

LIBRARY
Michigan State
University

This is to certify that the

dissertation entitled

EXAMINATION OF THE USDE

NORM-REFERENCED EVALUATION MODEL

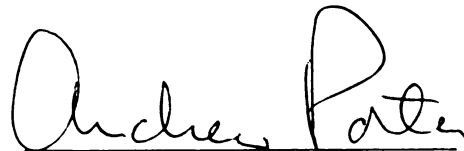
presented by

Irene Mary Leland

has been accepted towards fulfillment

of the requirements for

Ph.D. _____ degree in _____ Educational Psychology



Major professor

Date 11/11/87



RETURNING MATERIALS:
Place in book drop to
remove this checkout from
your record. FINES will
be charged if book is
returned after the date
stamped below.

~~03-15-90~~
4-1-92

EXAMINATION OF THE USDE
NORM-REFERENCED EVALUATION MODEL

by

Irene Mary Leland

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology
and Special Education

1987

Copyright by

IRENE MARY LELAND

1987

ABSTRACT

EXAMINATION OF THE USDE NORM-REFERENCED EVALUATION MODEL

By

Irene Mary Leland

The norm-referenced model replaces the control group with the norming population of a standardized test. It assumes natural growth follows the pattern observed in this norming population. The model lacks validity when local patterns of growth differ systematically from those of the norming population. Due to its wide usage in evaluating Chapter 1 Projects there remains a need for additional research to provide a clearer understanding of the model. Determination of whether local conditions result in patterns of growth sufficiently different from the national norms to invalidate the model is needed.

This study approached the norm-referenced model as implying a model of growth in measured achievement over successive school years. Patterns of growth implied by the model and the norms for two standardized tests were compared.

Local data from six school districts selected for differences in use of out-of-level testing, effectiveness of overall school program and proportion of low SES students were examined. Local patterns of growth were used to test the norm-referenced model. Consequences of differences

between local and national growth patterns on apparent gains, Type I errors and power were determined. For two districts longitudinal data were compared with expectations based on cross-sectional norms.

The patterns of growth implied by the norm-referenced model and norms of the standardized tests studied were found to be curvilinear with the rate of growth highest in the early grades. Significant differences were found in the growth curves for the two tests.

The model is robust with respect to differences in percentile rank. Significant differences between tests were found at Grade 6 in reading. In mathematics significant differences were found for specific grade levels and districts.

Local patterns of growth affect both Type I errors and power. Small true gains are unlikely to be detected at the local level but educationally significant gains usually will be. Results of the longitudinal analyses indicate at least some cases where local longitudinal data do not match the results of cross-sectional data for the national or local populations.

This study identifies several areas that may cause bias which should definitely be considered when interpreting local data.

Dedicated to my parents
Walker M. and Emily S. Dawson

ACKNOWLEDGEMENTS

I would like to acknowledge all of those who have helped me in this endeavor. First, I would like to thank Dr. Andrew C. Porter who served as my major advisor and as chairman of my committee. His encouragement enabled me to achieve more than I would have thought possible before I started this program. Then I want to thank the members of my committee: Dr. Norman T. Bell who encouraged me to believe I could undertake and complete a doctoral program at this point in my life; Dr. William W. Farquhar who helped me to relate the research methods I studied in my coursework to my work in evaluation at the Michigan Department of Education; Dr. Roy V. Erickson who helped with the statistical problems I encountered in the analyses for this dissertation and Dr. Lawrence I. O'Kelly who was always willing to listen when I needed to clarify my thoughts on some problem.

Next I want to thank all my coworkers at the Michigan Department of Education who encouraged me in this undertaking. Special thanks go to Dr. David L. Donovan and Dr. Daniel E. Schooley. Without their encouragement and support I could not have completed this program.

I also want to thank Ms. Caryl Basel, Ms. Marsha Bowers, Mr. LaVerne Dittenber, Mr. George Johnson, Dr. Franci Moorman and Mr. Donald Peters for their assistance in collecting the data for this study.

Then I want to thank my family for their understanding and support during the time I have been working on this project. Without their help, especially that of my two youngest children, Robert and Carolyn, I could not have completed this endeavor.

And finally, I want to thank my many friends who have encouraged me in this undertaking.

