the end of the school year in the spring (Trismen et al., 1975). No

Table 3.2

DESCRIPTION OF THE TESTS ADMINISTERED

	Grade Levels	Test Periods		Battery Level by Grade Level/Pre- Post*								
				1	2	3	4	5	6	7	8	9
MAT Norm Subpopulation	2-8	Fall, spring; longitudinal within grade	Pre:	-	PI	PII	E	I	I	A	A	
			Post:	-	PII	E	E	I	I	A	A	
Follow Through: Cohort III (SRI)	1,2,3	Spring; longitudinal across grades		PI	PII	E	-	-	-	-	-	
Compensatory Reading Program Study (ETS)	2,4,6	Fall, spring; longitudinal within grade	Pre:	-	PI	-	E	-	E†	-	-	
			Post:	-	PI†	-	E	-	E†	-	-	
PIP Evaluation (1975-1976)	1-8	Fall, spring; longitudinal within grade	Pre:	Pr	PI	PII	E	E†	E†	I†	A	A
			Post:	PI	PII	E	E	I	I	A	A	A

*Pr = Primer
PI = Primary I
PII = Primary II
E = Elementary
I = Intermediate
A = Advanced

†One level below that used in standardization program

data were available from ETS on the precise dates on which testing took place. However, if the typical school year beings in the first week of September and continues into mid-June, then fall testing would have taken place in late September or early October and spring testing would have taken place between late April and mid-May.

The examination of the FT and CR data consisted of generating summary statistics on pre- and posttest performance. Summary statistics included means and standard deviations of standard scores, the percentile ranks of the mean, the mean and standard deviation of the change in percentile, and the percent of children with a loss in percentile rank between pre- and posttests. The relationship between the pre- and posttest standard scores found in the data was also described by plots of so-called empirical growth curves. An empirical growth curve is the function that describes the mean posttest standard score given the pretest standard score. It can be compared with the equipercentile curve to assess the fit of the equipercentile curve to the data.

Demographic and socioeconomic factors related to change in percentile rank were examined by means of cross-tabulations of the distribution of change in percentile by demographic and socioeconomic characteristics, tabulations of summary statistics, and regression analysis.

Summary statistics and empirical growth curves were also generated for the MAT longitudinal data to compare the longitudinal empirical curves with the cross-sectional equipercentile curves. The relationship between the longitudinal and cross-sectional approach to deriving the expected posttest scores was also examined analytically.

Relationship Between the Cross-Sectional and Longitudinal Norms

The equipercentile growth curve is based on the standard score to percentile conversions printed in the MAT Teacher's Handbooks (Durost et al., 1971). This curve is based on cross-sectional norms in that even though many children were included in both the fall and spring standardization program, the conversions from standard score to percentile were calculated separately for fall and spring. Analytically, the relationship between the cross-sectional and longitudinal equipercentile growth curve is quite simple. If we assume that the pretest (x) and posttest (y) standard scores on a particular test have a bivariate normal distribution with means, μ_x and μ_y, standard deviations, σ_x and σ_y, and correlation coefficient, ρ, then the functions f and g may be expressed approximately as follows:

$$g(x) = 100. \; \Phi\left[\frac{x-\mu x}{\sigma_x}\right]$$

$$f(y) = 100. \; \Phi\left[\frac{y-\mu y}{\sigma_y}\right]$$

where $\Phi(x)$ is the cumulative distribution function for the normal distribution with a mean of 0 and a standard deviation of 1. The equipercentile growth curve, $f^{-1}[g(x)]$, is then:

$$f^{-1}[g(x)] = \mu_y + \frac{\sigma_y}{\sigma_x}(x-\mu_x) \qquad (1)$$

On the other hand, the expectation of the posttest score given the pretest score would appear to be an alternative to the equipercentile growth curve if the standard scores have a bivariate normal distribution. In that case, it is well known that:

$$E(Y|X = x) = \mu_y + \frac{\sigma_y}{\sigma_x}\rho(x-\mu_x) \qquad (2)$$

The correlation coefficient ρ, as estimated from the MAT longitudinal data, ranges between about .75 and .91. This would indicate that the equipercentile growth curve may be too low for pretest scores below the norm mean and too high for pretest scores above the norm mean under the assumption that one is sampling from the norm population and the pretests and posttests have a bivariate normal distribution.

The summary statistics for the MAT longitudinal data are displayed in Table 3.3 for Total Math and Total Reading. A norm-referenced analysis was conducted on the longitudinal data to assess the change in standard score relative to the equipercentile model of normal growth. The change in standard scores on Total Reading appears to be greater than expected for grades 2, 3, 5, 7, and 8. For Total Math, gains in grades 2, 5, 7, and 8 again appear higher than expected, and the gains for third grade appear to be less than what was expected. For both Total Reading and Total Math, grades 4 and 6 had gains that were close to what was expected. With the large sample size at each grade level, only small differences between expected and observed values could give t values that appear to be significant, however. Also, the conversion from standard score to raw score is extremely sensitive to the standard score around the 50th percentile. As a result, a difference of about a half a standard score point can translate to a difference of 4 percentile ranks, such as that found in second grade Total Reading.

Nevertheless, the empirical growth curves do indicate that the equipercentile growth curve tends to underestimate expected posttest scores for extremely low pretest socres and tends to overestimate expected posttest socres for extremely high pretest scores across grade levels and tests. Figure 3.1 through 3.8, for example, show plots of the equipercentile growth curve and the empirical growth curve for grades 2, 4, 6, and 8 on Total Reading and Total Math. The solid line in each figure represents the equipercentile growth curve and the dots represent the empirical growth curve. The number of pupils at each pretest standard score point varies. Because the number of pupils at the extremes may be extremely low, perhaps one or two in many cases, it is the patterns within each grade level and test and across grade levels and tests that are of interest.

Table 3.3

SUMMARY OF TRENDS IN PERCENTILE GROWTH FOR THE MAT STANDARDIZATION SUBGROUP

Test	Grade	N	Pretest Standard Score Mean	Pretest Standard Score S.D.	Pretest Percentile*	Posttest Standard Score Mean	Posttest Standard Score S.D.	Posttest Percentile*	Norm-Referenced Statistics: Expected Posttest Standard Score	t	Gain Statistics: Change in Percentile of Mean
Total Reading	2	2854	46.4	10.3	57.6	54.9	10.8	61.4	54.3	2.80	+3.8
	3	1638	58.4	11.5	53.6	62.9	12.9	57.6	61.9	3.09	+4.0
	4	2175	66.8	13.2	53.2	71.7	14.1	52.8	71.8	- .32	- .5
	5	2624	75.3	13.2	51.2	79.5	12.9	53.0	78.8	2.62	+1.8
	6	2732	82.9	14.5	55.6	85.6	14.0	56.4	85.4	.70	+ .8
	7	2154	86.8	14.6	49.2	89.4	16.1	54.8	87.8	4.62	+5.6
	8	2030	93.4	15.5	51.2	95.7	16.9	53.4	94.6	2.92	+2.2
Total Math	2	2853	48.8	11.9	51.6	59.7	11.2	60.8	57.4	9.58	+9.2
	3	1609	62.9	11.5	55.8	71.4	12.2	53.6	72.0	-1.89	-2.2
	4	2130	74.1	11.0	50.6	82.0	12.5	50.0	82.3	-1.11	- .6
	5	2595	83.9	10.0	53.6	88.7	11.0	56.8	87.9	3.66	+3.2
	6	2699	92.1	11.5	58.4	96.0	12.9	58.0	96.1	- .40	- .4
	7	2120	97.8	12.3	53.2	100.3	13.1	55.2	99.6	2.43	+2.0
	8	2002	103.7	12.7	51.4	105.9	13.7	53.8	104.7	3.91	+2.4

*Transform of mean standard score using linear interpolation

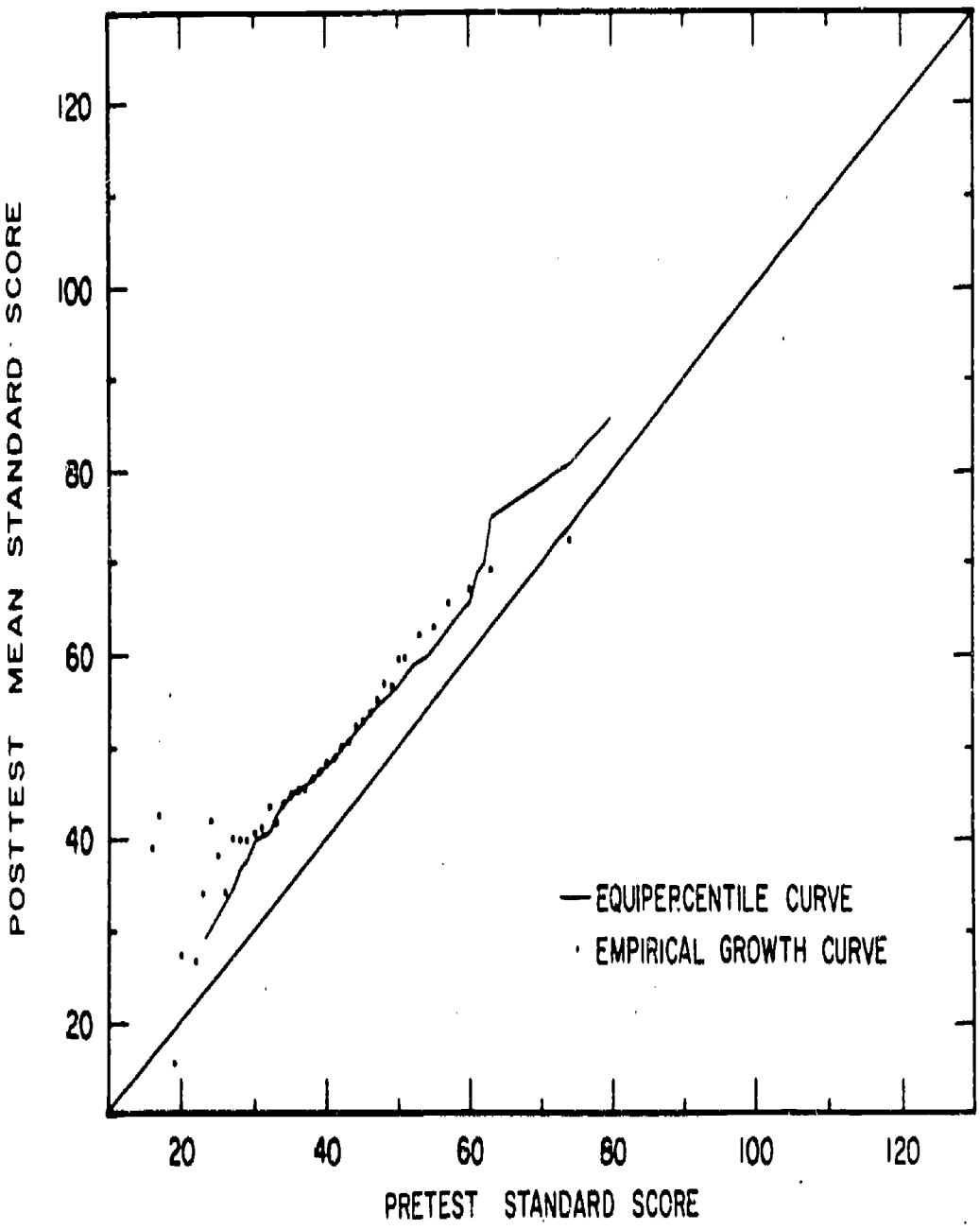


FIGURE 3.1 EMPIRICAL GROWTH CURVE FOR THE SECOND GRADE MAT LONGITUDINAL GROUP ON MAT TOTAL READING: FALL PRETEST AND SPRING POSTTEST

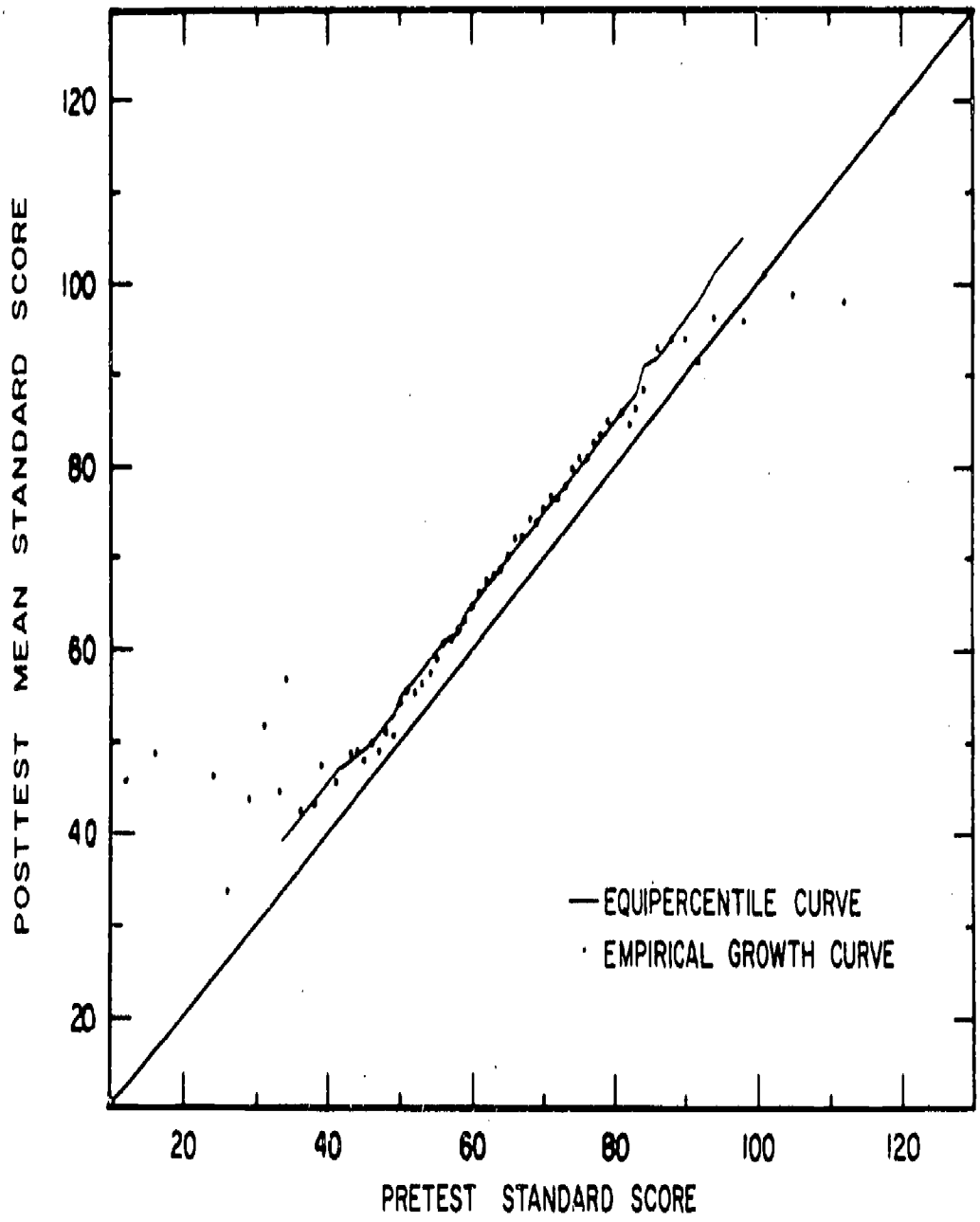


FIGURE 3.2 EMPIRICAL GROWTH CURVE FOR THE FOURTH GRADE MAT LONGITUDINAL GROUP ON MAT TOTAL READING: FALL PRETEST AND SPRING POSTTEST

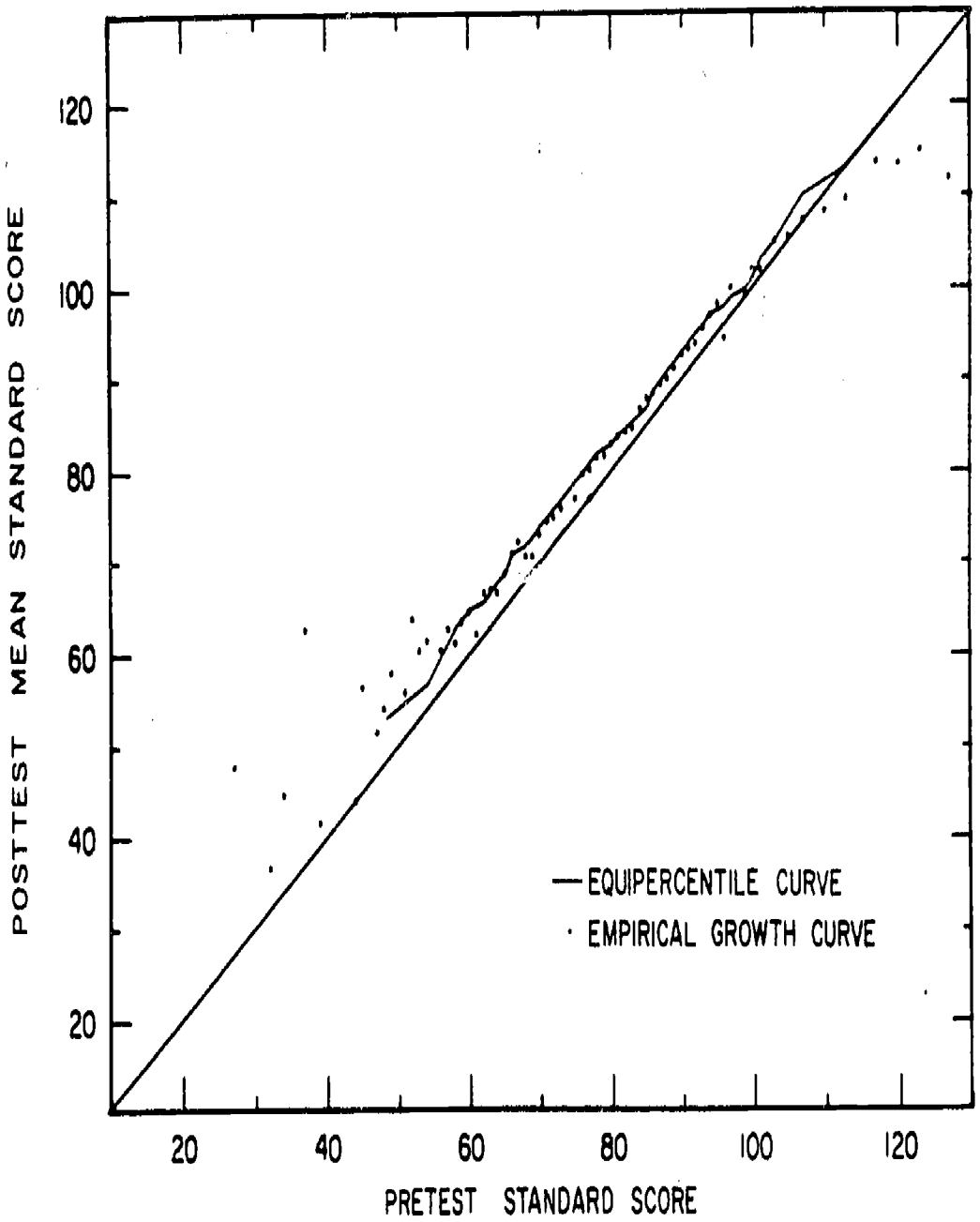


FIGURE 3.3 EMPIRICAL GROWTH CURVE FOR THE SIXTH GRADE MAT LONGITUDINAL GROUP ON MAT TOTAL READING: FALL PRETEST AND SPRING POSTTEST

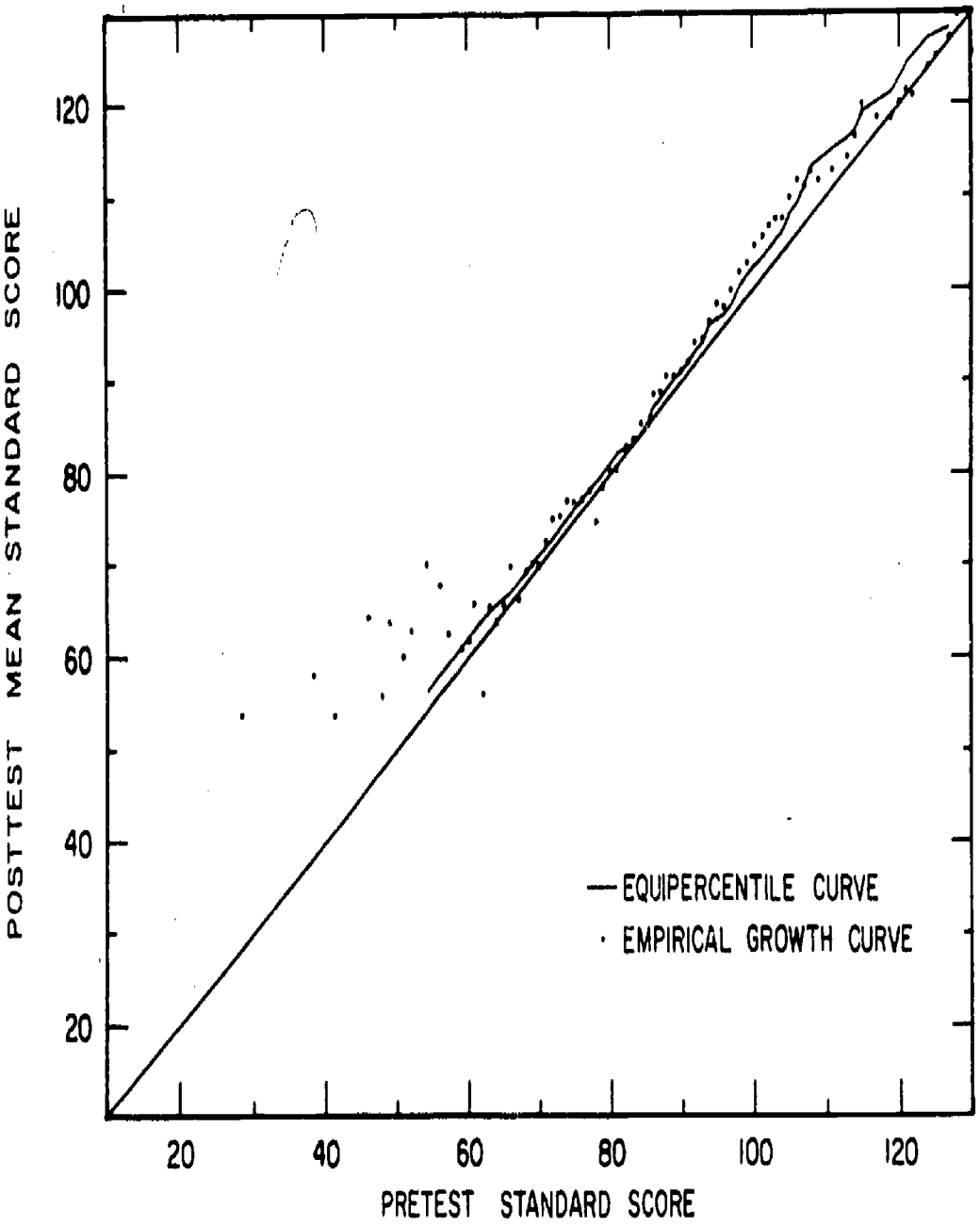


FIGURE 3.4 EMPIRICAL GROWTH CURVE FOR THE EIGHTH GRADE MAT LONGITUDINAL GROUP ON MAT TOTAL READING: FALL PRETEST AND SPRING POSTTEST