TABLE OF CONTENTS

	<u>Page</u>
LIST OF TABLES	x
LIST OF FIGURES.	xi
 Chapter	
I. INTRODUCTION.	1
Need for this Research	3
Overview of the Study.	3
Definitions of Important Terms	5
II. THE NORM-REFERENCED MODEL AND ITS ASSUMPTIONS ABOUT NATURAL GROWTH.	7
III. REVIEW OF THE LITERATURE.	10
IV. PROCEDURES	24
Development and Comparison of Patterns of Growth Using the Assumptions of the Norm-referenced Model	24
Effect of Variables Hypothesized to Affect the Norm-referenced Model's Validity at the Local Level	30
Selection of Districts	31
Collection and Processing of Local Data.	33
Development of Growth Curves Based on Local District Data.	35
Effects of Local District Growth Patterns on the Validity of the Norm-referenced Model	36
Determination of No-treatment Gains	36
Probability of Type I Errors.	38
Probability of Detecting True Gains	41
Comparison of Longitudinal Data with Expectations Based on Cross-sectional Norms	42

Chapter	Page
V. RESULTS OF THE STUDY.44
Comparison of Growth Curves for the CAT and MAT44
Analysis of the Growth Curves for the CAT and MAT.54
Comparison of National and Local Norms55
Stability of Local Norms63
Effect of Local Growth Patterns on Gains Measured by the Norm-referenced Model63
Effects on Expected No-treatment Growth .63	
Effects on Likelihood of Type I Errors. .68	
Effects on Power to Detect True Gains . .74	
Effects on Size of Sample Needed to Detect True Gains.77
Comparison of Longitudinal Data with Cross- sectional Norms77
VI. DISCUSSION81
Assumptions the Norm-referenced Model Makes About Natural Growth.81
Comparison of Growth Curves for the CAT and MAT82
Robustness of the Norm-referenced Model with Respect to Variations Found in Local Patterns of Growth83
Stability of Local Growth Patterns from Year to Year86
Adequacy of Cross-sectional Norms for Estimating Longitudinal Patterns of Growth. .87	
Implications of These Findings88
Areas Needing Additional Research.90
VII. SUMMARY AND CONCLUSIONS93
Conclusions.94
Appendix	
A. GRAPHS OF LOCAL GROWTH CURVES97
B. YEARLY DISTRICT DATA.	109
REFERENCES	113

LIST OF TABLES

Table	<u>Page</u>
1. Summary of Powers'(1983) Data18
2. Total Number of Students Tested in 1983, 1984 and 198534
3. Reading 4-way ANOVA Summary37
4. Mathematics 4-way ANOVA Summary38
5. CAT and MAT Scale Scores, Reading and Mathematics.45
6. CAT and MAT Rescaled Scale Scores51
7. Differences Between CAT and MAT Growth Curves	.54
8. National and Local Growth Curves.56
9. Comparison of National and Local Percentile Ranks57
10. Size of District and Stability of Local Norms	.64
11. Significance of Expected Zero Gain Scores . .	.65
12. Effect on Power of Local Growth Patterns. . .	.75
13. Effect on Sample Size of Local Growth Patterns.	.78
14. Longitudinal Data79
15. Yearly District Data: CAT - Reading.	109
16. Yearly District Data: MAT - Reading.	110
17. Yearly District Data: CAT - Mathematics. . .	111
18. Yearly District Data: MAT - Mathematics. . .	112

LIST OF FIGURES

Figure	<u>Page</u>
1. Distribution of Hypothetical Group.40
2. CAT Reading Growth Curves46
3. MAT Reading Growth Curves47
4. CAT Math Growth Curves.48
5. MAT Math Growth Curves.49
6. Rescaled Reading Growth Curves: CAT vs MAT .	.52
7. Rescaled Math Growth Curves: CAT vs MAT. . .	.53
8. National and Local Growth Curves: CAT - Reading.58
9. National and Local Growth Curves: MAT - Reading.60
10. National and Local Growth Curves: CAT - Math	.61
11. National and Local Growth Curves: MAT - Math	.62
12. Expected Zero Gain Scores: Reading67
13. Expected Zero Gain Scores in Math: CAT69
14. Expected Zero Gain Scores in Math: MAT70
15. Expected Zero Gain Scores in Math: Out-of- Level Districts.71
16. Expected Zero Gain Scores in Math: Improving MEAP Districts72
17. Expected Zero Gain Scores in Math: Low SES Districts.73
18. COL Growth Curves: Reading97
19. CIM Growth Curves: Reading98

Figure		<u>Page</u>
20.	CLS Growth Curves: Reading99
21.	MOL Growth Curves: Reading	100
22.	MIM Growth Curves: Reading	101
23.	MLS Growth Curves: Reading	102
24.	COL Growth Curves: Mathematics	103
25.	CIM Growth Curves: Mathematics	104
26.	CLS Growth Curves: Mathematics	105
27.	MOL Growth Curves: Mathematics	106
28.	MIM Growth Curves: Mathematics	107
29.	MLS Growth Curves: Mathematics	108

EXAMINATION OF THE USDE NORM-REFERENCED EVALUATION MODEL

CHAPTER I

INTRODUCTION

A number of evaluation models have been suggested for use in estimating program effects when a randomly assigned control group is not feasible. In response to legislative mandate RMC Research Corporation developed several models for use in evaluating federally funded programs (Tallmadge and Wood, 1976). One of these, the norm-referenced model has since become almost universally used in evaluating Chapter 1* Projects (Reisner et al, 1982). The norm-referenced model replaces the control group with the norming population of a standardized test. It assumes that natural growth, (i.e. growth that would occur in the absence of any special program), follows the same pattern as that observed in the norming population.

The norm-referenced model lacks validity when local patterns of natural growth differ systematically from those of national norming populations. A number of conditions can affect the apparent growth observed in a local group (e.g. the similarity between local and national populations, local and national school practices, test scheduling and the match of the test to the local school

* Formerly Title I, in July 1982 ESEA Title I was replaced by ECIA Chapter 1. The models originally developed under Title I have continued to be used under Chapter 1.

curriculum) when using national norms. There is a need to determine whether local conditions actually do result in patterns of natural growth sufficiently different from those of the national norms to invalidate this method. It is important to estimate the effects of specific conditions and the amount of bias likely to be incurred under each. Local school district norms need to be compared to national norms to determine whether there is evidence of bias.

The norm-referenced model compares longitudinal program data with cross-sectional norming data. There is a need to determine whether cross-sectional data provide a useful measure of the growth patterns found in longitudinal data.

This study had a threefold purpose.

1. To examine the assumptions of the norm-referenced model with respect to the nature of natural growth and to determine whether the growth curves implied by different standardized tests are congruent.

2. To describe conditions other than program effects that would be expected to affect the gains actually observed, and to examine, using local district data, their effects on local patterns of growth and the implications for the validity of the norm-referenced model.

3. To compare the growth actually found in longitudinal data with that implied by cross-sectional norms.

Need for this Research

This study is important because the norm-referenced model has come to enjoy wide use for evaluating program effects. If it is found to have questionable validity then its popularity is not warranted. It is important to examine the robustness of the norm-referenced model with respect to local deviations from conditions pertaining to national norming populations. School practices and other conditions that can affect the pattern of local natural growth need to be identified. Examination of the validity of using cross-sectional norms to evaluate longitudinal data is also important.

Overview of the Study

The following research questions represent the primary thrust of the project.

1. What assumptions does the norm-referenced model make about natural growth? More specifically what patterns of natural growth are implied by the assumptions?

2. Are the growth curves implied by the model the same for all standardized tests?

3. How robust is the norm-referenced model with respect to variations found in local patterns of growth?

4. How stable are the local patterns from year to year?

5. How adequate are cross-sectional norms for estimating longitudinal patterns of growth?

From these questions the following plan for testing the robustness of the model was developed.

1. The patterns of growth implied by national norms for two tests would be determined.

2. Selected local districts would be examined for differences on factors believed to influence patterns of growth. Local patterns of growth implied by local norms for these districts would be determined.

3. Consequences of differences between local and national growth patterns on observed normal curve equivalent (NCE) gains would be calculated and the effect on Type I errors and power of tests for program effects determined.

4. For two sample districts patterns of growth based on longitudinal data would be compared with patterns of growth obtained using cross-sectional data.

The local data would be used to test the norm-referenced model. If the national norms provide a valid comparison group for students in local compensatory education programs the expected growth of the local students should be the same whether local or national norms were used. The research hypothesis is that students growing in achievement at the rate implied by the local norms would maintain the same percentile rank from year to year when evaluated with national norms.

Definitions of Important Terms

At this point definitions of the following terms used in this paper are included.

1. Natural growth - The growth in achievement that would take place normally in an educational setting in the absence of any special program.

2. Expanded standard scores or scale scores - Scores on a continuous scale designed in such a way that scores at each test level are normally distributed (as are standard scores) and that scores on different levels at successive time points represent growth across levels. The scaling method is based on Thurstone's (1925) absolute scaling technique. Expanded standard scores are called by different names in the various test manuals. The manuals for the tests used here refer to this type of score as scale scores so that term will be used in this paper. The scales are considered to be equal interval so that a difference of 5 points represents the same difference anywhere on the scale and so that successive scores can be used as a measure of growth over time.

3. Growth curve - A curve depicting the relationship between achievement in scale scores and grade level placement.

4. Normal curve equivalents (NCE's) - Scores converted to a normalized standard score scale with a mean of 50 and a standard deviation of 21.06. NCE's are the same as percentile ranks at the 1st, 50th and 99th percentiles.

Between these points they are distributed on an interval scale whereas percentiles are not.

5. Functional level testing - Testing students with a test level on which they may be expected to answer between 30 and 80 percent of the items correctly. In some cases this may require out-of-level testing, (i.e. testing with a test level other than that recommended for a student's grade level placement, or use of different test levels at pretest and posttest times). Scores for students tested out-of-level are converted to scale scores. These in turn are converted to in-level percentiles or NCE's. Thus, the performance of students tested out-of-level is still compared with that of their grade-level peers.

CHAPTER II

THE NORM-REFERENCED MODEL AND ITS ASSUMPTIONS ABOUT NATURAL GROWTH

The norm-referenced model uses students' growth relative to that implied by national norms as the basis for evaluating the effectiveness of an educational program. The no-treatment expectation is that students will maintain the same relative status, (i.e. percentile rank or NCE score), over a period of time.