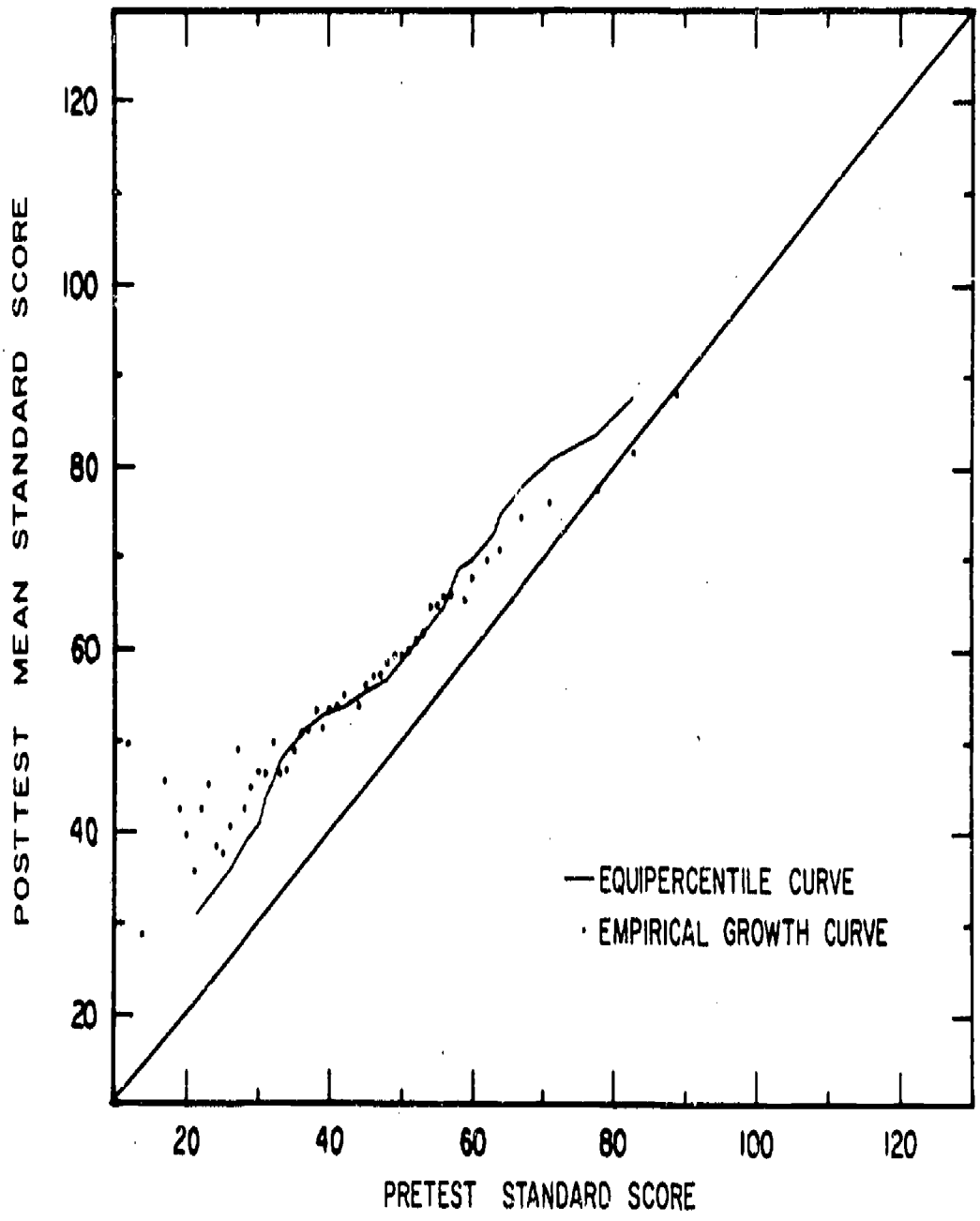


FIGURE 3.5 EMPIRICAL GROWTH CURVE FOR THE SECOND GRADE MAT LONGITUDINAL GROUP ON MAT TOTAL MATH: FALL PRETEST AND SPRING POSTTEST

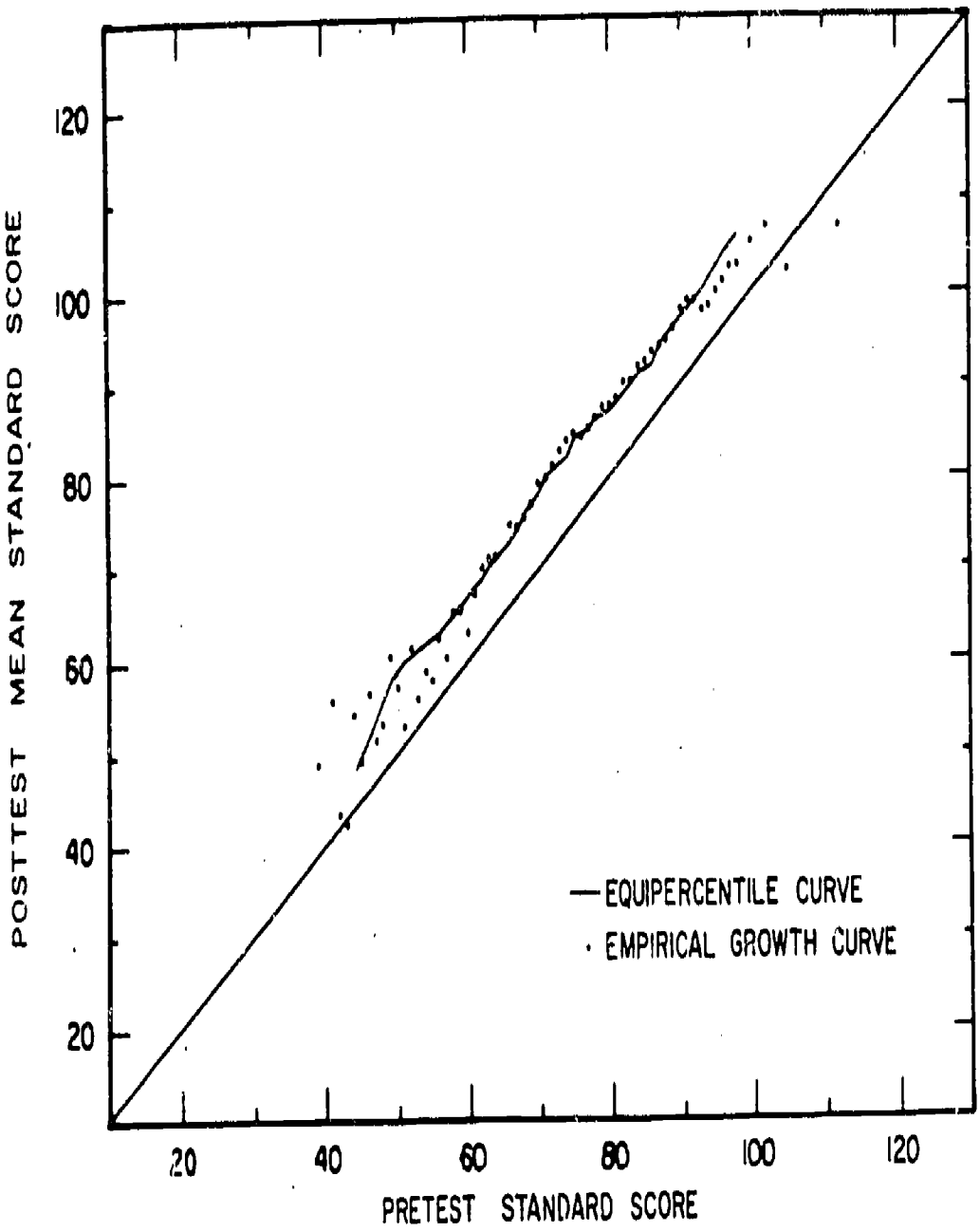


FIGURE 3.6 EMPIRICAL GROWTH CURVE FOR THE FOURTH GRADE MAT LONGITUDINAL GROUP ON MAT TOTAL MATH: FALL PRETEST AND SPRING POSTTEST

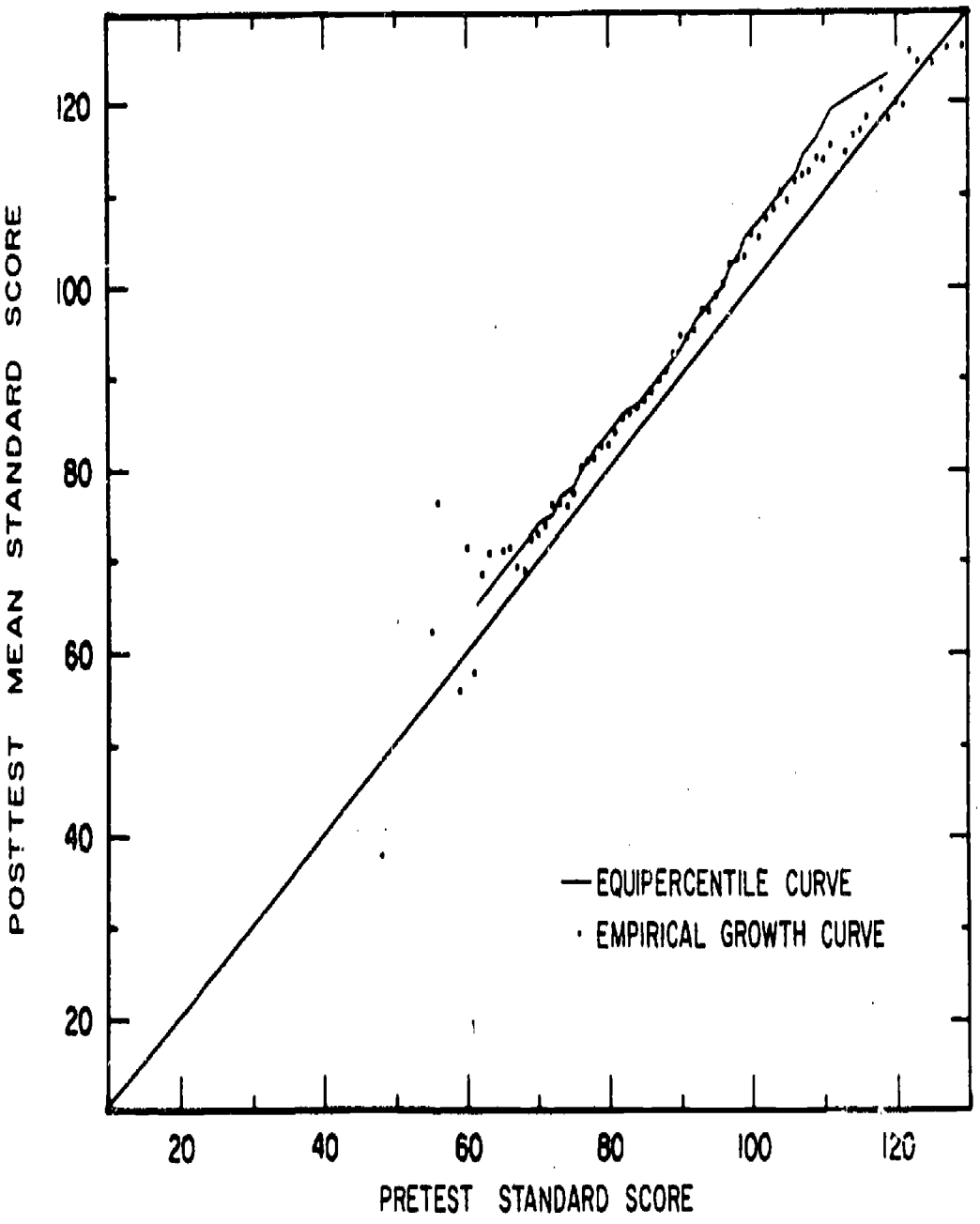


FIGURE 3.7 EMPIRICAL GROWTH CURVE FOR THE SIXTH GRADE MAT LONGITUDINAL GROUP ON MAT TOTAL MATH: FALL PRETEST AND SPRING POSTTEST

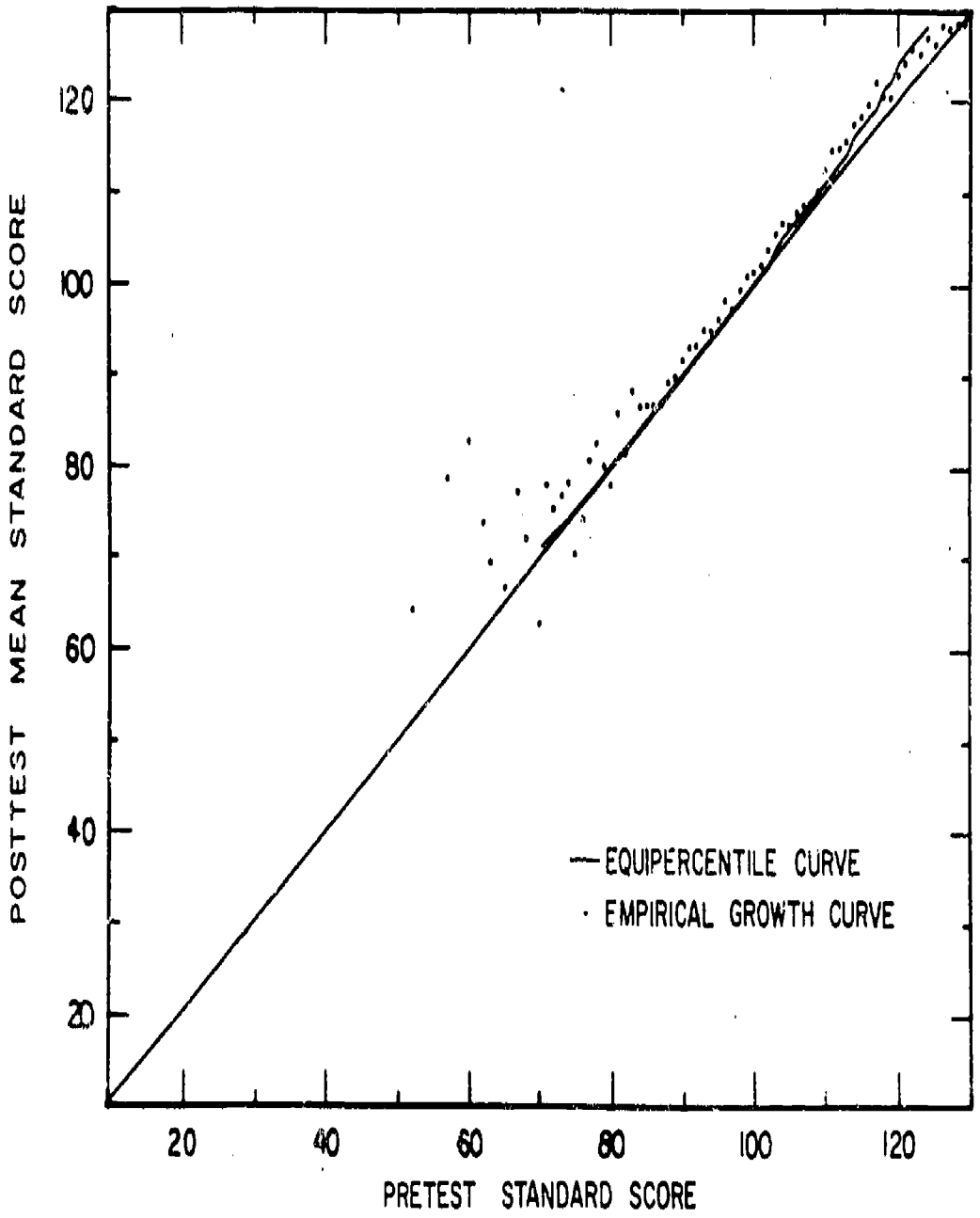


FIGURE 3.8 EMPIRICAL GROWTH CURVE FOR THE EIGHTH GRADE MAT LONGITUDINAL GROUP ON MAT TOTAL MATH: FALL PRETEST AND SPRING POSTTEST

In almost every case illustrated, the points in the empirical curve tend to lie above the equipercentile curve for low pretest scores and they tend to lie below the equipercentile curve for extremely high pretest scores. The two curves do track each other quite closely in the midrange of pretest scores. For the eighth grade, however, as indicated in the summary tables, the equipercentile curve consistently lies below the empirical growth curve.

The relationship between the equipercentile and empirical curves that was predicted from the bivariate normal assumption is not consistently evident in the plots. The bivariate normal model, as indicated above, predicted that the empirical curve should lie above the equipercentile curve for pretest scores below the mean pretest score and should lie below the equipercentile curve in the region above the mean pretest score. With the exception of the extreme pretest scores this does not appear to be the case (e.g., Total Reading for Grade 2, Total Math for Grade 6). This could be caused by any of a number of factors--the high correlation between the pre- and posttests, the use of only a subset of the standardization population, or the finite range of possible standard scores.

The results using the longitudinal subset of the norm population, then, indicate that with the exception of extreme pretest scores, the equipercentile model appears to be adequate in predicting spring scores given fall scores within the same school year. For extremely low pretest scores, the posttest score that would be predicted from the equipercentile curve, or from an extension of the curve, appears to be much lower than observed empirically in a number of cases.

Results Regarding the FT and CR Data

The summary statistics describing the distribution of pretest and posttest standard scores and the changes in percentile rank are presented for Total Reading and Total Math in Table 3.4 for the NFT and CR groups. Pupils in the CR evaluation were administered only the reading portions of the MAT.

For the NFT group, there is a consistent pattern of declining percentiles from the spring of first grade to the spring of third grade for Total Reading with an average loss of just over six percentile points in each year. For Total Math, there was an average drop in the percentile rank of over 6 points between first and second grade corresponding to a drop in the normative percentile rank of the mean of almost 19 points. The changes between second grade and third grade were more consistent with the equal percentile assumption, with the mean percentile change being about zero and the change in the percentile of the mean standard score increasing by 2.6 points.

For the CR evaluation data, the gains in percentile rank for second graders are uniformly large, with the average gain for the CR/SL group being in excess of 9 percentile points. For fourth grade the gains are not as spectacular, but they are still substantially higher than would

Table 3.4

SUMMARY OF TRENDS IN PERCENTILE GROWTH FOR THE NON-FOLLOW THROUGH AND COMPENSATORY READING GROUPS

Data Base/Test	Initial Grade	Subgroup	Pre/Post Duration	N	Pretest Standard Score Mean	Pretest Standard Score S.D.	Pretest Percentile*	Posttest Standard Score Mean	Posttest Standard Score S.D.	Posttest Percentile*	Gain Statistics: Change in Percentile of Mean	Gain Statistics: Change of Percentile Mean	Gain Statistics: Change of Percentile S.D.	Gain Statistics: Percent with Percentile loss
Non-Follow Through														
Total Reading	1		Sp 1973/Sp 1974	7096	39.4	10.1	49.6	50.9	10.4	43.4	- 6.2	- 6.9	17.5	65%
	2		Sp 1974/Sp 1975	2043	51.3	10.7	45.2	57.7	12.6	38.8	- 6.4	- 2.4	13.9	55
	3		Sp 1973/Sp 1975	2368	39.5	10.3	50.0	57.4	12.7	37.6	-12.4	- 9.1	20.0	69
Total Math	1		Sp 1973/Sp 1974	2012	38.8	11.3	41.6	51.7	11.1	22.8	-18.8	- 6.3	18.8	60
	2		Sp 1974/Sp 1975	1934	52.0	11.3	24.0	63.3	12.9	26.6	+ 2.6	.0	15.5	47
	3		Sp 1973/Sp 1975	2254	38.9	11.2	41.8	63.3	12.7	26.6	-15.2	- 5.6	20.8	58
Compensatory Reading Study														
Total Reading	2	CR/SL	Fall/spring	2842	34.8	8.9	15.6	46.1	10.8	20.3	+ 4.7	+ 9.3	20.6	32
		CR/NSL	Fall/spring	2638	37.6	8.6	22.4	49.6	10.6	36.4	+14.0	+11.3	21.2	30
		NCR/SL	Fall/spring	608	41.9	10.6	37.6	55.5	11.9	54.0	+16.4	+11.5	22.0	31
	4	CR/SL	Fall/spring	3038	51.7	12.4	15.7	58.3	12.4	18.6	+ 2.9	+ 2.8	10.9	38
		CR/NSL	Fall/spring	2150	55.5	11.3	23.0	62.7	11.8	29.4	+ 6.4	+ 4.8	12.7	35
		NCR/SL	Fall/spring	55[illegible]	60.7	14.2	36.8	68.0	15.0	42.0	+ 5.2	+ 4.9	12.0	33
	6	CR/SL	Fall/spring	2568	61.4	13.2	9.2	66.2	13.9	10.2	+ 1.0	+ 1.7	11.5	36
		CR/NSL	Fall/spring	2140	70.1	14.1	22.2	74.9	14.9	22.9	+ .7	+ 2.6	15.7	41
		NCR/SL	Fall/spring	545	76.8	15.9	37.6	80.7	16.3	38.8	+ 1.2	+ 1.5	17.3	44

*Transform of mean standard score using linear interpolation

be expected on the basis of the equal percentile assumption. Finally, the gains for the sixth grade groups are still greater than expected, but substantially less than those found in second or fourth grades.