The model as approved by the U. S. Department of Education (USDE) has several requirements designed to improve the validity of the results.

First, students are required to be pretested and posttested at dates near the empirical norming dates of the test being used. If the evaluation is based on a fall-to-spring testing schedule a test with both fall and spring empirical norms must be used.

Second, to counter the effects of statistical regression the model requires that students be selected for program participation on some basis other than their pretest scores.

Finally, to be properly implemented the model requires functional level testing. For many low-achieving

students this means using a test level that was normed on students in a different (lower) grade. When tests are normed, test levels are administered to students on the basis of their grade placement rather than their functional level.

Certain assumptions about natural growth are implicit in the norm-referenced model. The norm-referenced model compares observed growth with an estimate of what students' achievement would have been under conditions of natural growth. It assumes cross-sectional norms provide a valid estimate of natural growth. For example, consider a group of students who achieved at the 30th percentile at the end of third grade. Their expected achievement at the end of fourth grade would be the same as that of the norm group students at the end of fourth grade who achieved at the 30th percentile. Thus, the difference between the achievement of third and fourth grade students at the 30th percentile represents a year's expected growth for students at that percentile rank and grade level.

Put another way, under the equipercentile assumption percentile ranks and NCE scores for a group of students would be expected to remain constant across grade levels in the absence of special treatment effects. Students in a grade-level cohort with normal growth would be expected to show zero NCE gains from year to year. Grade level placement determines the norm group with which students'

scores are compared and, hence, their percentile ranks and expected achievement.

CHAPTER III

REVIEW OF THE LITERATURE

Models suggested in the literature for use in estimating program effects when a randomly assigned control group is not feasible and subjects are expected to change in the absence of any treatment are basically of two types:

1. models that use a non-equivalent comparison group for estimating expected gains, and
2. models that use information about the students' prior performance to estimate expected gains.

The models developed by RMC Research Corporation (Tallmadge and Wood, 1978) for evaluating Title I programs are examples of the first type. Olejnik's (1977) model and the value-added model (Bryk and Weisberg, 1976; Bryk, Strenio and Weisberg, 1980) are examples of the second type.

The three RMC models all evaluate gains using a type of standard score, the normal curve equivalent or NCE (Tallmadge and Wood, 1978). Thus, all make use of some kind of comparison group.

The RMC model referred to in the Title I literature as Model A is the norm-referenced model, the model being

evaluated here. The model assumes that, without treatment, students' percentile ranks and hence NCE scores will remain the same over time. The no-treatment expectation for the posttest score is the group's mean pretest NCE score. Tallmadge and Wood (1978) state (p.44), "Those children in the (norm) sample who obtained the same test scores as Title I children serve as a kind of control group." The analogy with a control group is not completely accurate because: 1) in nearly all cases the same children are not included in both the pretest and posttest norming samples, and 2) the composition and treatment with respect to relevant variables of the subsample being compared with the Title I children is not known. Nevertheless, data from the norming population are used as a substitute for control group data in the norm-referenced model.

The norm-referenced model was developed and refined by RMC Research Corporation for use with the Title I Evaluation and Reporting System or TIERS (Reisner et al, 1982). In a sense it was an outgrowth of earlier Title I evaluation decisions based on the concept of a "year's growth" using test scores expressed as grade equivalent units (Tallmadge, 1982).

The model was intended to provide information for Congress on the effectiveness of federally funded programs. A major concern was in obtaining data that could

be aggregated to provide an overview of program effects at the state or national level. The data so aggregated include results for large numbers of students involved in a variety of programs. Analyses of data at the national level identify small differences in gains as statistically significant while glossing over large differences among the programs aggregated (Reisner et al, 1982). In contrast, at the local level where single projects are evaluated small sample sizes may lead to large standard errors. Reisner et al(1982) describe the typical project as having fewer than 30 students and note that many projects have fewer than ten. With the large amount of variance present in student gains and the typically small sample sizes statistical power is often a problem.

The norm-referenced model as approved for use by USDE relies solely on the difference between observed and expected means and makes no provision for testing interaction effects. For example, Reisner et al(1982) note that the analyses performed when using the model give no information about how effective programs are in raising the floor of the distribution as opposed to the mean. Since the implied purpose of any compensatory education program is to remediate educational deficiencies, it seems possible that in many cases educational programs having their greatest effects on the lowest achieving students have been implemented. Such programs might result in substantial gains for the neediest students even though the mean

program gain is not large.

Of particular importance to the proposed study, the model makes no provision for variations in the effect of the regular school program. Where the regular program is markedly more or less effective than the national average, patterns of natural growth in a district would differ from those of the national norm group.