The dramatic decline in percentile ranks for the NFT group over the two-year period supports the straggler hypothesis--that in the absence of intervention programs, children targeted for compensatory education programs for the disadvantaged will lose ground relative to the norm group. The data from the CR evaluation, on the other hand, is not unequivocal in denying the straggler hypothesis. For one thing, children in the CR/SL and CR/NSL groups were participating in compensatory reading programs that in fact may have reversed a decline in percentiles. The NCR/SL group that appeared to have the promise of serving the role of comparison group had percentile gains that equalled or exceeded those of the two CR groups. Furthermore, the NCR/SL group has characteristics that would call into question the assumption that these children are typical of those that would be in compensatory programs. They are disproportionately from the South, from moderate size cities, and nonminority. Finally, none of the NCR/SL students, as contrasted with 44% of the CR/SL students, were in schools where the estimated percent of pupils from families receiving public assistance exceeded 25% of the school population.

Out-of-level testing probably does not explain the extraordinary gains in second grade. Pelavin and Barker (1976) have found that when pupils are tested within a short time period on both the Primary I and Primary II reading subtests that standard scores on Primary II tend to be higher than standard scores on Primary I. This would mean that the out-of-level testing in the spring may have in fact suppressed the gains in standard scores on Total Reading. One factor that might have contributed to the high gains are the extremely low pretest scores. However, the empirical growth curves presented below show gains greater than expected across a broad range of pretest scores. Another factor might be the time at which the tests were administered. It appears from the descriptions of the test administration that the children in the CR evaluation were tested a bit earlier than the standardization sample on the pretest and a bit later than the standardization sample on the posttest. It does not appear to be plausible, however, that the slight differences in scheduling tests could result in gains that were so much larger than expected.

Figures 3.9 through 3.12 present the empirical growth curves for NFT for Total Reading and Total Math. The four figures show a consistent pattern. For extremely low pretest scores, the empirical curve lies above the equipercentile curve. For the remainder of the range, the empirical curve is generally below the equipercentile curve. In both Total Math and Total Reading, the drop below the equipercentile line is greater between grade 1 and grade 2 than between grade 2 and grade 3.

Figures 3.13 through 3.18 present the empirical growth curves for the CR/SL and NCR/SL subgroups on Total Reading. The CR/NSL empirical growth curves, which are not shown, are similar to the CR/SL curves.

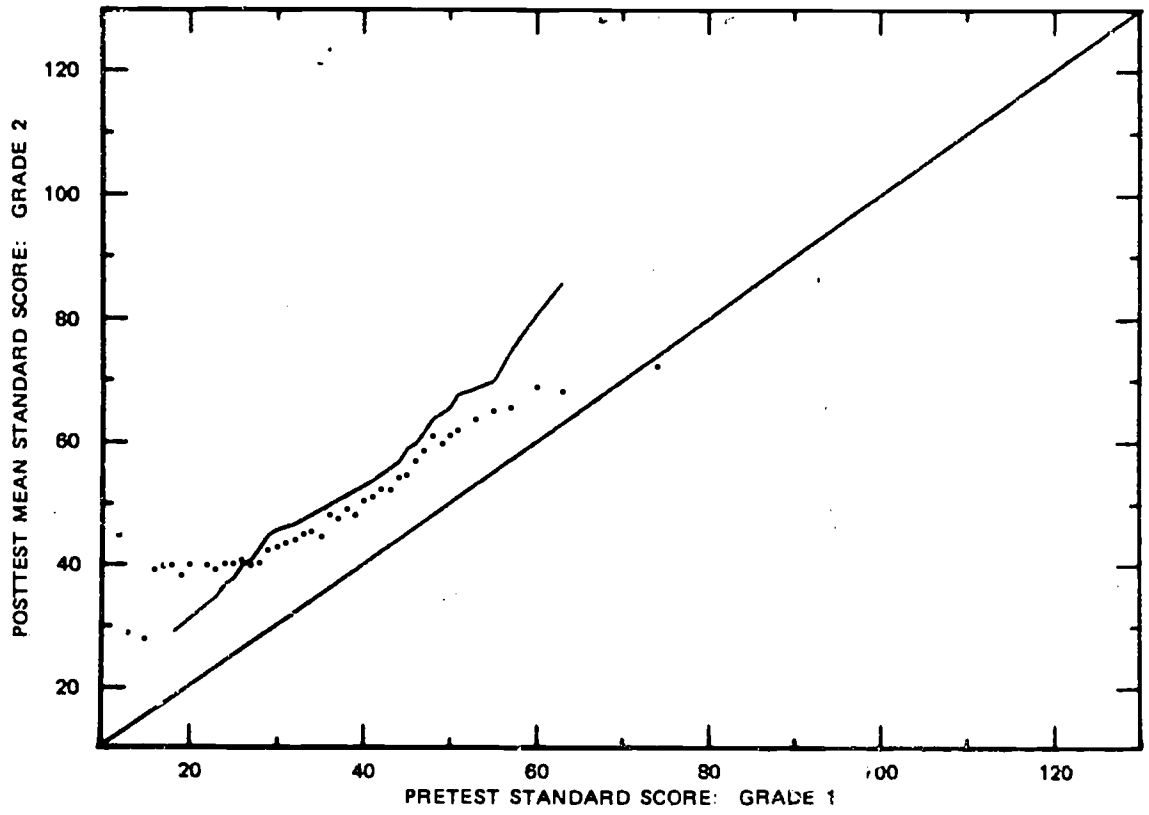


Figure 3.9 EMPIRICAL GROWTH CURVE OF THE NFT GROUP, FIRST TO SECOND GRADE, ON MAT TOTAL READING: SPING PRETEST AND POSTTEST

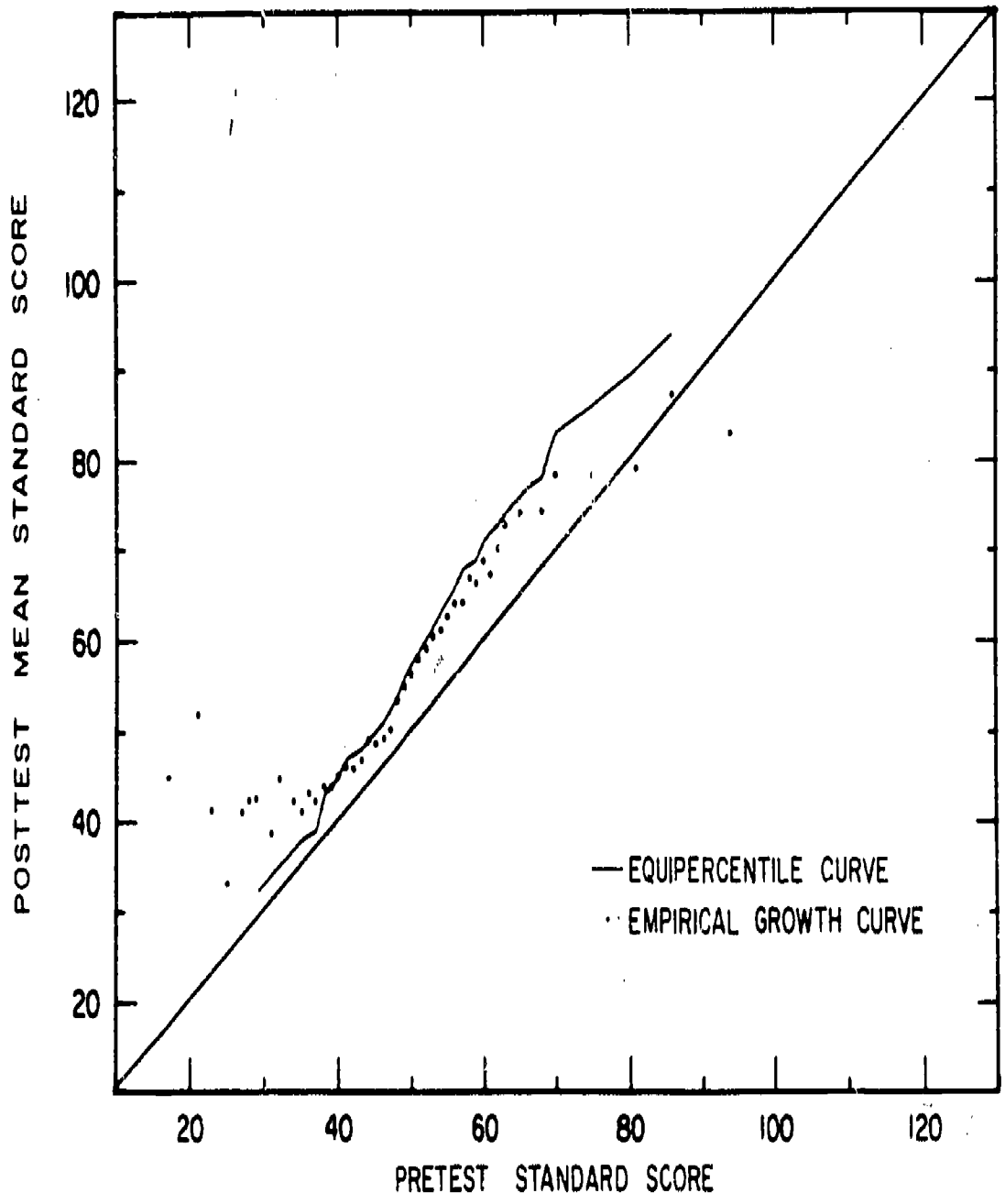


FIGURE 3.10 EMPIRICAL GROWTH CURVE FOR THE NFT GROUP, SECOND TO THIRD GRADE, ON MAT TOTAL READING, SPRING PRETEST AND POSTTEST

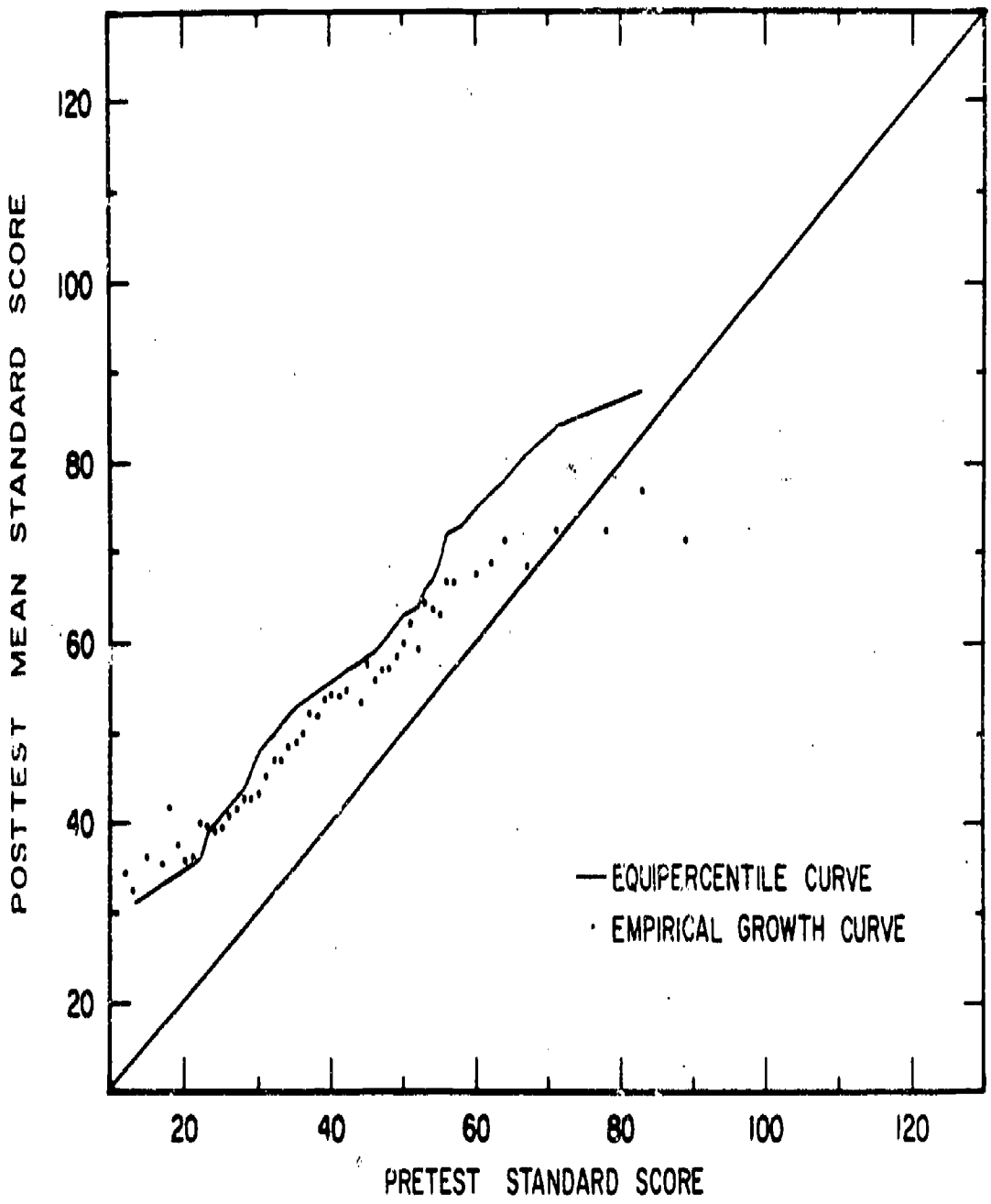


FIGURE 3.11 EMPIRICAL GROWTH CURVE FOR THE NFT GROUP, FIRST TO SECOND GRADE, ON MAT TOTAL MATH, SPRING PRETEST AND POSTTEST

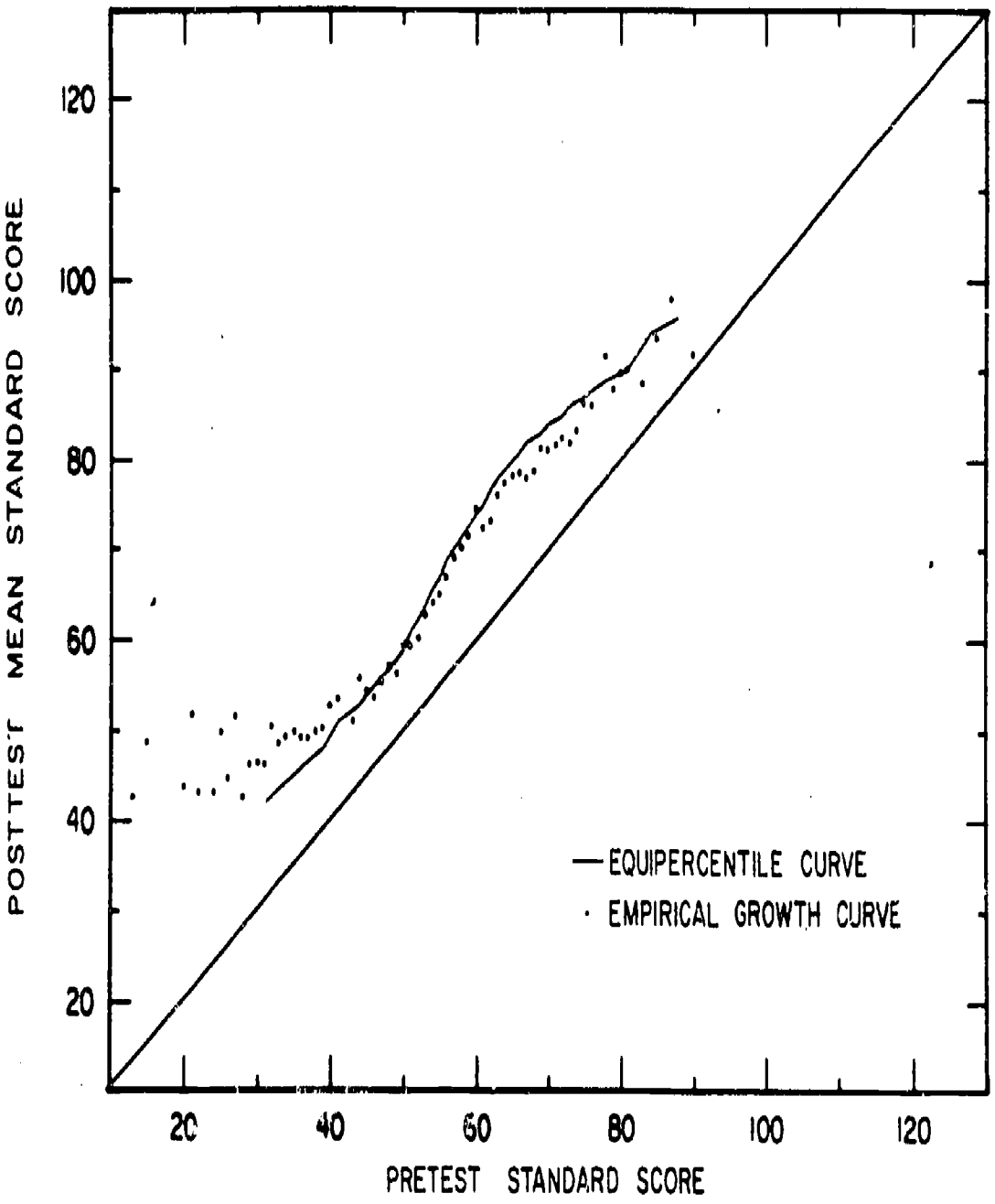


FIGURE 3.12 EMPIRICAL GROWTH CURVE FOR THE NFT GROUP, SECOND TO THIRD GRADE, ON MAT TOTAL MATH, SPRING PRETEST AND POSTTEST

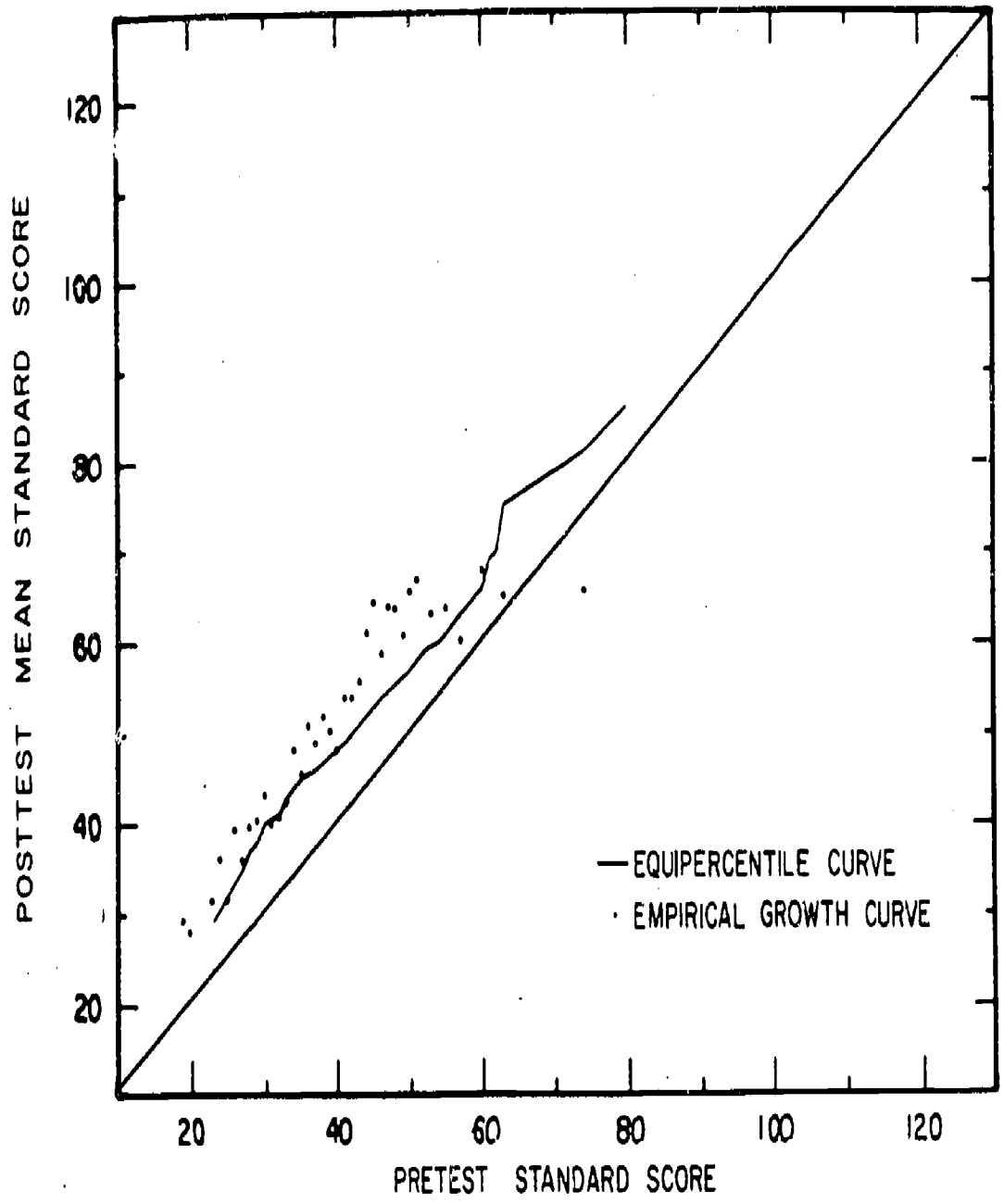


FIGURE 3.13 EMPIRICAL GROWTH CURVE FOR THE SECOND GRADE NCR/SL GROUP ON MAT TOTAL READING: FALL PRETEST AND SPRING POSTTEST

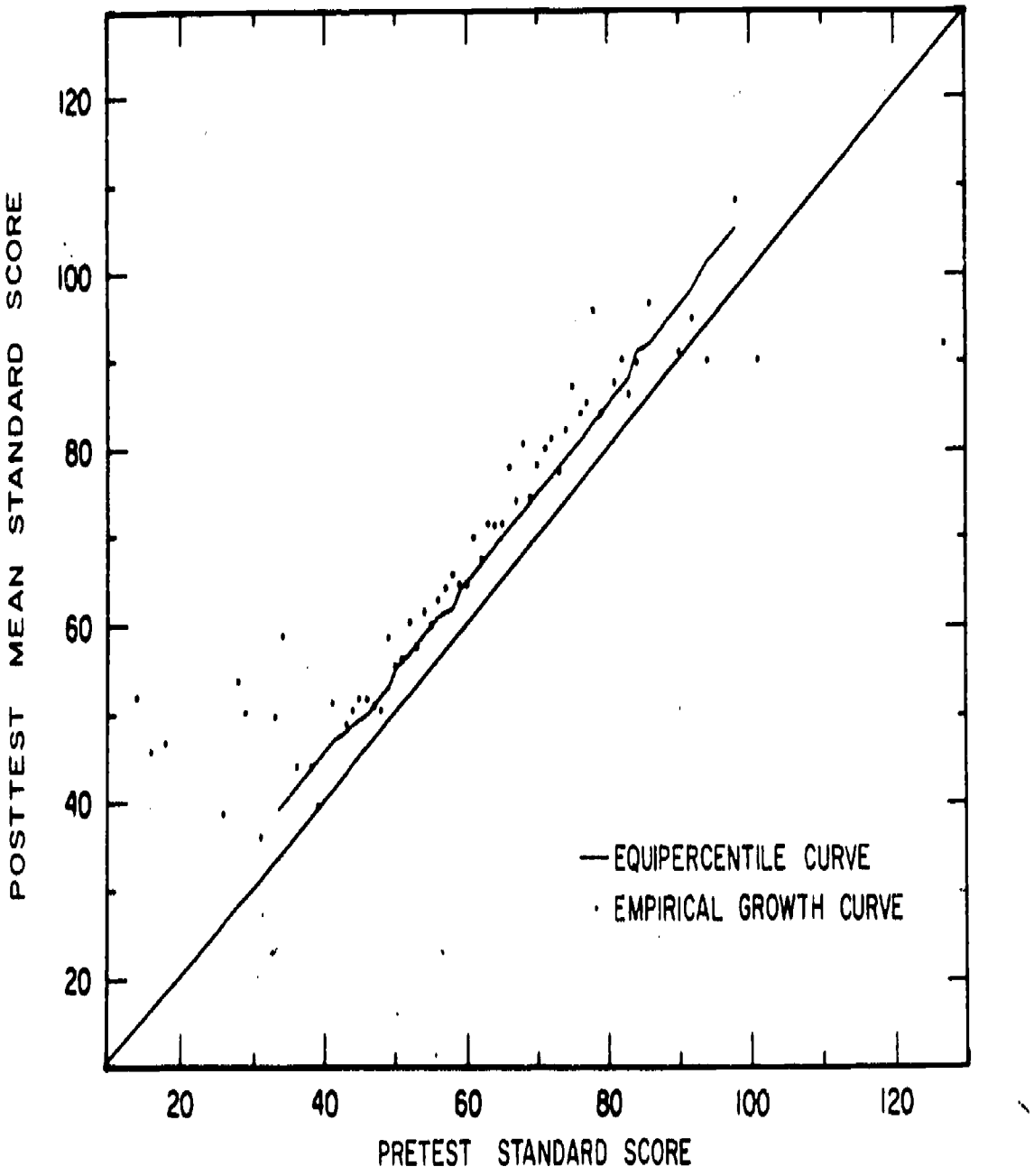


FIGURE 3.14 EMPIRICAL GROWTH CURVE FOR THE FOURTH GRADE NCR/SL GROUP ON MAT TOTAL READING: FALL PRETEST AND SPRING POSTTEST

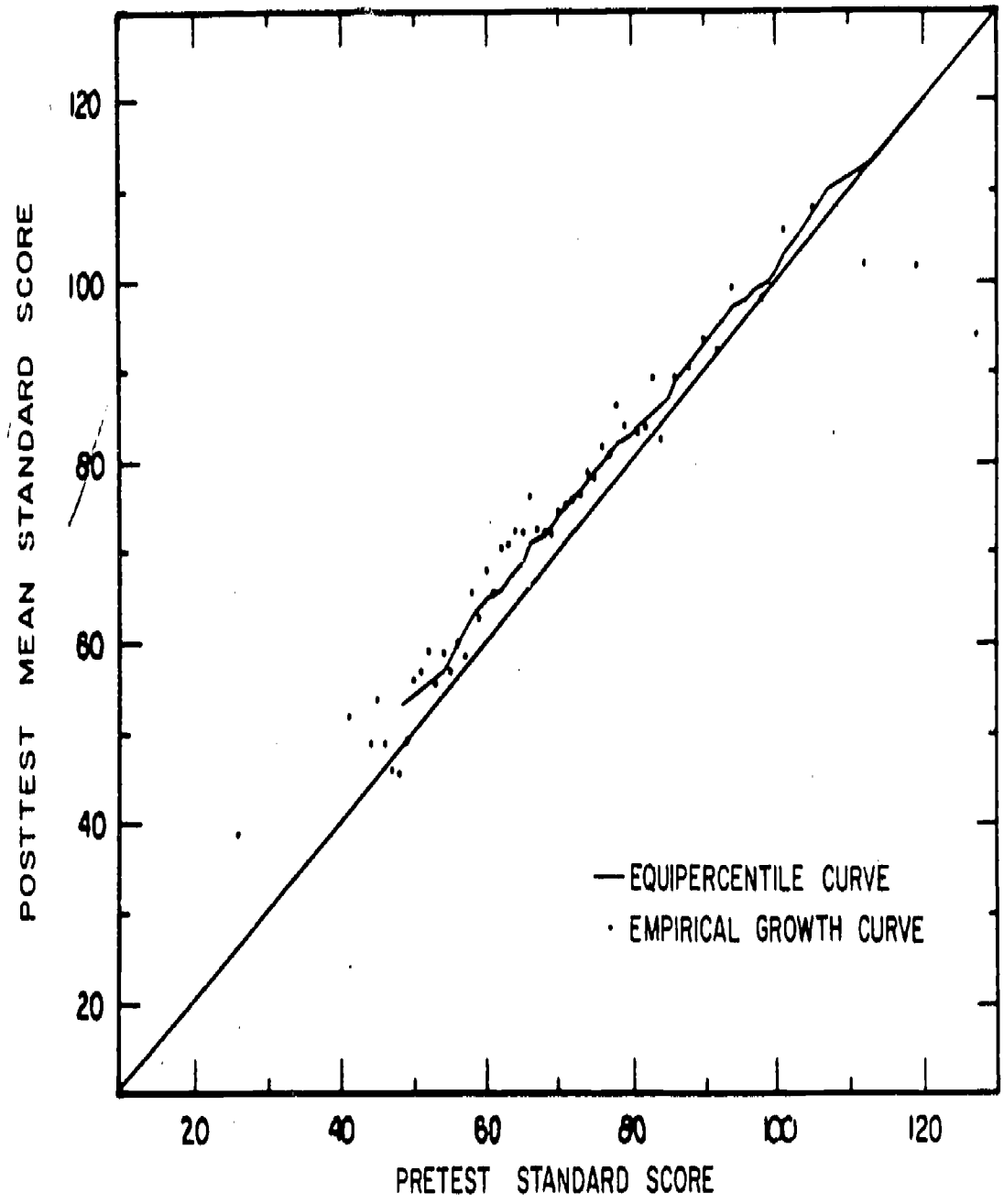


FIGURE 3.15 EMPIRICAL GROWTH CURVE FOR THE SIXTH GRADE NCR/SL GROUP ON MAT TOTAL READING: FALL PRETEST AND SPRING POSTTEST

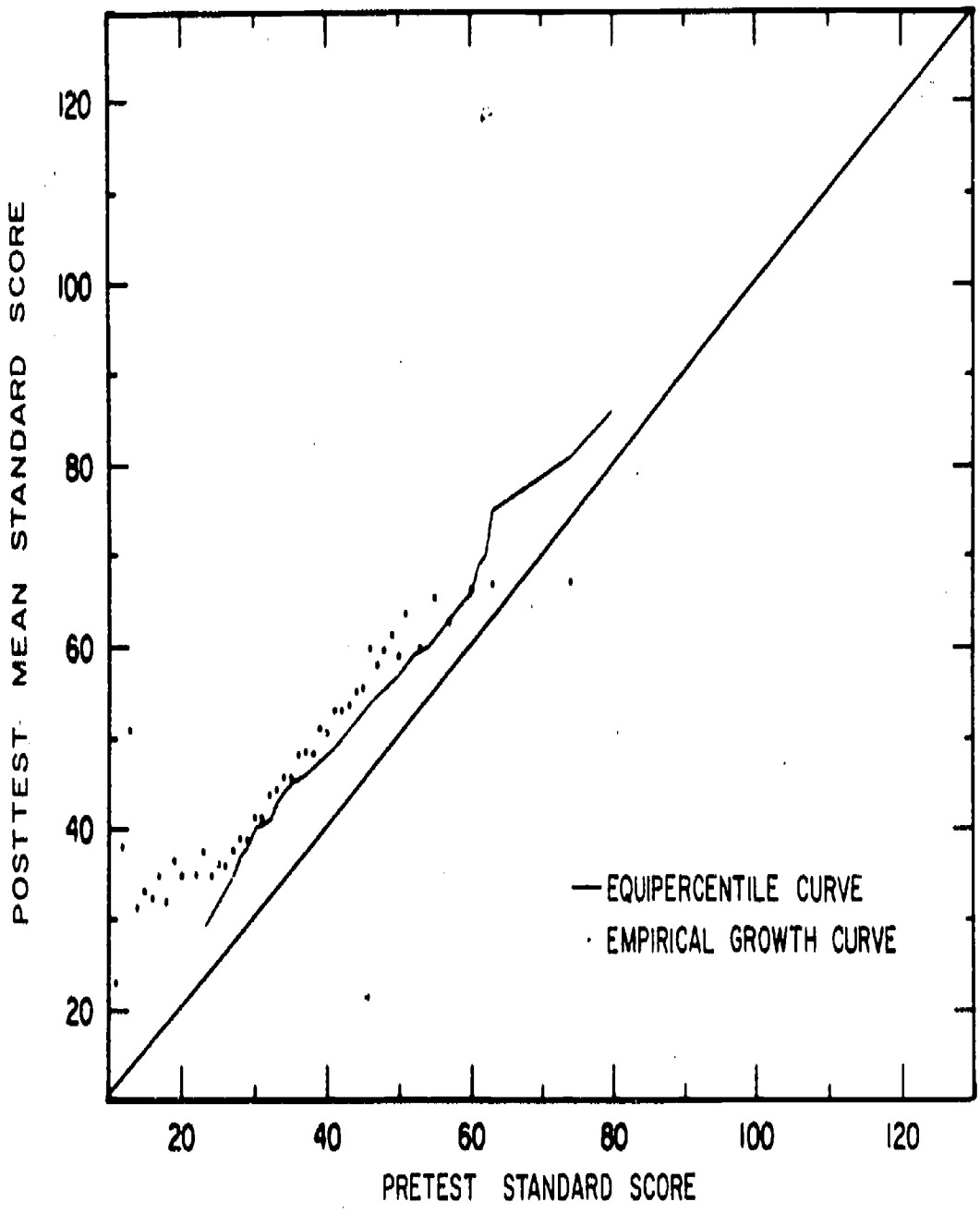


FIGURE 3.16 EMPIRICAL GROWTH CURVE FOR THE SECOND GRADE CR/SL GROUP ON MAT TOTAL READING: FALL PRETEST AND SPRING POSTTEST

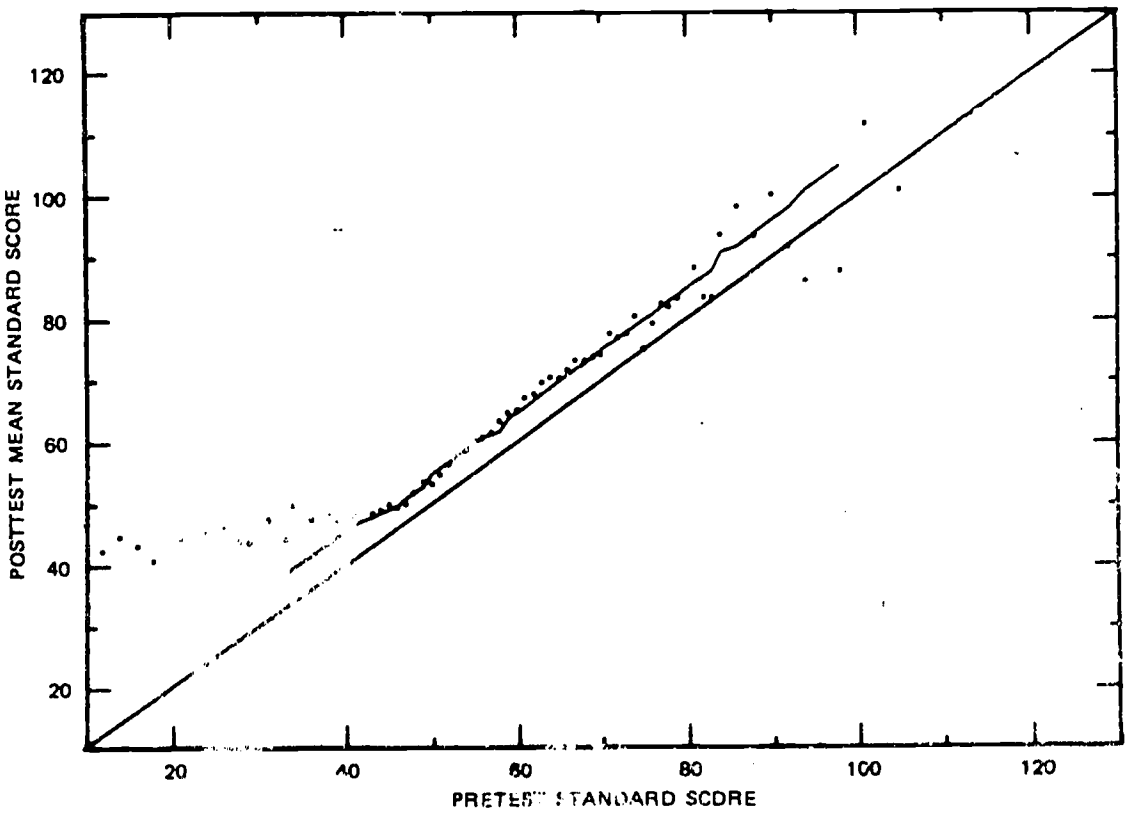


Figure 3.17 EMPIRICAL GROWTH CURVE FOR THE FOURTH GRADE CR/SL GROUP ON MAT TOTAL READING: FALL PRETEST, SPRING POSTTEST

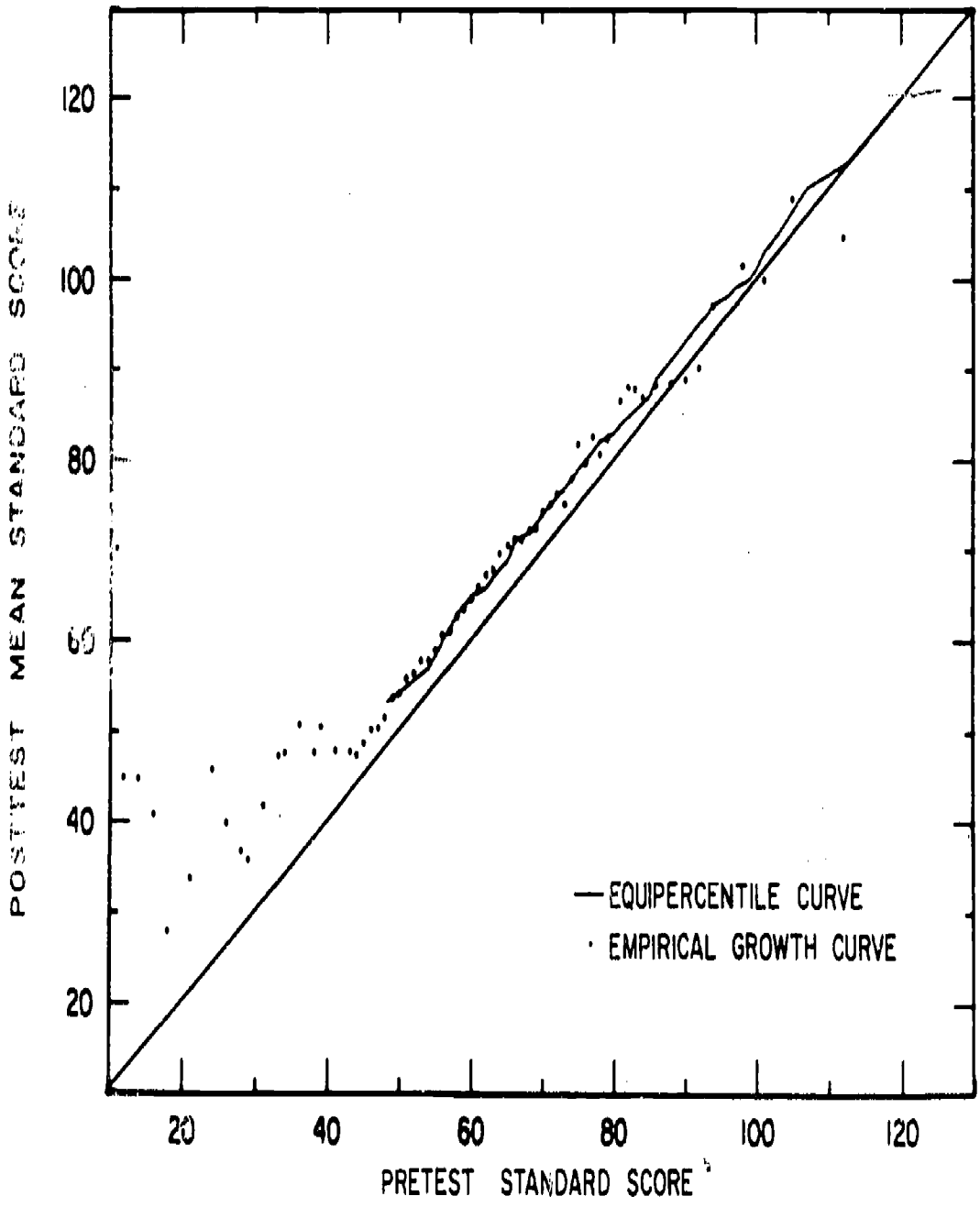


FIGURE 3.18 EMPIRICAL GROWTH CURVE FOR THE SIXTH GRADE CR/SL GROUP ON MAT TOTAL READING: FALL PRETEST AND SPRING POSTTEST

The empirical curves are consistently above the equal percentile curve for the second grade CR/SL and NCR/SL groups. For grades 4 and 6, the empirical curves track the equipercentile curve very closely in the midrange of pretest standard scores, but they tend to lie above the equipercentile curve for extremely low pretest standard score values.

As would be expected, pupils in the NCR/SL group have much higher scores than their peers in the CR/SL group. Their empirical growth curves tend to lie above the equipercentile curve across the range of pretest scores. Because the number of NCR/SL pupils was relatively low, the NCR/SL empirical curve appears to be much less smooth than other curves displayed.

Results by Minority Status

Criticism of the use of a norm-referenced procedure has been the strongest with respect to its use in the evaluation of programs that have a high proportion of minority students. Because the norm groups for most standardized tests are intended to be representative of the entire school population, they are composed predominantly of white middle-class students, whose performance, according to the critics, may differ considerably from that of the minority subpopulation.

For the norm-referenced procedure used in the PIP evaluation, the question is whether the function that describes the expected posttest standard score given the pretest standard score is different for minority pupils than for white pupils and whether the difference is large enough to make a difference in the results of the analysis. In this section, therefore, we will look at the relationship between pretest and posttest standard scores separately by minority status. Ideally, we would want to examine the effect of minority status controlling for other demographic and socioeconomic characteristics such as income, age, size of place of residence, and so forth. The differences in the CR and NFT groups on these types of variables are large. For example, the median household income for white NFT children was in the range between $7,700 and $8,000 and the median household income for minority NFT children was in the range between $5,000 and $5,200. Most white NFT children resided in relatively small towns and rural areas, whereas most minority NFT children resided in cities with a population of 200,000 or more.