The varying effectiveness of regular school programs is perhaps one of the most important variables affecting the validity of the model. Tallmadge (1982) explores the question of school effects on the model in one of his analyses using scores for treatment and control students matched by school building. He found evidence of school effects in the data for Grade 2 but not for Grades 4 and 6. He notes that the treatment evaluated by the norm-referenced model is the students' total school experience. Thus an ineffective Chapter 1 program in an effective school system might show gains larger than those in the national norms. In contrast, for an effective program in an ineffective system the observed gains might be less than the no-treatment expectation.

A number of other concerns about the validity of the model have been raised in the literature. The problem receiving the most attention is that of regression to the mean. It was recognized when the model was first considered as a problem if the same test results were used for selection of students into the program and as pretest

scores (Tallmadge and Horst, 1976, Tallmadge, 1977). As a result the model approved by USDE for use with Title I programs requires selection of students on the basis of different test scores than those used for the pretest (Tallmadge and Wood, 1976).

Concern has been expressed by a number of writers that there is still a residual regression artifact present in the observed gains so long as the selection test correlates more closely with the pretest than with the posttest (Linn, 1980, Roberts, 1980, Trochim, 1982). An adjustment could be made for this residual regression if the selection-pretest and selection-posttest correlations were known. For the norming population they are not. Even the pretest-posttest correlation for the norming population is not known since the same individuals are not involved in the pretest and posttest norming. In the case of local studies, the correlations could be calculated and the adjustment made for any residual regression. This procedure has rarely been followed in actual practice.

A number of studies have addressed the validity of the equipercentile assumption, that in the absence of special treatment students will maintain the same relative positions with respect to one another over time. Included are studies by Tallmadge(1985,1982), Hiscox and Owen (1979), Powers et al(1983), and Linn(1979). In several cases evidence was found to dispute the validity of the equipercentile assumption in school and district data.

Linn (1979) used data originally collected by Van Hove et al (1970). The Van Hove study looked at achievement test results at two grade levels for schools in six urban school districts. Unweighted average percentile ranks for schools categorized by percent of minority students were compared. While the main purpose of the original study was to compare performance in the six school districts, Linn used the data to consider the validity of the equipercentile assumption. He converted the average percentile ranks from Van Hove's study into NCE's and examined the implications of the results for the norm-referenced model. He found that in nearly all of the cities the NCE scores for the later grade (6) were lower than for the earlier grade (either 3 or 4). He notes that the data are cross-sectional rather than longitudinal and that averaging percentiles across parts of a battery may conceal "interesting" trends. More to the point averaging percentile ranks is not considered an acceptable measurement procedure. While it's difficult to estimate the effect that averaging percentile ranks had on them, these results do raise doubts about the validity of assuming that urban minority students would maintain a constant NCE score (or percentile rank) from year to year in the absence of a compensatory education program.

Hiscox and Owen (1978) found evidence of significant variability in the percentile ranks of Portland (Oregon) students over a four year period. Using longitudinal data

from the Portland Public Schools they followed the Comprehensive Test of Basic Skills (CTBS) scores for groups of low-scoring fourth and seventh grade students for up to four years. Students were grouped both by the percentile rank of their first year's score and by the number of years (zero to four) that they were eligible for Title I services. Hiscox and Owen imply that all eligible students received Title I services. However, they provide no evidence that they checked beyond students' test scores to determine that this was actually so. Included in their data were scores for students attending eight elementary schools offering Title I programs. No information is given about the nature of the Title I programs at these schools or even whether all the schools provided the same type of Title I program. Their results showed year-to-year differences in the groups' percentile standings ranging from zero to nine percentiles with most differences in the two to three percentile range. From this they conclude that the "noise" level in a local norm-referenced study may be greater than the potential student gains. This suggests that results from data for a single year (or cohort) may, at the local level be unduly influenced by the presence of local variation.

Powers et al (1983) in their study of Tucson(Arizona) students found that groups of these students gained consistently in percentile rank over a fall-to-spring interval. Seventh and ninth grade students, from schools that

did not participate in Title I projects, were grouped on the basis of stanines and also into deciles. When students were grouped by stanines into low (1-3), middle (4-6) and high (7-9) achieving students the mean NCE gains for all the groups (low, middle and high at each grade level) were consistently greater than zero. From this he concluded that use of the equipercentile assumption is inappropriate.

Tallmadge (1985) reanalyzed Powers' data and concluded that they actually provide support for the equipercentile assumption. He notes that when the selection test scores are included in the analysis a different picture results. In Table 1 may be found the overall mean seventh and ninth grade scores for the selection test, pretest and posttest along with the corresponding N's and standard deviations as reported by Powers (1983, p.301).

Normally in using the norm-referenced model only the pretest and posttest scores would be considered. However, in this case the selection test scores were used only to subdivide the group for analysis and not for selecting students into the study. Further, the selection test scores were obtained two years prior to the pretest. Under these conditions, Tallmadge argues, it is reasonable to consider the mean selection scores for the total seventh and total ninth grade groups as providing evidence for or against the validity of the equipercentile assumption. He finds the selection and pretest scores obtained

Table 1. Summary of Powers'(1983) Data

Grade	N	Test	Mean Score	SD
7	1327	Selection	59.24	18.59
		Pretest	58.61	19.20
		Posttest	62.11	18.11
9	1897	Selection	59.45	19.14
		Pretest	58.88	18.33
		Posttest	61.03	18.86

two years apart to be "virtually identical." The posttest scores obtained the following spring are, however, significantly higher. Tallmadge examines a number of possible alternative hypotheses. He concludes that the most likely explanation is one that attributes the apparent gains to an artifact resulting from misapplication of the norm-referenced model, e.g. "stakeholder" bias. This results in bias when the data are interpreted using the norm-referenced model.

A study by David and Pelavin (1978), found evidence that fall-to-spring gains were not maintained over the summer. Analysis of Title I data nationally (Reisner et al, 1982), suggests that gains based on fall-to-spring testing are artificially inflated relative to those ob-

tained when following an annual testing schedule. A number of sources of bias that may contribute to the inflation of fall-to-spring gains are identified. Thus, Powers' findings may reflect sources of bias present in gains based on fall-to spring testing that are not present when growth is measured annually.

An earlier study by Tallmadge(1982) using a national data base provides support for the validity of the equipercentile assumption. He uses data from the Sustaining Effects Study* and from the 1977 norming of the California Achievement Tests. An unknown number of students in the CAT norming file were participants in compensatory education programs. Students were selected in a manner designed to simulate the procedures in implementing the norm-referenced model and the randomized control group design. Specifically, he divided students in the data base into high- and low-achieving groups on the basis of a selection test cutoff score. The low-achieving group was randomly assigned to simulated treatment and control groups. (Of course, neither group actually received any treatment.) The data used were for students tested in both the fall and spring of the same year. Tallmadge reports the results of three analyses using scores for

 *The Sustaining Effects Study (SES) is a longitudinal study of compensatory education funded by the U.S. Department of Education. The SES data analysis file used by Tallmadge was one made up of scores of students who had not participated in compensatory education programs.

students in grades 2, 4 and 6. In all of these he compared gains for the treatment group obtained using a simulated control group model. Although his results for the norm-referenced model indicate a positive bias of about 1 NCE for the Title I groups, he found them to be comparable in degree of accuracy with those obtained from the simulated control group model. (The variation for the randomized control group was, however, about equally positive and negative.) As a result of his study he concluded that "the norm-referenced model yields gain estimates that are reasonably comparable to those derived from the randomized control group design"(p. 110).

One aspect of the validity of the equipercentile assumption concerns the use of cross-sectional norms to generate the no-treatment expectation for project evaluation. Local data collected for the norm-referenced model are longitudinal with pretest and posttest scores for the same students. Norming data are usually based on scores of different students for each grade level and norming date (Reisner et al 1982).

The norm-referenced model assumes that the cross-sectional data provide a valid representation of longitudinal growth. Specifically, the use of cross-sectional norms assumes that the students in the norm population represent the same population at different points in time (Murray et al 1979). When evaluating local projects the use of the model assumes that the norm population is representative

of the local population at the pretest and posttest time points, or at least that any differences are constant across the pretest to posttest time period. Thus, the growth rate of students in a project would, in the absence of any special program, be the same as that for students in the norming population.

Variables that have been addressed as possibly affecting the validity of using a national norming population as a control group include: differences in school policies relative to promotion of students (Tallmadge, 1977), differences in the racial/ethnic composition (Tallmadge, 1977), differences resulting from comparing longitudinal project data with cross-sectional norming data (Reisner et al, 1982; Trochim, 1982), and use of the appropriate test level (Tallmadge & Wood, 1978). Other studies in contexts unrelated to the norm-referenced model have examined factors that affect patterns of growth in school achievement. For example, Langer et al (1984) found age at school entry related to the pattern of growth in later school achievement. Bryk et al (1980) found achievement at day care centers related to a number of background variables including sex, race and SES as well as to age. Porter et al (1978) found that standardized tests differ in content and hence in match to any given curriculum. The match between test and curriculum could affect the pattern of growth observed. The studies reported by Linn (1979), Powers et al (1983) and Tallmadge (1982) suggest that the

effectiveness of the overall school program may be an important factor in the results obtained.

Nationwide 99 percent of school systems use the norm-referenced model to evaluate Chapter 1 programs. Most of those using another model use RMC's special regression model, Model C (Reisner et al 1982). Trochim (1982) found enough school districts in Florida using Model C to permit a study comparing results obtained using the norm-referenced and special regression models. On the basis of a meta-analysis of evaluations of districts using either of the two models Trochim found that programs evaluated with the norm-referenced model tended to show positive NCE gains while those evaluated with the special regression model were more likely to show zero or negative NCE gains. Though Trochim concludes that the norm-referenced model overestimates the effectiveness of programs while the regression model underestimates it, his study provides no evidence of the extent to which the difference in results is due to sources of bias present in each of the models or to bias in the districts electing to use a given model.