Similar differences were found for the CR/SL groups. Only 5% of the white students lived in large cities, as compared to 28% of the minority students. About 75% of the white students were in schools where the estimated percent of children on public assistance is less than 25% compared to the 34% minority children in such schools. Therefore, this section should be considered an exploratory study. The results are presented by minority status to illustrate the differences in the normal growth curves that could result by such an approach.

Table 3.5 presents the summary data on the differences in gains between white and minority students in the NFT and CR evaluation groups. In every case, the mean pretest standard score for the white students is

Table 3.5

TRENDS IN PERCENTILE GAINS BY MINORITY STATUS

Data Base/Test	Initial Grade	Subgroup	Pre/Post Duration	Minority Status	N	Pretest Standard Score Mean	Pretest Standard Score S.D.	Pretest Percentile*	Posttest Standard Score Mean	Posttest Standard Score S.D.	Posttest Percentile*	Gain Statistics: Change in Percentile of Mean	Gain Statistics: Change of Percentile Mean	Gain Statistics: Change of Percentile S.D.	Gain Statistics: Percent with Percentile loss
Non-Follow Through															
Total Reading	1		Sp 1973/Sp 1974	White	806	43.6	10.6	69.2	55.6	10.7	64.4	- 4.8	- 5.8	16.9	65%
				Minority	1289	36.6	8.9	37.8	48.0	9.0	28.0	- 9.8	- 7.6	17.9	66
	2		Sp 1974/Sp 1975	White	847	56.2	10.8	66.8	63.9	12.2	59.8	- 7.0	- 1.5	12.8	54
				Minority	1194	47.8	9.1	27.0	53.3	10.9	25.2	- 1.8	- 3.0	14.6	63
	1		Sp 1973/Sp 1975	White	903	44.5	10.5	72.0	64.5	12.1	62.0	-10.0	- 7.1	19.5	68
				Minority	1462	36.5	8.9	37.0	53.1	11.0	24.4	-12.6	-10.3	20.2	70
Compensatory Reading Study															
Total Reading	2	CR/SL	Fall/spring	White	1473	36.8	8.7	19.6	48.7	10.6	32.2	+12.6	11.7	22.0	32
				Minority	1337	32.5	8.5	11.0	43.2	10.2	12.4	+ 1.4	6.7	18.6	34
		CR/NSL	Fall/spring	White	2094	38.3	8.3	24.6	50.4	10.5	40.2	+15.6	12.0	21.5	29
				Minority	574	34.8	9.2	15.6	46.3	10.2	20.9	+ 5.3	8.2	19.4	33
		NCR/SL	Fall/spring	White	458	43.7	10.6	44.8	55.8	11.0	65.2	+20.4	12.8	22.7	28
				Minority	146	36.4	8.5	18.8	46.4	11.7	21.2	+ 2.4	7.2	19.3	38
	4	CR/SL	Fall/spring	White	1551	55.9	11.8	23.8	62.6	12.3	29.2	+ 5.4	3.5	12.0	37
				Minority	1445	47.1	11.5	7.1	53.7	10.5	10.7	+ 3.6	1.9	9.5	38
		CR/NSL	Fall/spring	White	1702	56.3	11.2	24.6	63.7	11.6	31.4	+ 6.8	5.1	12.9	33
				Minority	423	52.0	10.5	16.0	58.4	11.7	18.8	+ 2.8	3.5	11.7	40
		NCR/SL	Fall/spring	White	394	63.9	13.4	45.6	71.6	14.3	52.4	+ 6.8	5.7	12.3	31
				Minority	134	51.6	12.4	15.2	58.4	12.2	18.8	+ 3.6	2.7	11.0	39
	6	CR/SL	Fall/spring	White	1151	66.7	12.9	16.7	71.5	13.6	17.0	+ .3	2.2	14.4	41
				Minority	1386	56.9	11.6	4.9	61.6	12.3	5.1	+ .2	1.3	8.4	31
		CR/NSL	Fall/spring	White	1740	71.4	14.0	24.8	76.2	14.8	26.4	+ 1.6	2.7	16.5	42
				Minority	383	64.1	12.9	12.2	68.7	13.5	13.4	+ 1.2	2.0	11.4	37
		NCR/NSL	Fall/spring	White	467	79.0	15.5	44.0	82.9	16.0	45.6	+ 1.6	1.6	18.0	44
				Minority	76	63.6	11.5	11.6	67.6	11.3	11.6	.0	.5	12.2	40
Non-Follow Through															
Total Math	1		Sp 1973/Sp 1974	White	794	42.7	11.3	51.4	56.7	10.9	48.2	- 3.2	- 3.1	18.0	55
				Minority	1217	36.2	10.6	32.8	48.5	10.0	13.0	-19.8	- 8.3	19.0	63
	2		Sp 1974/Sp 1975	White	831	56.9	11.1	49.4	68.8	12.7	43.2	- 6.2	- .5	16.2	49
				Minority	1101	48.3	10.0	12.6	59.2	11.4	16.4	+ 3.8	.4	15.1	46
	1		Sp 1973/Sp 1975	White	888	43.5	11.2	55.0	69.4	12.5	45.6	- 7.4	- 3.9	19.9	56
				Minority	1363	36.0	10.2	32.0	59.4	11.1	16.8	-12.2	- 7.2	21.2	60

*Transform of mean standard score using linear interpolation

substantially higher than the mean for the minority students. On Total Reading, minority pupils consistently gained less (or lost more) in the percentile of the mean and in the mean percentile ranks than white pupils. This was true for every group with the exception of the NFT group's change in the percentile of the mean standard score between second and third grade. Even here, however, the mean change in percentile was -3.0 for the minority students and -1.5 for white students. The minority groups had a consistently higher percent of students who had a lower percentile rank in the spring than in the fall, with the exception of the grade 6 CR groups, as shown in the last column of Table 3.5. The difference between the white and minority groups on percent decline appears to be extremely small, however.

For Total Math, the differences between white and minority NFT students are similar to those found for Total Reading. Minority students as a group lost substantially more than white students between the spring of grade 1 and the spring of grade 2, but the difference appears to have reversed between spring of grade 2 and spring of grade 3.

Figure 3.19 and 3.20 present the empirical growth curves separately for white and minority NFT children on the MAT Total Reading and Total Math, respectively. The pretest was taken in the spring of grade 1 and the posttest was taken in the spring of grade 3. With a few exceptions for some extremely low pretest scores, the empirical curve for the white population lies above that for minorities for Total Reading. For Total Math, the two curves are close, up to a pretest standard score of 34 (24th percentile), and then tend to diverge.

Figures 3.21, 3.22 and 3.23 present the corresponding empirical growth curves for the CR/SL students in grades 2, 4, and 6, respectively, on Total Reading. For each grade, there is a range in which the empirical growth curve for white students is consistently higher than that for minority students. Across grade levels, the lower end of the range corresponds to about the second percentile. The upper end of the range appears to decrease over grade level in terms of percentile rank from about 68 in grade 2, to about 48 in grade 4, to about 12 in grade 6. This corresponds to a progressive decrease in the percentile rank of the mean standard score for minority CT/SL students from 11.0 in grade 2, to 7.1 in grade 4, to 4.9 in grade 6.

The differences shown for the NFT groups in Figure 3.19 and 3.20 are larger than those illustrated for the CR/SL groups because of the differences in the length of time between pre- and posttest, of course. For the single year time periods the differences exhibited in the NFT groups between white and minority empirical growth curves are consistent with those shown for the CR/SL groups, however.

In general, it appears from this data that the difference in the expected normal growth between white and minority students is between 1 and 2 standard score points between a fall pretest and spring posttest within the same academic year. The significance of such a difference in the norm-referenced analysis would depend on the number of pupils in the evaluation and the grade level at which the evaluation takes place. The analyses that were conducted above were comparisons within particular NFT and CR groups. If other variables could have been controlled so as

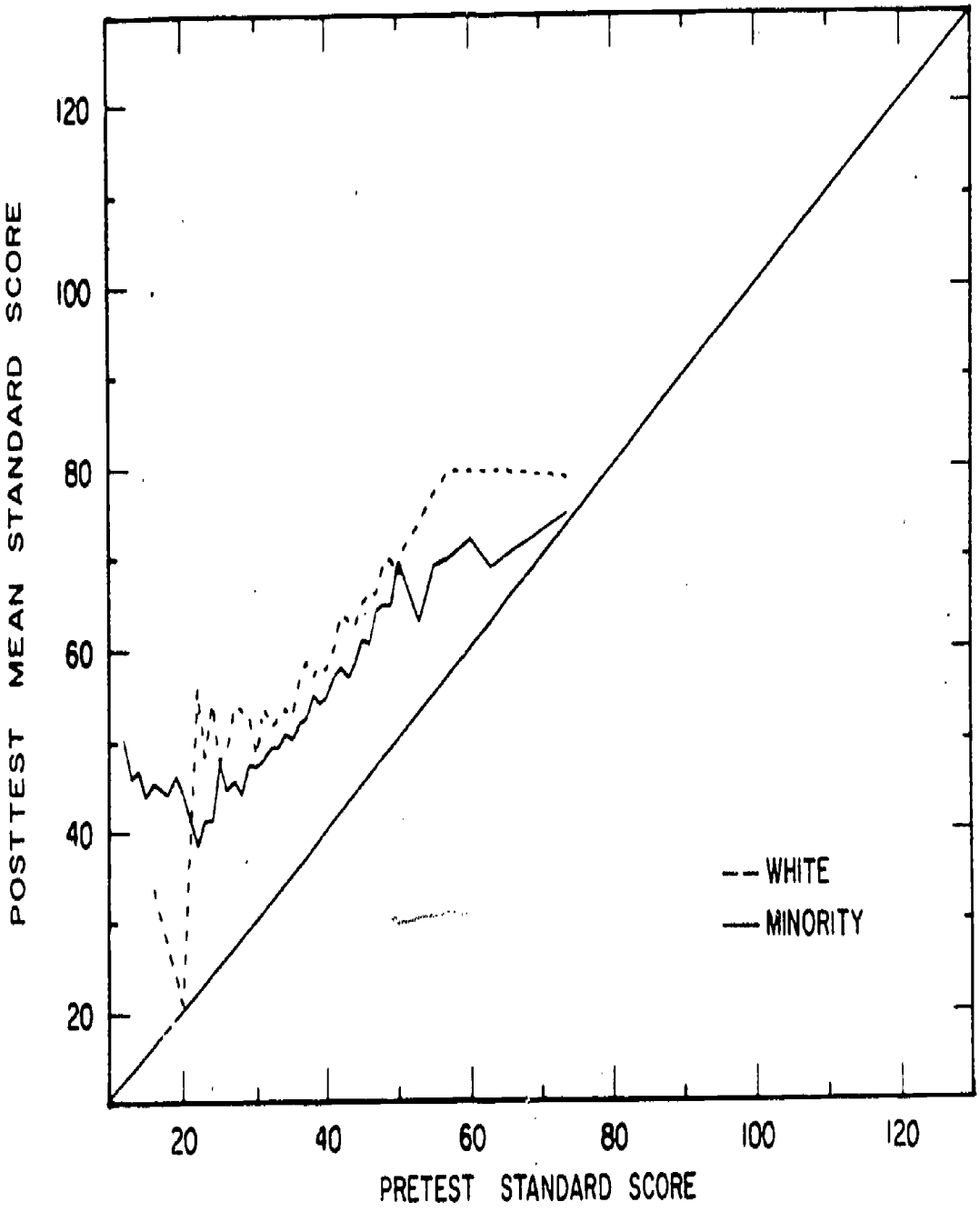


FIGURE 3.19 EMPIRICAL GROWTH CURVES FOR NFT GROUPS, FIRST TO THIRD GRADE, ON TOTAL READING BY MINORITY STATUS: SPRING PRETEST AND POSTTEST

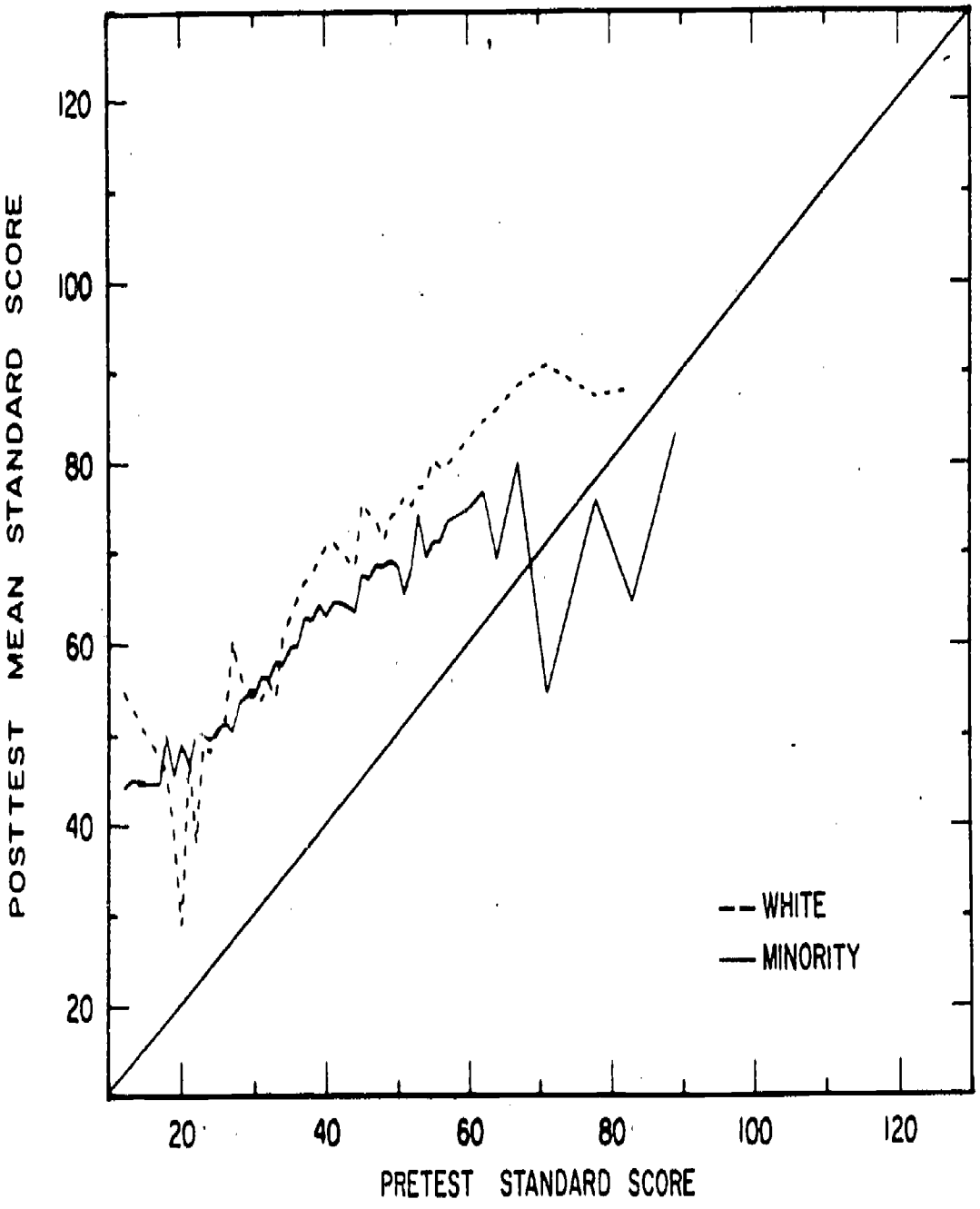


FIGURE 3.20 EMPIRICAL GROWTH CURVES FOR NFT GROUPS, FIRST TO THIRD GRADE, ON TOTAL MATH BY MINORITY STATUS: SPRING PRETEST AND POSTTEST

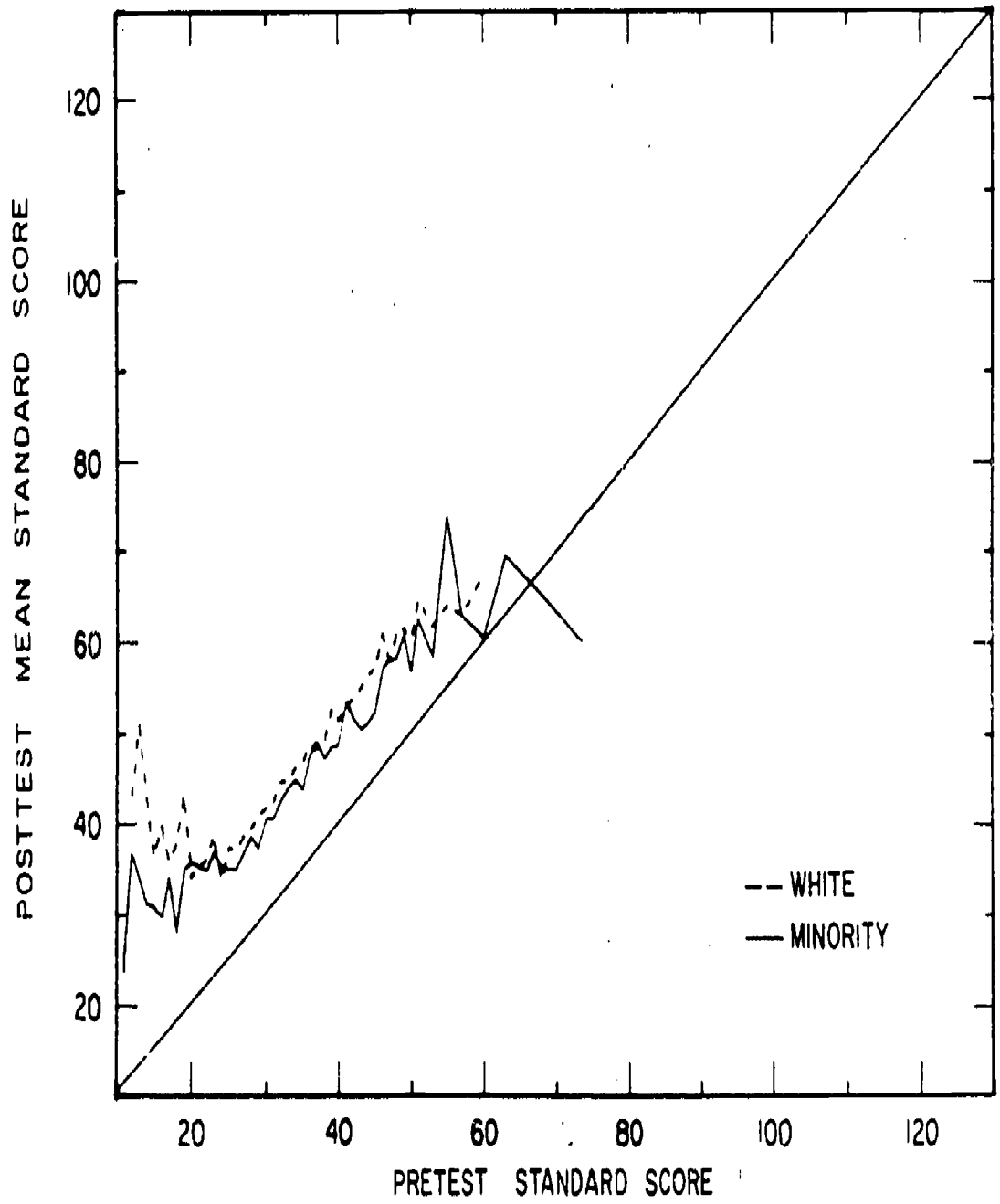


FIGURE 3.21 EMPIRICAL GROWTH CURVES FOR THE SECOND GRADE CR/SL GROUP ON MAT TOTAL READING, BY MINORITY STATUS: FALL PRETEST, SPRING POSTTEST

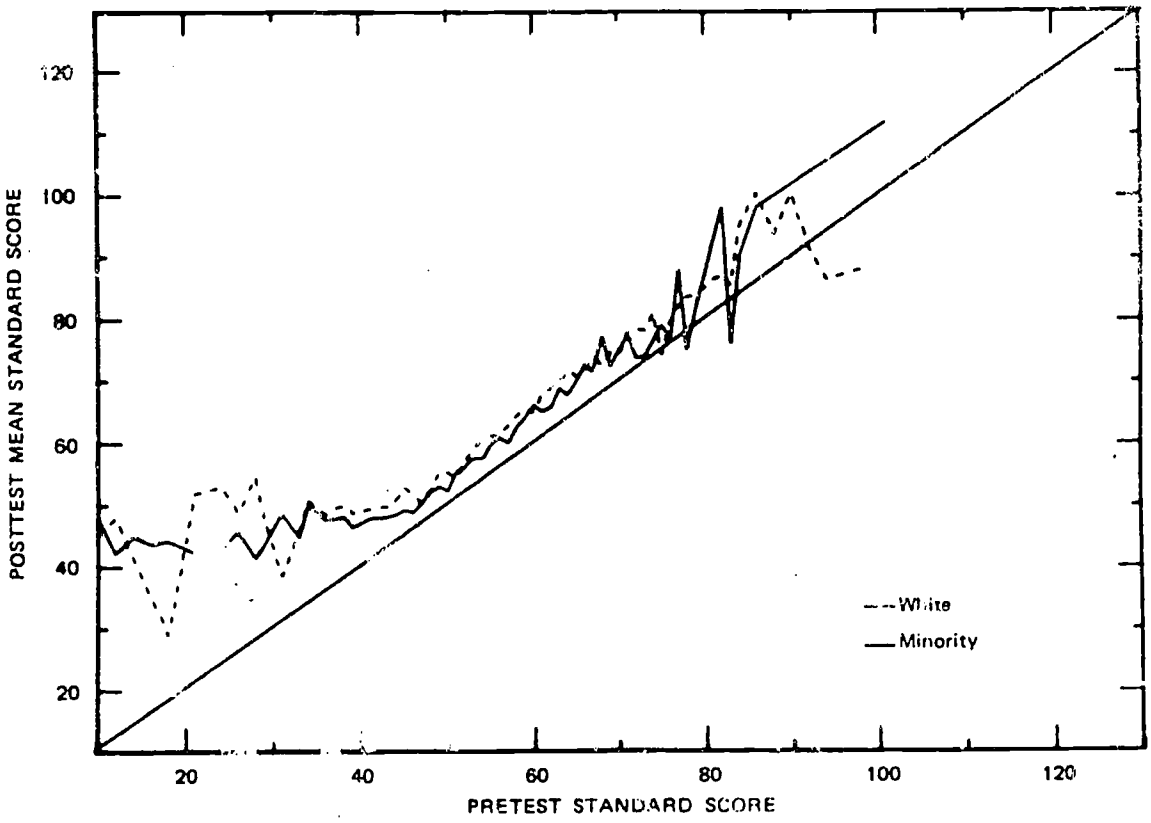


Figure 3.22 EMPIRICAL GROWTH CURVE FOR THE FOURTH GRADE CR/SL GROUP ON MAT TOTAL READING BY MINORITY STATUS: FALL PRETEST, SPRING POSTTEST

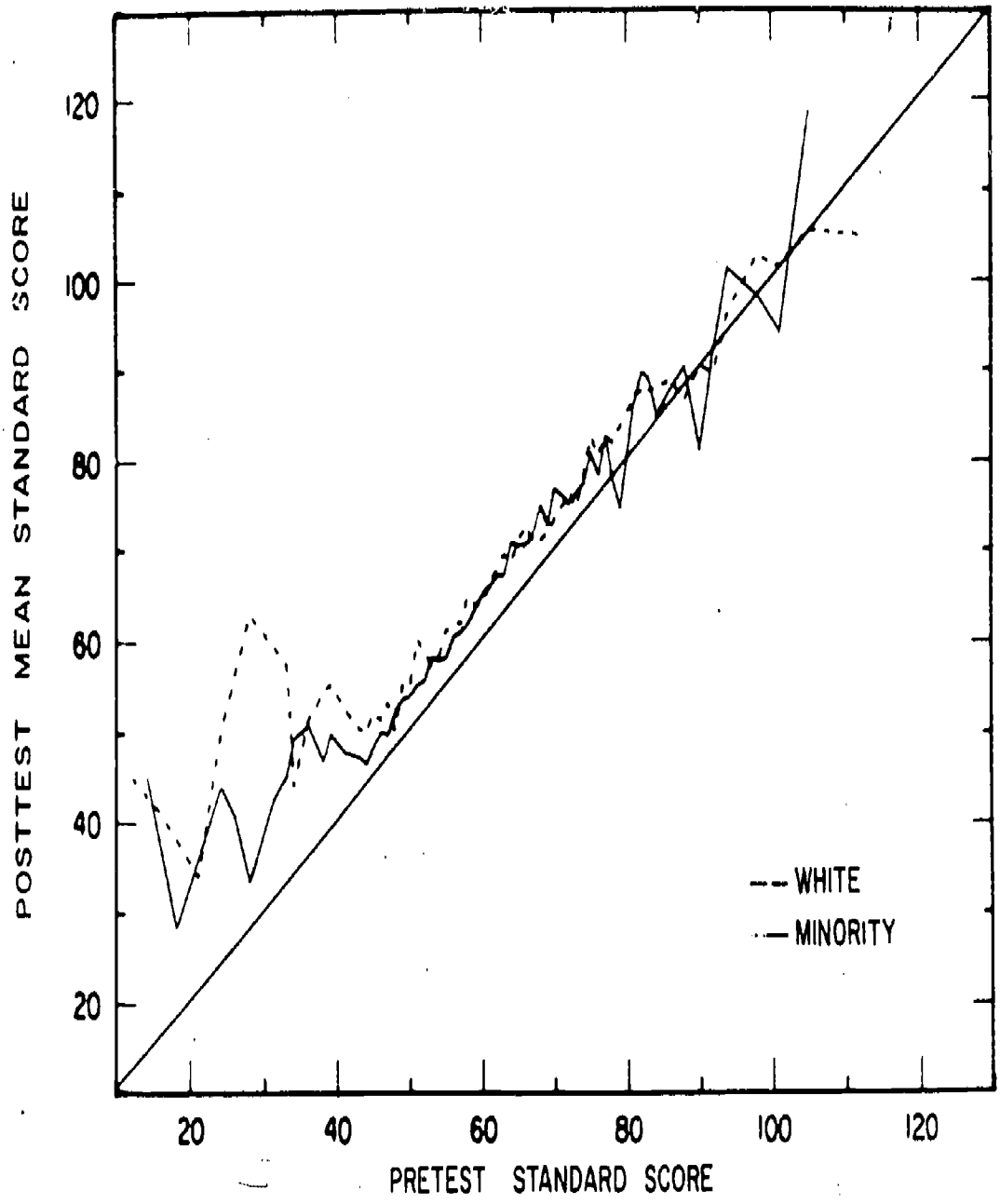


FIGURE 3.23 EMPIRICAL GROWTH CURVES FOR THE THIRD GRADE CR/SL GROUP ON MAT TOTAL READING, BY MINORITY STATUS: FALL PRETEST AND SPRING POSTTEST

to make the minority and white populations more similar, the differences would be expected to be less than those observed. On the other hand, if the performance of the minority students had been compared to a group that was more diverse, the differences would probably be even larger, as observed in comparing the minority NFT students' performance to the norm group.

Summary

In summary, there appear to be two countervailing factors affecting the outcome of the norm-referenced analysis:

(1) For pupils with extremely low pretest scores, the equal percentile assumption leads to a predicted posttest standard score that is much lower than what was observed in the MAT longitudinal data, the NFT data, and the CR data.

(2) Use of the norms based on the standardization group will lead to an expected posttest score that will be too high for students ordinarily in compensatory programs, especially minority students who have pretest scores that are not extremely low.

Depending on the distribution of children on pretest scores, then, the equal percentile assumption could lead either to the conclusion that a program did have an impact when in fact it did not or to the conclusion that the program did not have an impact when in fact it did.

The problem of invalid scores because of extremely high or extremely low scores has been recognized both by the MAT test developers (Prescott, 1973) and evaluators such as Horst et al. (1975). The usual recommendation to minimize such effects is to test very low achievers (those most likely to be encountered in a compensatory-type program) one or two levels below that used to standardize the test. The rule of thumb is to select the battery level so that pupils score near the middle of the range of possible raw scores.

Testing out of level requires greater reliance on the standard score metric because it is this metric that links various levels of the test and allows the transformation of a raw score on an out-of-level test to a percentile rank. Although the linkage between the tests may be imperfect, Horst et al., for one, prefer out-of-level testing when the alternative in-level testing may result in a substantial number of children having extremely high or low raw scores.

For the PIP evaluation, as with most evaluations, only one level of the test could be administered per grade level at a given project for administrative reasons. In some grade levels the tests that were administered were one level below that recommended by the MAT developers, but some children still had extreme raw scores.

Out-of-level testing does appear to diminish the phenomenon of the predicted posttest score being too high for extremely low scores, at least for CR students in Total Reading at sixth grade (see Figure 3.18). Whether such a salutary effect can be generalized to other grade levels and tests is still open.

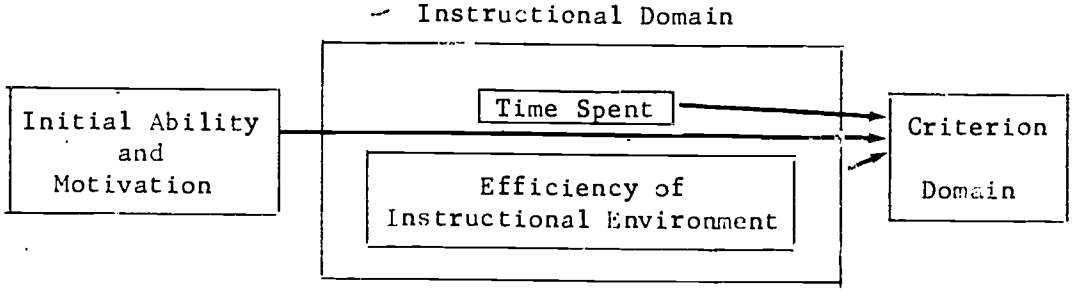
IV STRINGENCY OF THE NORMAL GROWTH AND EDUCATIONALLY SIGNIFICANT GROWTH CRITERIA

Ideally, the criteria of normal growth and educationally significant growth should be established and implemented so that the degree of difficulty in attaining the criteria is independent of grade or pretest score. It should be about as difficult to demonstrate program effectiveness at the second grade as at the seventh grade; for children whose pretest scores are at the fourth percentile as for those at the 50th percentile; and for programs in math as for programs in reading. Variation in stringency could lead to policy decisions that are based on artifacts of the evaluation procedure rather than the educational impacts of a program.

Concept of Stringency

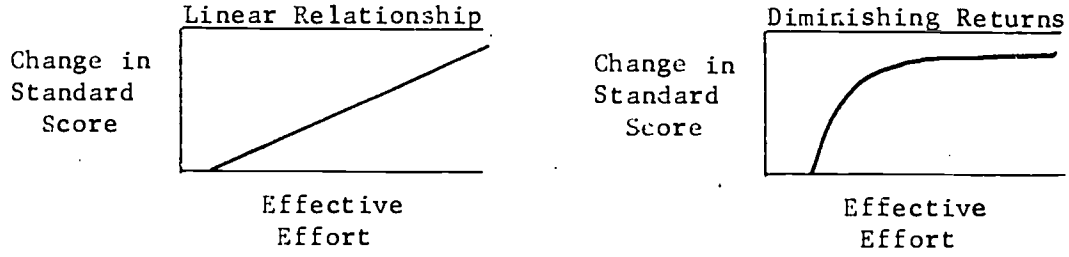
The standard score on a standardized test is conventionally interpreted as a measure of achievement. Changes in standard score from pre- to posttest may be attributed to a combination of errors in measurement, learning, forgetting, and so forth. Some portion of the increase in standard score may be attributed to the instructional process that takes place in the classroom. The notion of stringency of a criterion is related to the amount, intensity, and effectivenss of instruction necessary to achieve the criterion. The more effort required to pass a criterion, the more stringent it is. To describe fully the difficulty of attaining normal growth or educationally significant growth, one would require a model of the instructional process that included measures of effort describing the relationship of instructional events to performance criteria.

For example, Cooley and Lohnes (1976) proposed an instructional model that related performance on a set of criteria to the initial ability and motives of the learner, the time spent by the learners, and the efficiency of the instructional environment. This is represented schematically below:



This model suggests that an instructional environment could be described in terms of "effective effort"--in this case time spent adjusted or standardized by the efficiency of the instructional environment. The relationship between the effective effort and change in

standard score would be described statistically rather than deterministically because of factors other than the instructional process and because of variance in measuring achievement due to item selection and standard errors of measurement. Nevertheless, for a satisfactory definition of effort, a monotome nondecreasing relationship between effective effort and increase in standard scores would be necessary. As effective effort increases it would be expected that the gain in standard score would also increase. Assuming some measure of effective effort were available, we could investigate the relationship between this measure and changes in standard score. We could hypothesize a linear relationship between effective effort and gain in standard score, or a "diminishing return" relationship. These two relationships are illustrated below:



The linear relationship model would indicate an increase in effective effort would result in a proportional increase in standard score points. The diminishing returns model hypothesizes that increments in effective effort result in diminishing increments in standard score.

The difficulty or stringency of a criterion in educational terms, then, is related to the effort necessary to achieve the necessary gains, where effort is measured in terms of educational resources and time needed for students to acquire the necessary skills. If reliable measures of effective effort were available then difficulty would be expressed mathematically as $\frac{\text{effort}}{\Delta \text{ standard score}}$, the amount of effort necessary to achieve a specified gain in standard score.

The amount of effort necessary for a specified gain in standard score no doubt depends on a student's initial abilities, his motivation, the subject matter, and the level of the test. The principle of equal stringency for a criterion such as educationally significant growth calls for a specification of the gains in standard score such that the effort necessary to achieve the criterion is constant across grade level, pretest scores, etc.

The problem in assessing stringency is that educational theory cannot come near to providing a measure of effective effort. While the standard score metric is purported to be an equal interval scale of achievement, a specified gain in standard score probably does not require the same amount of effort independent of grade level or initial standard score. Also, effort per se is not enough; it must be effort directed to the acquisition of skills measured by the MAT.

Under the equipercentile assumption, the gains in standard score necessary for normal growth vary considerably over grade level. For Total Reading, for example, a gain of between 7 and 9 standard score points is necessary to maintain normal growth at the second grade between fall and spring. This decreases to a gain between 1 and 4 standard score points for eighth grade. In several instances, mostly in the upper grade levels, the equal percentile assumption dictates that zero gain between fall and spring is sufficient for normal growth. Below the 50th percentile, the specified standard score gains necessary for normal growth are quite uniform across percentile ranks within grade.

The substantial differences across grade may be attributed to any of a number of factors:

- The test items become increasingly irrelevant to the type of skills being taught in the upper grades.
- Students tend to reach an asymptote in their acquisition of reading skills and additional gains require much more effort than at the lower grade levels.
- The standardization procedure was defective.
- Less time and effort are spent in the upper grades in acquiring the skills measured by the MAT.

If it is assumed that the standardization program produced valid norms, then either pupils appear to reach some asymptote in the upper grade levels on reading achievement or the curriculum at the upper grade level is irrelevant to the skills measured by the MAT. If the curriculum is irrelevant then the observed low gains in standard scores may not indicate a great degree of difficulty in achieving growth, but rather that students are not spending much time learning skills relevant to items on the MAT.

To assess the difficulty of achieving specified gains on the MAT cannot be done with any degree of accuracy given the current state of educational theory. However, the gains by the MAT standardization group as indicated in the norm tables may be taken as representative of the output of programs with an average effort expended. The gains necessary for educationally significant growth may be compared to the expected gains under the normal growth assumption, the principle being that the gains necessary for educationally significant growth should have some systematic and reasonable relationship to the gains necessary for normal growth.

Stringency of the Criterion of Educationally Significant Growth Relative to Normal Growth

As was indicated in the introductory section, the criterion for educationally significant growth is that the gain in standard score be one-third of a standard deviation above that predicted by the equipercentile model of normal growth. This criterion was proposed as a rule

of thumb and very little was said in Horst et al. (1975) regarding its properties, in particular the relationship between educationally significant growth and normal growth.

Tables 4.1 and 4.2 show the gain in standard score points needed between fall and spring to achieve normal growth and educationally significant growth by grade level and fall percentile score for Total Reading and Total Math, respectively. The gains necessary for normal growth, as was indicated earlier, decrease with increasing grade level. For the most part, they are constant across fall percentile score within grade level, for the 50th percentile or less. Because the standard deviation of standard scores tends to increase over grade levels, we find that the additional gain necessary for educationally significant growth is an increasing fraction of normal growth. From the point of view of the equal stringency principle, this indicates that educationally significant growth is much more difficult to attain in the upper grade levels when equipercentile growth is taken as the base.

As another approach to assessing the relationship between educationally significant growth and expected normal growth on Total Reading for various percentiles, a polynomial regression equation was fit to the normative standard scores and percentiles. The basic idea was to simultaneously find two polynomials, $P_1(t)$ and $P_2(t)$, such that for a standard score y:

$$f(z,t) = y = P_1(t) + zP_2(t) \quad ,$$

where z is the point on the standard normal distribution corresponding to the percentile for y and t represents time in school in months. In this representation, $P_1(t)$ describes changes at the mean as a function of time and P_2 describes changes in the standard deviation.

After some preliminary runs, a second degree polynomial was selected for P_1, and a fifth degree for P_2. The resulting equations are:

$$\mu(t) = P_1(t) = [.17 + (8.2 \times 10^{-3})\, t - (2.8 \times 10^{-5})\, t^2] \times 132$$

$$\sigma(t) = P_2(t) = [.12 - (7.4 \times 10^{-3})\, t + (3.7 \times 10^{-4})\, t^2 - (7.3 \times 10^{-6})\, t^3 + (6.3 \times 10^{-8})\, t^4 - (2 \times 10^{-10})\, t^5] \times 132 \quad ,$$

where t is time in months from beginning of kindergarten.

We have shown the coefficients to two places. The five-place equation we fit by BMD 07R (Dixon, 1973) has a coefficient of determination of .995 on 756 degrees of freedom. While this coefficient is large enough for our present purposes, the reader is cautioned that errors as large as 10% can be found fairly frequently, when predicted norm standard scores are compared with actual. Overall, however, predicted standard

Table 4.1

GAIN IN STANDARD SCORE POINTS NEEDED IN THE SPRING FOR NORMAL GROWTH AND EDUCATIONALLY SIGNIFICANT GROWTH BY GRADE AND FALL PERCENTILE SCORE: MAT TOTAL READING FALL TO SPRING

Fall Percentile Score		Grade 2	3	4	5	6	7	8
4	NG*	9	0	5	5	4	1	2
	ES**	3.6	4.3	4.8	4.3	4.5	5.3	5.6
	ES/NG	.4	†	.96	.86	1.12	5.3	2.65
10	NG	9	3	4	5	4	1	1
	ES	3.6	4.3	4.8	4.3	4.5	5.3	5.6
	ES/NG	.4	1.43	1.20	.86	1.12	5.3	5.60
20	NG	9	3	5	6	4	0	1
	ES	3.6	4.3	4.8	4.3	4.5	5.3	5.6
	ES/NG	.4	1.43	.96	.72	1.12	†	5.60
30	NG	8	4	5	5	3	0	1
	ES	3.6	4.3	4.8	4.3	4.5	5.3	5.6
	ES/NG	.45	1.07	.96	.86	1.50	†	5.60
40	NG	8	4	5	4	4	0	1
	ES	3.6	4.3	4.8	4.3	4.5	5.3	5.6
	ES/NG	.45	1.02	.96	1.08	1.12	†	5.6
50	NG	7	4	5	3	3	1	1
	ES	3.6	4.3	4.8	4.3	4.5	5.3	5.6
	ES/NG	.51	1.07	.96	1.43	1.50	5.3	5.6
60	NG	7	5	5	3	2	2	1
	ES	3.6	4.3	4.8	4.3	4.5	5.3	5.6
	ES/NG	.51	.86	.96	1.43	2.25	2.65	5.6
70	NG	7	6	5	3	3	3	2
	ES	3.6	4.3	4.8	4.3	4.5	5.3	5.6
	ES/NG	.51	.67	.96	1.43	1.50	1.77	2.3
80	NG	6	6	5	4	3	3	3
	ES	3.6	4.3	4.8	4.3	4.5	5.3	5.6
	ES/NG	.60	.71	.96	1.08	1.50	1.77	1.87
90	NG	8	5	7	4	1	4	4
	ES	3.6	4.3	4.8	4.3	4.5	5.3	5.6
	ES/NG	.45	.86	.69	1.08	4.5	1.33	1.40

*NG = Gain necessary for normal growth.

**ES = Gain necessary over normal growth for educationally significant growth

†NG = 0

Table 4.2

GAIN IN STANDARD SCORE POINTS NEEDED IN THE SPRING FOR NORMAL GROWTH AND EDUCATIONALLY SIGNIFICANT GROWTH BY GRADE AND FALL PERCENTILE SCORE: MAT TOTAL MATH FALL TO SPRING

Fall Percentile Score		Grade 2	3	4	5	6	7	8
4	NG*	11	7	9	6	4	1	0
	ES**	3.7	4.0	4.0	4.1	4.2	4.3	4.8
	ES/NG	.34	.57	.44	.68	1.05	4.3	†
10	NG	14	6	7	4	3	1	0
	ES	3.7	4.0	4.0	4.1	4.2	4.3	4.8
	ES/NG	.26	.67	.57	1.03	1.40	4.3	†
20	NG	15	8	8	5	4	1	0
	ES	3.7	4.0	4.0	4.1	4.2	4.3	4.8
	ES/NG	.25	.50	.50	.82	1.05	4.3	†
30	NG	13	9	8	4	3	1	0
	ES	3.7	4.0	4.0	4.1	4.2	4.3	4.8
	ES/NG	.28	.44	.50	1.03	1.40	4.3	†
40	NG	10	10	9	4	3	1	0
	ES	3.7	4.0	4.0	4.1	4.2	4.3	4.8
	ES/NG	.37	.40	.44	1.03	1.40	4.3	†
50	NG	9	9	8	4	3	1	1
	ES	3.7	4.0	4.0	4.1	4.2	4.3	4.8
	ES/NG	.41	.44	.50	1.03	1.40	4.3	4.8
60	NG	8	9	8	4	4	2	1
	ES	3.7	4.0	4.0	4.1	4.2	4.3	4.8
	ES/NG	.46	.44	.50	1.03	1.05	2.15	4.8
70	NG	9	11	8	4	4	2	1
	ES	3.7	4.0	4.0	4.1	4.2	4.3	4.8
	ES/NG	.41	.36	.50	1.03	1.05	2.15	4.8
80	NG	10	10	7	5	6	2	2
	ES	3.7	4.0	4.0	4.1	4.2	4.3	4.8
	ES/NG	.37	.40	.57	.82	.70	2.15	2.40
90	NG	10	8	7	6	6	3	4
	ESES	3.7	4.0	4.0	4.1	4.2	4.3	4.8
	ES/NG	.37	.50	.57	.68	.70	1.43	1.20

*NG = Gain necessary for normal growth.