None of the above studies approaches the norm-referenced model as providing a model of natural growth in measured achievement. Though the assumptions of the norm-referenced model when combined with the percentile norms and expanded standard score scale for a given test implies a model of expected growth in the achievement measured,

this aspect of the model has not received the attention of researchers. Yet it is important. As Linn (1981,p. 183) notes, "Most work on the measurement of change has devoted little or no attention to models of growth. But good models of growth seem crucial to measuring and interpreting measures of change.... Developing sound models of growth, especially for growth in measured achievement over the school years, will require considerable research."

CHAPTER IV

PROCEDURES

The procedures for carrying out this study can be grouped into three parts:

1)those involved in development and comparison of the patterns of growth implied by the norm-referenced model;

2)those involved in collecting and analyzing the local cross-sectional data; and

3)those involved in collecting and analyzing the longitudinal data.

Development and Comparison of Patterns of Growth Using the Assumptions of the Norm-referenced Model

Although not explicitly specified by the norm-referenced model, expected patterns of growth are implied when the equipercentile assumption is applied to empirical data. Defining these patterns of growth required data from multiple time points, scores expressed on a developmental scale across grade levels and an equivalence between percentile ranks at each grade level and scores on a developmental scale, such as the expanded standard score scale used for the scale scores.

Procedures specified under the model for comparing students tested out-of-level with in-level norms requires

that raw scores be converted to in-level NCE's (Tallmadge & Wood, 1976). The scores are converted using scale scores. This assumes that for every scale score there is an equivalent in-level NCE and at any grade level there is a scale score equivalent of every NCE score. Both NCE's and scale scores are designed to provide equal interval scales at all grade levels. NCE's remain the same across grade levels, whereas scale scores permit a measure of growth over time. The expected "normal growth" pattern across grade levels for students attaining a given percentile rank or NCE score on a pretest can be represented by the scale score equivalents of that rank or score. No-treatment growth can be expressed in terms of expanded scale score equivalents across grade levels for any NCE score or percentile rank.

Two nationally standardized test batteries, the California Achievement Tests, 1978 edition(CAT) and the Metropolitan Achievement Tests, 1978 edition(MAT), were selected for use in this study. These two tests are among the tests most widely used for evaluating Michigan Chapter 1 programs. For these tests the patterns of growth implied by the equipercentile assumption and the tests' norms tables (California Achievement Tests, 1978; Prescott, Balow, Hogan and Farr, 1978) were examined. Scores for Total Reading and Total Math batteries were used. Using empirical spring norms the scale scores corresponding to the 10th, 30th, and 50th percentile ranks at each grade

level were determined. These percentile ranks cover the range of achievement usually encountered in Chapter 1 programs. Scale scores were listed and plotted across grade levels for the selected percentile ranks. The scale scores represent expected achievement on the tests' empirical spring norming dates each year.

Use of spring norms assumes a spring-to-spring test interval which provides a more valid determination of yearly growth than fall-to-spring testing. It eliminates both the fluctuations that result from differences in the growth rate during periods of vacation and schooling and what Tallmadge(1985) refers to as "stakeholder bias." This is the bias that results when the classroom teacher who has a stake in the students' achievement gains administers both the pretest and the posttest. Approximately 80 percent of Michigan school districts evaluate Chapter 1 programs using a spring-to-spring test schedule. Although fall-to-spring testing is more common in other states, USDE recommends that districts be encouraged to use annual testing for program evaluation (Reisner et al 1982). Rarely do school systems test students district-wide more than once a year.

Scale scores for the 10th, 30th and 50th percentiles based on spring empirical norms for grades Kindergarten through 12 were determined for the CAT and MAT. The scale scores for each percentile rank were then plotted against grade level to produce a set of growth curves for each

test.

Next the resulting growth curves for these two tests were compared to determine whether they appeared to be similar. The two reading tests purport to measure achievement in the same subject and so do the two mathematics tests. It was hypothesized that they would show similar growth curves. Differences between the curves for the two tests would indicate differences content between the two tests or lack of an interval scale for one or both tests. If there are differences the success of a local program being evaluated with the norm-referenced model would depend on the test used and on its match to the local curriculum. Further, the norm-referenced model is used to allow reading and mathematics data to be aggregated statewide and nationally across tests and at the state level to compare the achievement of districts using different tests. Implicit in these procedures is the assumption that the tests used measure achievement in a common content for which there is a common growth curve on an interval scale.

Since the two tests use different scales it was impossible to compare the expected scores for the two tests directly. To make possible a comparison of the scores for the two tests they were rescaled by setting the 50th percentile score for Grade 2 at 600 and that for Grade 6 at 800. The difference between the 50th percentile scores for Grades 2 and 6 on the new scale(200) was

divided by the differences between the scores for the same points on the original scales to obtain the coefficients for converting other points on the curves to the new scale. The following formulae were used in the rescaling.