**ES = Gain necessary over normal growth for educationally significant growth.

†NG = 0

scores for the five-place equation are quite close to the actual scores found in the norm tables. Presumably this good fit reflects the normalizing transformation used to construct the standard scores.

Table 4.3 shows the rates of change of the mean, $\frac{dP_1(t)}{dt}$, and standard deviation $\frac{dP_2(t)}{dt}$, for the fitted normative standard scores. Compared to changes in the mean, changes in the standard deviation are small. The fitted data indicate there is a slight tendency for the standard deviation to decrease at the higher grades; essentially, however, the one-third the MAT norm standard deviation criterion for educationally significant growth is constant for all grades. Because average growth decreases as a function of grade, the criterion for educationally significant growth is an increasing function of average growth.

If we take the growth required by the equipercentile assumption as the measure of gains under average effective effort, then the results imply that, in terms of average growth, the difficulty of achieving educationally significant growth increases with grade level.

The question whether the difficulty of achieving educationally significant growth varies across pretest scores within grade level was examined by computing the rate of change in standard score as a function of percentile. If the rate of change is close to constant across percentile, then we may conclude that the effort needed to increase standard scores is about constant and that because the criterion for educational significance is constant for a given grade level, then the difficulty of the criterion is constant across percentile rank for a given grade level.

In Table 4.4 the values of the reciprocal of the rate of change in standard score over a nine-month time period,

$$9\left[\frac{dP_1(t)}{dt} + Z\,\frac{dP_2(t)}{dt}\right]^{-1}$$

are given for selected percentile ranks for grades 2 and 8 on Total Reading.

It can be seen from Table 4.4 that the ratio of the criterion to the growth of the fitted data is relatively constant across percentiles within grades, but not between grades. However, the relatively little variation within grades probably should not be neglected, because at the second grade the criterion is a larger fraction of growth of the fitted scores at low percentiles than at high ones, while the reverse is true at the eighth grade. The main explanation of this is shown in column three of Table 4.4; that is, for the fitted data, the MAT reading skills for children at the 90th percentile of the norm group in the second grade are growing almost 6 times faster than the MAT reading skills of the eighth grade norm children at the same percentile, while the MAT reading skills of second grade children at the 10th percentile are only growing about 1½ times faster.

Table 4.3

TOTAL RATE OF CHANGE OF MEAN AND STANDARD DEVIATION
FOR FITTED STANDARD SCORES ON TOTAL READING

Grade	Months from Kindergarten	Monthly Rate of Change of $\hat{\mu}$ (1)	Rate of Change of $\hat{\mu}$ per nine months (2)	Montly Rate of Change of $\hat{\sigma}$ (3)	Rate of Change of $\hat{\sigma}$ per nine months (4)	Ratio of (3) to (1) (5)
1	14	.979	8.81	-.090	- .81	-.091
	20	.935	8.42	.065	.59	.071
2	26	.890	8.01	.133	1.20	.150
	32	.846	7.61	.140	1.26	.170
3	38	.802	7.22	.111	1.00	.14
	44	.757	6.81	.063	.57	.083
4	50	.713	6.42	.013	.12	.019
	56	.668	6.01	-.029	- .26	-.043
5	62	.624	5.78	-.056	- .50	-.087
	68	.560	5.04	-.065	- .59	-.117
6	74	.535	4.82	-.057	- .51	-.106
	80	.491	4.42	-.039	- .35	-.079
7	86	.447	4.02	-.019	- .17	-.042
	92	.402	3.62	-.012	- .11	-.030
8	98	.358	3.22	-.035	- .32	-.099
	104	.314	2.83	-.110	- .99	-.350
9	110	.269	2.42	-.262	-2.36	-.975

Table 4.4

RECIPROCAL OF THE RATE OF CHANGE PER NINE MONTHS OF FITTED STANDARD SCORES FOR SELECTED PERCENTILES AND GRADES

Percentile	Spring Second Grade (1)	Spring Eighth Grade (2)	Ratio of (2) to (1) (3)
10	.167	.244	1.461
20	.153	.273	1.784
30	.144	.299	2.076
40	.137	.325	2.372
50	.131	.353	2.695
60	.126	.387	3.071
70	.121	.432	3.570
80	.115	.500	4.348
90	.108	.640	5.926

Thus, the detailed answer to the question of whether the criterion is easier to achieve relative to some percentiles than to others must be given on a grade-by-grade basis. We have found two grades at which the answer is affirmative. It is not clear that one would desire a criterion of educationally significant growth to be harder to obtain relative to some percentiles than to others, especially if which percentile is harder depends on grade.

The preceding arguments regarding stringency are based on the assumption that the equipercentile assumption in conjunction with the norm data may be used to establish baseline measures of gains under average effort. In an earlier section, the validity of the equipercentile assumption was questioned, especially with regard to children in compensatory education programs. In addition, the norm data were not collected and the standard scores were not established with a longitudinal design. The longitudinal data presented in Section III, however, do support the conclusions regarding the differences in change in standard score across grade level. Therefore, use of the norm data does substantiate earlier evidence that the rate of growth in standard scores varies across grade level and to a much lesser extent across percentile ranks within grade.

The question is still open, though, as to what ought to be the relationship between normal growth and educationally significant growth to the extent that this relationship is an indication of equal stringency. The procedure by which the standard score scale is established has no implications with regard to the difficulty of obtaining gains in standard score. Furthermore, the currently used criterion for educationally significant growth has absolutely no basis in educational theory or statistical theory. We have shown above, however, that the currently-used criteria have a differential stringency under certain plausible assumptions.

Statistical Considerations

Statistically, the issue of stringency is more clear-cut, at least at a theoretical level. The stringency of a criterion from a statistical point of view may be expressed in terms of the power function that describes the chance of passing the criterion given a particular gain. The power function, however, can only give an impression of the stringency because its validity depends on the ideal conditions underlying the application of the statistical test such as random assignment, independence, etc.

For the norm-referenced procedure suggested by Horst et al. (1975), and adopted for use in the PIP evaluation, the power function may be expressed as:

$$P(T \geq t_{.025,\ N-1} | \delta)$$

where T is the test statistic described above in Section I; $t_{.025\ N-1}$ is the .025 critical point of students' t distribution with N-1 degrees of freedom; and δ is the noncentrality parameter expressed as $\frac{\Delta}{\sigma}\sqrt{n.}$ In this formulation, Δ is the true mean difference between the observed and expected posttest scores and σ is the standard deviation of the differences. Figure 4.1 shows the power curves for a few selected sample sizes. In all cases, the probability of passing the normal growth criterion given that Δ is zero is .025. This means, that if the average "population" gain in standard score is exactly what is expected under the normal growth assumption, then in only 25 out of 100 replications will the normal growth criterion be passed. This means that the procedure is rather stringent in that a program must produce impacts that exceed the equipercentile expectation to have a chance of passing the normal growth criterion. Of course, the stringency of concluding that a program did pass the criterion is matched by the stringency that a program did not pass the criterion.

For a given value of $\frac{\Delta}{\sigma}$, the chances of passing the criterion increase as the number of students in the analysis increases. For $\frac{\Delta}{\sigma}$ = .3, for example, the chance of passing the criterion is greater than 8 out of 10 when the number of students is about 100 and is less than 4 out of 10 when the number of students is about 30.

From a statistical point of view, it is plausible and reasonable, of course, to have a more stringent requirement for normal growth as the sample size decreases. However, in most field evaluations, the number of children in the various programs are not under the control of the evaluator. Therefore, the stringency of the criterion depends to some extent on such extraneous factors as the number of sites where a program was implemented, the number of children that were found at the sites, and the optimum number of children that could be accommodated in the operational design of the program. As a result, large disparities in sample size may be found. For example, in the 1975-76 PIP evaluation,

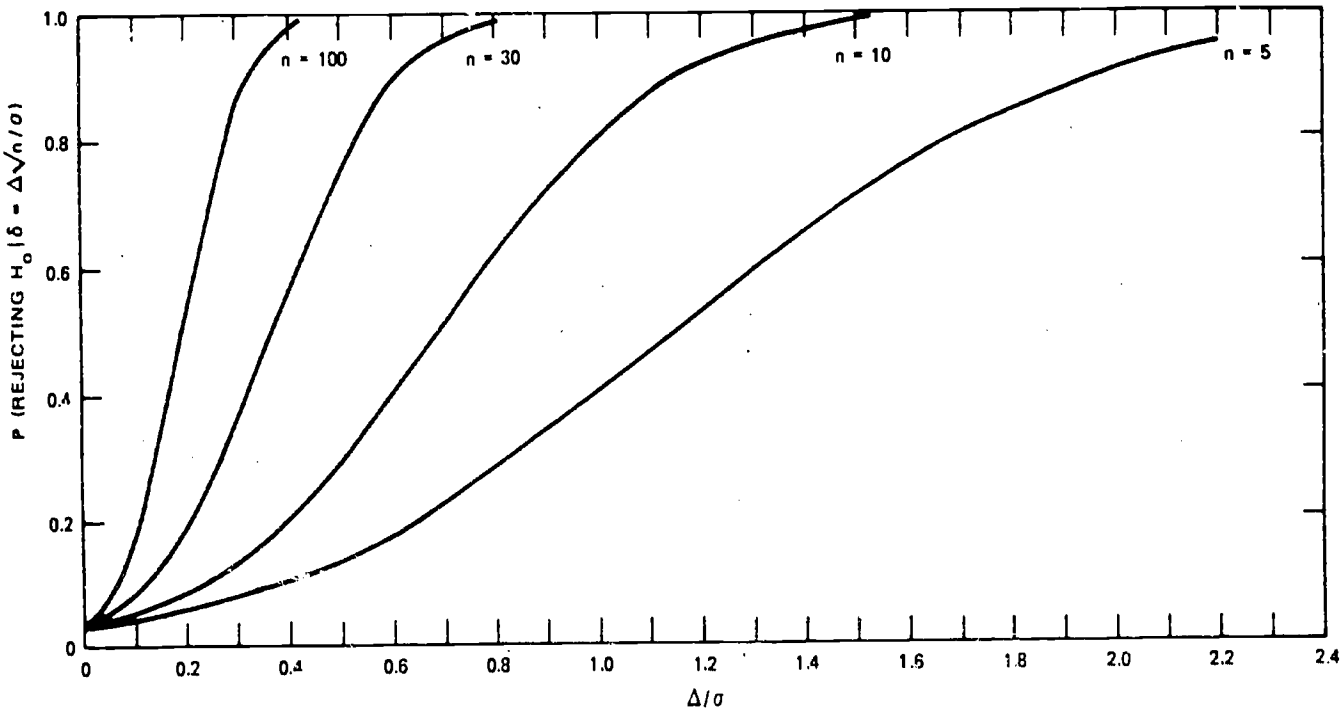


FIGURE 4.1 POWER CURVES FOR SELECTED SAMPLE SIZES, $\alpha = 0.025$

one PIP site had as few as 4 students and another had as many as 799 at the same grade level. Other things being equal, including the actual impact of a program on achievement, the program with the larger number of children has a much greater chance of demonstrating its effectiveness.

In the initial formulation of the norm-referenced procedure by Horst et al. (1975), the criterion of educational significance was added, apparently, in recognition that a gain could be statistically significant but not educationally significant. From this point of view, if a program passes what we have called the normal growth criterion, it has produced "statistically significant achievement gains (Horst et al., 1975)." Judgment as to educational significance did not depend on a statistical test, but was based simply on the magnitude of the difference between observed and expected posttest scores. This procedure would tend to protect the evaluator against the possibility of concluding that a program did have an impact based solely on statistical significance. The use of this procedure, however, magnifies the problem of determining the size of gains that are educationally significant. As we have stated several times, the one-third standard deviation criterion can be called crude at best.

V STATISTICAL PROPERTIES OF THE NORM-REFERENCED PROCEDURE

As was indicated in the introductory section, two questions were examined with regard to the statistical properties of the norm-referenced procedure:

- How sensitive is the test to the unit of analysis?
- How good an approximation is the standard deviation of the difference between the pretest and posttest scores to the standard deviation of the difference between the posttest and expected posttest scores?

The answers to both questions depend in part on the properties of the function that yields the expected posttest score given the pretest score that we have called the equipercentile function.

Properties of the Equipercentile Function

The equipercentile function, e(x), is defined as:

$$e(x) = f^{-1}[g(x)]$$

where g is the transform from pretest standard score to percentile rank; f^{-1} is the transform from posttest percentile rank to standard score.

Neither g nor f^{-1} are well-defined in a mathematical sense from the tables provided in the MAT Teacher's Handbooks (Durost, 1971). A percentile rank is not given for every standard score and every percentile rank does not have a corresponding standard score. The convention prescribed by the MAT developers for conversion from standard score to percentile when a standard score does not appear is to use the percentile rank of the next higher standard score. No convention was prescribed for the transformation from percentile to standard score. Furthermore, in the norm-referenced analysis, the mean standard score of students in a class or site is used. Because the pretest mean is generally not a whole number, a convention must be adopted for using the norm tables to derive the expected posttest mean score. (This was handled in the PIP evaluation by fitting a curve through the points defined by the standard score to percentile rank conversion tables.) In general, these are relatively minor technical problems, but in some cases the results of the analysis could depend on the conventions used to define the equipercentile function.

In Section III, it was shown that if pretest and posttest standard scores are both normally distributed in the norm population, then the equipercentile function is linear in the pretest standard score:

$$e(x) = a + bx,$$

where a and b are constants depending on the means and standard deviations of the two distributions. With a longitudinal design, if the pretest and posttest standard scores have a bivariate normal distribution, the expected posttest score given the pretest score is also a linear function, but with different coefficients.

The standard score to percentile conversions printed in the MAT Teacher's Handbook (Durost, 1971) were empirically derived, however. That is, they are based on the distribution of standard scores observed in the norm group and not on the theoretical distribution of standard scores used to derive the standard score scale. Furthermore, the standard score scale was derived using the fall test results only.

Although the equipercentile functions are not exactly linear they are very close to being so (see for example, Figures 3.1 through 3.8 in Section III). Table 5.1 gives the regression statistics for a linear fit of the equipercentile function by grade and test. Only points in the norm tables with a standard score and percentile score for both fall and spring were included in the analysis. In every case the percent of variation explained by the regression, the square of the correlation coefficient, was 93% or more, and in most cases it exceeded 99%. The fit appears to be a bit worse in the lower grades than in the upper grades, but not by much.

Because the equipercentile function is approximately linear, changing the unit of analysis for the numerator of the test statistic would have little if any impact on the analysis, all other things being equal. However, as was indicated in Section III, the equipercentile function appears to be inappropriate for use in evaluations of programs targeted for disadvantaged children. If the alternative for the equipercentile function is not linear, then the question of the sensitivity of the procedure to the unit of analysis remains open.

The Variance Estimate

It is implied in the norm-referenced procedure that the test statistic, T, defined in Section I has a central t distribution under the null hypothesis that there is no difference between the mean posttest standard score and the expected mean posttest standard score under the normal growth assumption.

The central t distribution is defined as the quotient of a standardized normally distributed random variable and the square root of the quotient of an independently distributed χ^2 variable and its degrees of freedom. For the usual application, this would be:

$$t = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \Bigg/ \sqrt{\frac{(n-1)s^2}{(n-1)\sigma^2}} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where $\bar{x}$ and s are the sample mean and standard deviation of n random

Table 5.1

REGRESSION STATISTICS FOR FIT OF EQUAL PERCENTILE FUNCTION BY A STRAIGHT LINE BY GRADE AND TEST: FALL TO SPRING

Grade		Test: Word Knowledge	Reading	Total Reading	Math Computation	Math Concepts	Math Prob.Solv.	Total Math
1	r^2	-	.952	-	-	-	-	.975
	slope	-	2.00	-	-	-	-	1.23
	intercept	-	-5.24	-	-	-	-	16.30
2	r^2	.979	.990	.989	.964	.981	.934	.985
	slope	.910	1.02	.95	1.09	1.11	1.00	.90
	intercept	13.2	-4.87	10.1	6.98	7.80	13.5	15.84
3	r^2	.993	.991	.990	.987	.998	.993	.989
	slope	1.08	1.16	1.07	.978	1.03	1.02	1.03
	intercept	- .765	-5.78	- .31	9.78	5.54	5.68	6.41
4	r^2	.995	.997	.999	.991	.996	.993	.996
	slope	.986	1.00	1.02	1.07	.967	1.02	1.00
	intercept	5.55	4.22	3.63	4.01	7.91	4.51	7.21
5	r^2	.995	.994	.998	.992	.994	.996	.995
	slope	1.03	.908	.95	1.10	1.13	1.01	1.07
	intercept	1.76	10.4	7.60	- 1.54	- 5.65	2.43	- .86
6	r^2	.992	.996	.998	.994	.994	.998	.996
	slope	1.04	.869	.94	1.06	1.19	.950	1.07
	intercept	.707	13.4	8.14	- .71	-10.2	6.85	- 1.35
7	r^2	.995	.996	.997	.997	.987	.997	.999
	slope	1.06	1.06	1.08	1.09	1.04	1.02	1.04
	intercept	- 3.19	-3.01	- 4.58	- 6.27	- 1.55	1.23	- 1.85
8	r^2	.997	.996	.997	.989	.992	.997	.998
	slope	1.95	1.03	1.04	1.13	1.08	.977	1.08
	intercept	1.00	-1.37	- 1.58	-11.8	- 5.89	3.94	- 6.80
9	r^2	.991	.994	.996	.984	.992	.993	.997
	slope	.992	.969	.99	1.05	1.07	1.04	1.03
	intercept	3.34	5.44	2.90	- 2.72	- 4.87	- 1.00	- 2.64

variables from a normal distribution with mean μ and standard deviation σ.

If we assume that the corresponding random variables in the norm-referenced procedure are the differences between the observed and expected posttest scores and e(x) is linear, then the test statistic would be

$$T' = \bar{Y} - e(\bar{x}) \Big/ \sqrt{\frac{s_y^2 + b^2 s_x^2 - 2brs_x s_y}{n}}$$

where s_x is the sample standard deviation of the pretest standard scores,
s_y is the sample standard deviation of the posttest standard scores,
r is the sample correlation between the pre- and posttest standard scores,
and $e(x) = a + bx$.

This formula differs from that given in Horst et al. (1975) by the presence of b and by n rather than n-1 in the denominator. The latter difference was probably a typographical error in the RMC text and is of little consequence for n moderately large.

The effect of b for the MAT subtests depends on the values of the sample standard errors and correlation:

$$s_y^2 + b^2 s_x^2 - 2brs_x s_y > s_y^2 + s_x^2 - 2r_x s_y \qquad (1)$$

$$(b^2 - 1)s_x^2 > 2rs_x s_y (b - 1)$$

$$(b + 1)(b - 1) > 2rs_y (b - 1)$$
$$\frac{}{s_x}$$

A reexamination of Table 5.1 shows that for the MAT, as would be expected of any reasonable test, b is in the neighborhood of 1 and is always positive.

For $b > 1$, we have:

$$(b + 1) > 2rs_y$$
$$\frac{}{s_x}$$

as the condition for Equation (1) to be true. If the ratio of s_y to s_x is approximately one, then $b + 1$ will always be greater than $2r$ when $b > 1$. Therefore, it would appear that substituting the left-hand side of Equation (1) into the t statistic of the norm-referenced procedure would result in a slightly smaller t value. Because the values of b

shown in Table 5.1 are very close to 1 in most cases, it appears that the effect of using the denominator recommended by Horst et al. (1975) rather than the one on the left-hand side of Equation (1) would be inconsequential.

Empirical Results

Four sets of data on Project Catch-Up from the first-year PIP evaluation were reanalyzed to assess the impact of modifications in the statistical procedure on the results of the norm-referenced analysis. The transformation from fall score to expected spring score was applied at the student level rather than at the site level. The mean and variance of the difference between the student observed and expected spring scores were used in the calculation of the t statistic. The results are summarized in Table 5.2.

Table 5.2

COMPARISON OF RESULTS OF ORIGINAL AND MODIFIED NORM-REFERENCED PROCEDURES

			Original Procedure				Modified Procedure			
Grade	Test	n	Gain Over Expected Gain	S.D. Post-Pre	t Test	Meets Normal Growth	Gain Over Expected Gain	S.D. Post-Pre	t Test	Meets Normal Growth
3	Total Reading	18	1.60	4.33	1.56	Unknown	.94	5.04	.79	Unknown
5	Total Math	22	7.14	5.67	5.91	Yes	3.15	5.04	2.93	Yes
5	Total Reading	19	1.47	4.13	1.55	Unknown	2.05	4.70	1.90	Unknown
6	Total Math	27	1.11	6.77	.85	Unknown	1.06	6.70	.82	Unknown

For these four sets of data, the numerator of the t statistic, gain over expected gain, tends to be lower under the modified procedure than under the original procedure and the donominator tends to be the same under either procedure. As a result, the t values tend to be smaller under the modified procedure. In all of these cases, the conclusion regarding normal growth would have been the same under either procedure; however, in situations where normal growth is only narrowly achieved under the original procedure, it may not be achieved under the modified procedure.

Again, the results of this section indicate that the norm-referenced procedure is relatively insensitive to changes in the unit of analysis or the denominator of the test statistic given the equipercentile assumption of normal growth. If a different normal growth assumption were made so that the expected postest standard score function was not linear or did not have a slope close to 1, the the issues regarding the statistical nature of the procedure would need to be reexamined.

VI CONCLUSIONS

We have examined criticisms of the norm-referenced procedure at a general policy level and at a technical level. At the policy level the question is, "Why use standardized tests at all?" By using standardized tests, the evaluator is usually establishing a criterion of program impact that is removed from specific program goals and content. Under such circumstances, the evaluation question is not so much whether a program succeeded in achieving its intended objectives, but whether it achieved some externally specified goal, often specified after the program has been developed and implemented.

For policy purposes, the use of standardized tests is reasonable to the extent that it can be demonstrated that the content of the tests are related to specific policy goals. If the only policy goal is to determine whether a program achieved certain specific educational objectives, no standardized achievement test might be adequate. On the other hand, the evaluator and policy-maker need to be sensitive to the hazards of establishing extremely general goals such as "stimulating cognitive growth" or "improving reading skills" and then using standardized tests indiscriminantly to establish criteria for meeting these goals. It is necessary to at least examine specific program objectives and content to determine their relationship to the general policy goals and to recognize the unique aspects of a program in the process of selecting measures of program impact.

As a rather extreme example, let us say that a policy-maker specifies that the goal of any program in math should be the improvement of students' skills in math and that the criterion for program success is improvement on MAT Total Math as demonstrated in a norm-referenced analysis. If he now tries to evaluate a program targeted for eighth grade students he may or may not notice that based on the equipercentile model expected growth for this group between fall and spring is negligible (see Figure 3.8). Does this mean that math skills for eighth grade students are not expected to improve for students in conventional education programs? Or does it mean that the content of the test is irrelevant to what is being taught at this grade level?

The need for congruence between test content and program content has often been stated, but the process of specifying either policy or program objectives to the point where they may be used in test development or selection is difficult and tedious. Nevertheless, the process is necessary if standardized tests are to be used rationally. Also, policy-makers and evaluators need to face the possibility that for many educational programs standardized tests may not be an appropriate vehicle for evaluating impacts. At a minimum, under such circumstances, it is essential to distinguish between the degree of success of a program in accomplishing its own objectives and the degree of success in accomplishing some externally imposed objectives. The degree

of agreement between the two sets of objectives would have policy implications totally apart from the results of a norm-referenced analysis. In fact, if educationally significant gains on a specific standardized test is stated as a policy objective, then perhaps the most reasonable approach would be to foster educational programs with that as their explicit goal.

At the technical level, a number of potential weaknesses of the norm-referenced procedure were identified. These include:

- On some tests, the expected posttest standard score based on the equal percentile assumption is too low for students with extremely low pretest scores.
- There are indications that the expected posttest standard scores are too high for disadvantaged students, especially disadvantaged minority students.
- The criterion for educationally significant growth may not be of uniform stringency across grade levels.
- The stringency of the criteria for normal growth and educationally significant growth depend on the number of students in the evaluation.

This list does not touch upon some of the more basic criticisms of the norm-referenced procedure that were discussed above. If the evaluator agrees that the conceptual basis for the norm-referenced analysis is extremely weak, then no minor modifications of the procedure will be satisfactory. Nevertheless, to complete the description of the results of the norm-referenced analysis, the data in the 1975-76 PIP evaluation for fourth grade Total Reading were reexamined using a modified procedure. Fourth grade was selected because it was the only grade where the same battery level of the MAT had been administered to the MAT standardization group, the Compensatory Reading evaluation group, and the PIP evaluation group. The CR/SL group within the Compensatory Reading evaluation group was used to derive the function describing the expected posttest score given the pretest score. A separate function was derived for each of three ethnic/racial groups: whites, blacks, and Spanish surname. A regression analysis was used to derive the function in a pretest standard score range between 46 and 88, corresponding to a percentile range between 6 and 94. The summary statistics are presented in Table 6.1.

Table 6.1

SUMMARY STATISTICS FOR ESTIMATION OF EXPECTED POSTTEST TOTAL READING STANDARD SCORE GIVEN PRETEST STANDARD SCORE

Race/Ethnicity	n	Intercept	Slope	r^2	S.D.	Standard Error (Slope)
White	1238	3.32	1.04	.69	6.2	.02
Black	619	-1.60	1.11	.57	5.8	.04
Spanish surname	179	8.31	.93	.63	5.8	.05

We believe that the functions still over estimated posttest standard scores because of the possibility that the time interval between pre- and posttest may have been longer than the interval for the PIP evaluation by as much as a month. Also, these children had been in some form of compensatory reading program. It is not clear whether children who were in the PIP program would have been in other compensatory programs if the PIPs had not been implemented. If the children would have been in other compensatory programs, such as those funded under Title I, then the CR/SL group would appear to be a good comparison group. Otherwise, some adjustment to the functions would need to be made to account for the gains attributable to the compensatory reading program. For the current purposes, each intercept was decreased by 1.5 standard score points to represent the effect of about a month's increase in the interval between pre- and posttest and the effect of the compensatory reading program. Pelavin and Barker (1976) proposed a rate of growth of about .75 standard score points per month as a rule of thumb in their study of the MAT. Most studies of compensatory education program have found only small effects attributable to the program. Therefore, a cumulative effect of about 1.5 standard score points would appear to be reasonable for the purpose of this study.

For students below the 6 percentile lower bound, a constant expected spring score was postulated for each racial/ethnic group. Prior results (see Figure 3.2 and 3.17) had indicated that the expected spring score for a student with fall standard scores below the sixth percentile was approximately independent of the particular pretest score. The expected posttest score was found to be approximately equal to the regression lines evaluated at a pretest standard score of 45. A similar procedure was used for students with scores above the 96th percentile.

Figure 6.1 shows the relationship between two of the three modified normal growth curves and the equipercentile curve. Between a pretest standard score of about 43 and 75, corresponding to a percentile range between 5 and 74, the equipercentile curve lies above the modified

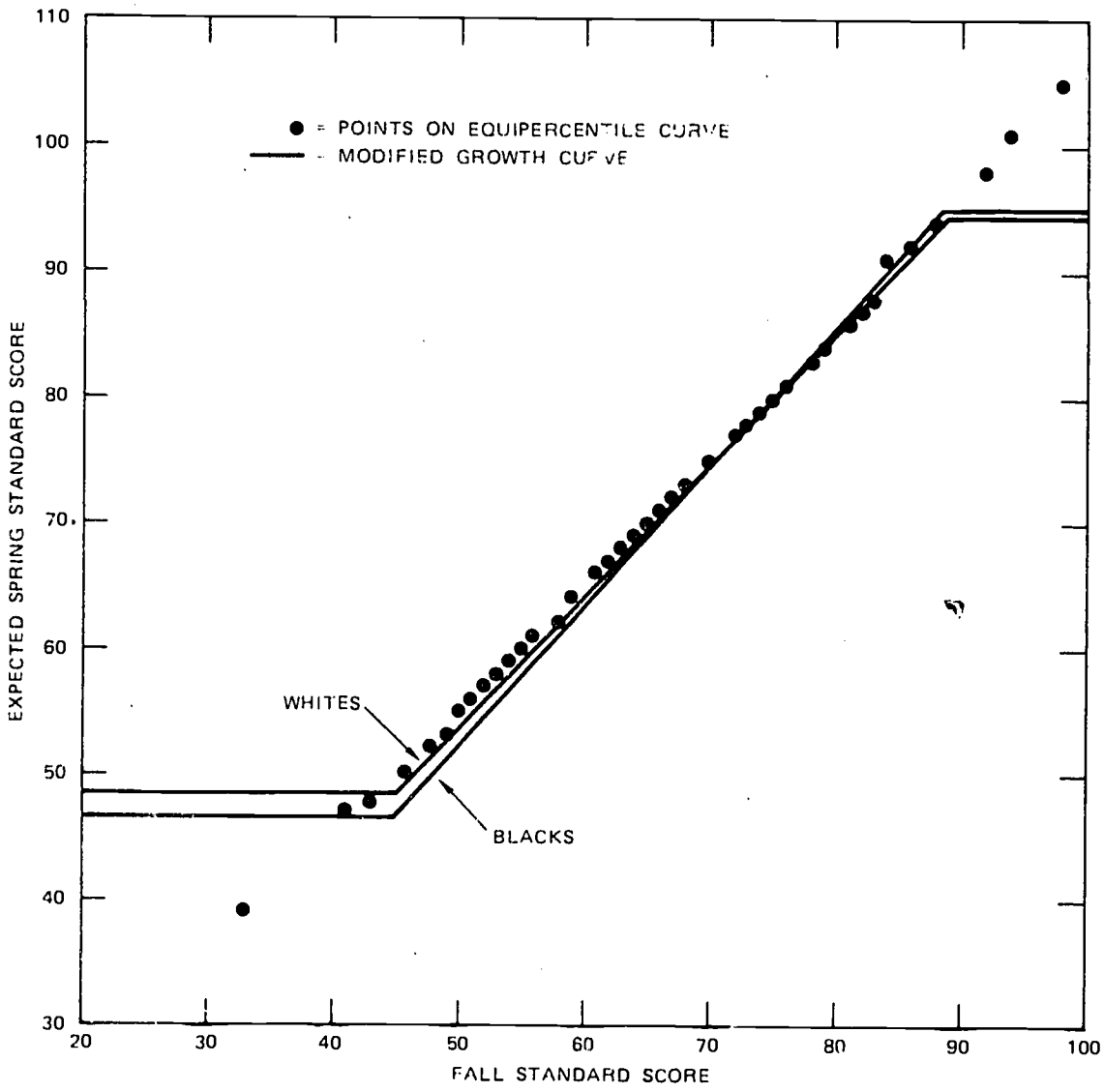


FIGURE 6.1 COMPARISON OF EQUIPERCENTILE GROWTH CURVE TO CURVES USED IN THE MODIFIED ANALYSIS: GRADE 4 TOTAL READING

curves. Between a pretest standard score of 43 and 62 (between the 5th and the 40th percentile), the difference between the modified and original curves is at least one standard score point. Below a pretest score of 43 the equipercentile and modified growth curves diverge so that the modified curves are substantially above the equipercentile curve. The modified curve for black students is two standard score points below the curve for white students for the lowest pretest scores. The divergence between the curves for whites and blacks decreases to zero at a pretest score of about 70, corresponding to the 62nd percentile.

The student was used as the unit of analysis. Each student's fall score, F, was used to derive an expected spring score, E. The difference, D = S - E, between the observed spring standard scores S and the expected spring scores E was then calculated. These differences were used to calculate the t statistic

$$T = \frac{\bar{D}}{SD/\sqrt{n}}$$

where $\bar{D}$ = mean difference
SD = standard deviation of the differences
n = number of students

Under the assumption that the deviations between observed and expected scores are approximately independent and normally distributed, T will have a Student's t distribution with n-1 degrees of freedom.

Table 6.2 gives the results of the original and modified analysis with respect to the normal growth criterion for fourth grade Total Reading. No major change in conclusions regarding normal growth is evident. Only one site, Wayne City, had a change from "no" to "unknown". The gains over what was expected increased under the modified procedure for all projects, but at a few sites the gains decreased because of the increase in expected spring scores for extremely low fall scores. For the 44 pupils scoring below the sixth percentile in the fall, the average gain was 8.4 standard score points, with a mean spring standard score of 48.18. For the remaining 181 pupils scoring at the sixth percentile or above, the average gain was only 3.3 standard score points.

If the test for educationally significant growth is carried out on the difference scores calculated in the modified analysis and maintaining the 1/3 σ criterion, the conclusions are almost identical to those made in the original analysis. The only change occurred at two Catch-Up sites, Brookport and Galax, where "unknown" was superseded by the conclusion that the gains did not meet the criterion for educational significance.

From the above results, it would appear that modifications in the procedure would tend to be inconsequential with regard to the

Table 6.2

RESULTS OF ORIGINAL AND MODIFIED NORM-REFERENCED PROCEDURE FOR FOURTH GRADE TOTAL READING BY PIP AND SITE

PIP/Site	No. Pupils	Gain Over Fall*	Original Analysis: Gain Over Expected*	Original Analysis: t Test	Original Analysis: Meets Normal Growth	Modified Analysis: Gain Over Expected*	Modified Analysis: t Test	Modified Analysis: Meets Normal Growth
Catch-Up	83	2.96	-2.04	-3.46	No	-1.48	-2.81	No
Bloomington	40	2.73	-2.28	-2.70	No	-1.75	-2.23	No
Brookport	8	3.75	-1.25	- .75	Unk	.01	.01	Unk
Galax	7	4.43	.43	.11	Unk	-2.49	- .95	Unk
Providence Forge	20	3.05	-1.95	-2.08	Unk	- .94	- .91	Unk
Wayne City	8	1.88	-2.25	-2.64	No	-2.11	-2.36	Unk
Conquest	108	5.56	1.08	1.98	Unk	1.24	2.46	Yes
Benton Harbor	28	7.36	3.19	2.75	Yes	3.45	3.53	Yes
Cleveland	53	4.77	.49	.65	Unk	.80	1.17	Unk
Gloversville	27	5.22	.22	.22	Unk	- .18	- .17	Unk
IRIT	34	3.44	-1.23	-1.39	Unk	-1.18	-1.58	Unk
Bloomington	28	4.39	.12	.12	Unk	- .46	- .56	Unk
Schenectady	6	-1.00	-6.00	-6.71	No	-4.56	-4.94	No

*In standard score units

conclusions of the norm-referenced analysis for the type of sites included in the PIP evaluation. There may be a few marginal cases where the conclusion would change one way or the other, but the pattern of conclusions remains substantially the same.

This is not surprising because the modifications appear relatively mild and they are not all in the same direction. The modifications in the normal growth curve, for example, increased the expected posttest scores for extremely low pretest scores and decreased the expected posttest scores in the moderately low-to-average range of pretest scores. These modifications, then, may have cancelled each other out in many cases. The changes in the unit of analysis and the denominator of the test statistic would appear to have a minor impact, if any, because the modified normal growth curves were linear over a broad range of pretest scores.

In summary, we list here some of the issues we have examined regarding the norm-referenced analysis and the results and conclusions of our study.

Use of Standardized Tests in Educational Evaluations

We agree with other critics of standardized tests that standardized tests should be used selectively, if at all. The demonstrated correspondence of test content with program or policy objectives is of utmost importance. If such correspondence cannot be shown, we feel that no basis for the use of standardized tests exists.

The Standard Score Metric

We were concerned about the cross-sectional standardization design and other details of the standardization procedure used by the MAT developers. Some peculiar properties of the equipercentile growth curves, in particular the "summer growth" phenomenon, lead us to conclude that the cross-sectional norms are probably not adequate for predicting longitudinal gains that span more than one grade level.

The Equipercentile Normal Growth Assumption

The equipercentile normal growth curves that yield the expected posttest score given a pretest score were shown to be approximately linear. The equipercentile growth curves were a relatively good fit to the empirical growth curves (mean posttest score given pretest score) for the MAT longitudinal subgroups, except for extremely high and low pretest scores.

Examination of data from the Follow Through Evaluation revealed that NFT students with pretest scores that were neither extremely high nor extremely low tended to drop in percentile rank from one spring test period to the next. The drop in percentile rank was not found for groups in the Compensatory Reading Evaluation, but gains in

percentile rank might be explained by time of testing, the type of students in the groups, and the effect of the compensatory programs. In practically all cases examined in the NFT and CR evaluations, the empirical curves for minority students were consistently below the corresponding curves for white students. Because other factors were not being controlled, the differences certainly cannot be attributed solely, if at all, to minority status. However, the results do indicate that the equipercentile assumption of normal growth would be less applicable to groups of disadvantaged minority children than to groups of disadvantaged white students.

Although the straggler hypothesis, that students for whom compensatory programs are targeted tend to lose in percentile rank from pre- to posttest, tended to be confirmed, it appeared in subsequent analysis that the drop was not substantial enough to change the results of the norm-referenced analysis for the PIP data we examined. This was at least partially caused by the offsetting effect that students with extremely low pretest scores tended to have larger than expected standard score gains on the posttest. Indications from the NFT data, however, were that percentile ranks continued to decline over time between first and third grade. This means that for longitudinal programs where the period between pre- and posttests is more than one school year, the phenomenon of declining percentiles will have a more serious effect on the analysis. This, coupled with our concerns regarding the validity of the norms as longitudinal predictors across grade level, leads us to the conclusion that the norm-referenced procedure should not be used in evaluations of programs that extend over more than one school year.

Stringency and the Educationally Significant Growth Criterion

Although we can conceptualize what it means for a procedure to have uniform stringency across grade levels, pretest scores, and subject matter, we have found it extremely difficult to translate the concept into operational terms. Starting from the point of view of using the normal growth criterion as a baseline and assessing the educationally significant growth criterion against this baseline, however, we found that the educationally significant growth criterion (using the MAT) was not uniformly stringent across grade levels.

This technical problem with the educationally significant growth criterion, however, is rather superficial relative to its conceptual problems. Horst et al. (1975) stress that the $1/3\sigma$ criterion is merely a rule of thumb with no basis in either statistical or educational theory. We believe that the lack of a conceptual foundation for the educationally significant growth criterion, paired with the technical problems of its potential lack of uniform stringency, call for abandoning this facet of the norm-referenced procedure. Rules of thumb have a tendency to become established practice over time. The negligible basis for this particular rule calls for its abandonment before it is widely accepted as the authoritative approach to assessing educational significance.

The Statistical Properties of the Procedure

Several aspects of the statistical procedure were examined and the impacts of modifications were assessed. The modifications by themselves and in conjunction with changes in the equipercentile model of normal growth appeared to have very little impact on conclusions at least with regard to projects such as those in the PIP evaluation.

REFERENCES

Armor, D., P. Conry - Oseguera, M. Cox, N. King, L. McDonnell, A. Pascal, E. Pauly, G. Zellman, "Analysis of the School Preferred Reading Program in Selected Los Angeles Minority Schools," prepared for the Los Angeles Unified School District (RAND Corporation, Santa Monica, California, August 1976).

Barker, P. and S. Pelavin, "Conserving Scores and Scale Transformations in Standardized Achievement Tests, Their Accuracy and Dependability for Individual and Aggregation: The Case of MAT70," a working note prepared for the National Institute of Education, WN-9161-NIE, (RAND Corporation, Santa Monica, California, 1975).

Beck, M. D., "Development of Empirical 'Growth Expectancies' for the Metropolitan Achievement Tests," paper presented at the 1975 convention of the National Council on Measurement in Education, Washington, D.C., 31 March 1975.

Bianchini, J. C., "Achievement Tests and Differentiated Norms," a paper presented at USOE Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation held in Washington, D.C., May 1976.

Coleman, J. S., et al., Equality of Educational Opportunity, U.S. Department of Health, Education, and Welfare, Office of Education (Government Printing Office, Washington, D.C., 1966).

Cooley, W. W. and P. R. Lohnes, Evaluative Inquiry in Education, Irvington Publisher, Inc., New York, New York, 1976.

Dixon, W. J. (editor), Biomedical Computer Programs, (University of California Press, Berkeley, California, 1973).

Durost, W. N., H. H. Bixler, J. W. Wrightstone, G. A. Prescott, and I. H. Balow, Metropolitan Achievement Tests Teacher's Handbook, from each MAT test battery (Harcourt Brace Jovanovich, New York, New York, L971).

Gamel, N., G. K. Tallmadge, C. T. Wood, J. L. Binkley, "State ESEA Title I Reports: Review and Analysis of Past Reports, and Development of a Model Reporting System and Format," prepared for USOE (RMC Research Corporation, Mountain View, California, October 1975).

Hoepfner, R., "Achievement Test Selection for Program Evaluation," a paper presented at the USOE Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, held in Washington, D.C., May 1976.

Horst, D. P., G. Tallmadge, and C. Wood, "A Practical Guide to Measuring Project Impact on Student Achievement," U.S. Department of Health, Education, and Welfare, Office of Education, Washington, D.C. (1975).

Jencks, C., Inequality: A Reassessment of the Effect of Family and Schooling in America (Basic Books, Inc., New York, New York, 1972).

Lord, F. M., and M. R. Novick, Statistical Theories of Mental Scores (Addison-Wesley Publishing Co., Inc., Reading, Massachusetts, 1968).

MAT Guidelines No. 1, Development of the Standard Score System for the 1970 Edition of Metropolitan Achievement Tests (Harcourt Brace Jovanovich, Inc., New York, New York, October 1972).

Mayeske, G. W., and A. Beaton, Jr., Special Studies of Our Nation's Students, (Government Printing Office, Washington, D.C., 1975)

Pelavin, S., and P. Barker, A Study of the Generalizability of the Results of a Standardized Achievement Test," based on the RAND Corperation study for the National Institute of Education under Contract B2C-5326, paper presented at the American Educational Research Association Meeting, San Francisco, California, 19-23 April 1976.

Prescott, G. A., Metropolitan Achievement Tests Manual for Interpreting (Harcourt Brace Jovanovich, New York, New York, 1973).

Stearns, M., Evaluation of the Field Test of Project Information Packages, Volume I: Viability of Packaging, prepared for USOE under contract OEC-0-74-9256 (Stanford Research Institute, Menlo Park, California, 1975).

Trismen, D. A., M. Waller, G. Wilder, A Descriptive and Analytic Study of Compensatory Reading Programs, Volume I of Final Report, prepared for USOE under contract No. OEC-0-71-3715 (Educational Testing Service, Princeton, New Jersey, 1975).

Tyler, R. W., "Discussion of Hoefner's Paper," a paper presented at the USOE Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, held in Washington, D.C., May 1976.

Wargo, M. J., "Achievement Testing of Disadvantaged and Minority Students for Education Program Evaluation: an Evaluator's Perspective," a paper presented at the USOE Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, held in Washington, D.C., May 1976.