CAT reading: $Y=600+1.4286(X-360)$
 MAT reading: $Y=600+1.7094(X-620)$
 CAT mathematics: $Y=600+1.4388(X-352)$
 MAT mathematics: $Y=600+1.0582(X-507)$

With the scores on the same scale it was possible to compare the variance (from the 50th percentile to the 30th and 10th) as well as the configurations of the growth curves for the two tests.

To test the null hypothesis that the growth curves are the same for both tests it was necessary to test both the configurations and variances. To test whether the configurations were the same Lord's(1957) test that the disattenuated correlation between two tests is 1.00 was used. ($H_0: \rho_{xy}=1.00$) Lord's test requires in addition to the correlation between the two tests a measure of the reliability of each of them. The problem here was that the growth curves used in this study were taken directly from the tests' norms tables and could be assumed to have a reliability of 1.00. Any difference would then be significant. What was really wanted was a test of whether the growth curves at a given percentile rank for the two tests differed significantly more from each other than they did from the growth curves for other percentile ranks for the same test. That is, the correlations between

curves for different percentile ranks for a single test represent the expected variation among curves for that test. Therefore, the correlation between points on the growth curves for two different percentile ranks for one test was used in place of the usual reliability coefficient. In the case of the 50th percentile curve the correlations between it and the 30th and 10th percentile curves were calculated and averaged. These correlations for the CAT and the MAT were then used with Lord's test in place of the reliability coefficients.

Let r_{xx} = the average correlation between the CAT
curve being tested and the other CAT
curves;

r_{yy} = the average correlation between the MAT
curve being tested and the other MAT
curves; and

r_{xy} = the correlation between the CAT and MAT
curves being tested.

Under H_0 assume $\rho^0_{xx} = \rho^0_{yy} = \rho^0_{xy} = \rho^0$
 ρ^0 cannot be measured directly but Lord(1957) demonstrates that $\hat{\rho}^0$ can be estimated using:

$$\hat{\rho}^0 = 1/6(r_{xx} + r_{yy} + 4r_{xy}).$$

The alternative hypothesis is that the curves are not the same. ($H_1: \rho_{xy} \neq 1.00$)

Under H_1 assume $\rho^1_{xx} = r_{xx}$;

$$\rho^1_{yy} = r_{yy}; \text{ and}$$

$$\rho^1_{xy} = r_{xy}.$$

The formula for Lord's test is then:

$$\chi_1^2 = (N-1) \log_e \frac{(1 - \hat{\rho}^0)^2 [(1 + \hat{\rho}^0)^2 - 4(\hat{\rho}^0)^2]}{(1 - r_{xx})(1 - r_{yy}) [(1 + r_{xx})(1 + r_{yy}) - 4(r_{xy})^2]}$$

To test the hypothesis that the levels of the two tests' growth curves are the same the SPSS^x subprogram T-TEST was used with a paired-samples design. The hypothesis tested was $H_0: \mu_{x-y} = 0$ for each of the percentile ranks.

Effect of Variables Hypothesized to Affect the Norm--referenced Model's Validity at the Local Level

A number of variables may affect the validity of the model's assumptions about natural growth. Three of them were selected for this study. The first variable was the use of out-of-level testing. Since the model requires functional level testing many districts test their low-achieving students out-of-level. The results are compared with norming data obtained from students tested in-level. To test the validity of this practice two school districts that test their Chapter 1 students with a test level below that recommended for their grade placement were selected for inclusion in this study.

Another factor was the effectiveness of the overall school program in reducing the proportion of low-achieving students in a school district. Districts showing a

pattern of improving achievement over a period of years might be expected to show changing patterns of growth. Two districts that have shown a decreasing proportion of students achieving less than 50 percent of the objectives on the Michigan Educational Assessment Program (MEAP) tests over the period from 1982 to 1984 were selected for inclusion in this study. A variety of background variables, (e.g. sex, race and SES), may affect patterns of growth if composition of the local population differs markedly from that of the norming population on any of them. A high proportion of low SES students has been most closely linked to Title I (now Chapter 1) programs. Eligibility for Chapter 1 funding is based on the proportion of low SES students in any given school building. Schools participating in test norming studies may be assumed to have an average number of low SES students overall. Two school districts with a high proportion of students eligible for Chapter 1 were included in the study.

Selection of Districts

A total of six school districts were selected, one that exemplifies each of three variables for each of the two tests. This resulted in the following design:

	Out-of-Level Testing Used		Improving MEAP Scores		High Prop. of Low SES Students	
	I-----I		I-----I		I-----I	
CAT	I	1	I	1	I	1
	I-----I		I-----I		I-----I	
MAT	I	1	I	1	I	1
	I-----I		I-----I		I-----I	

The following abbreviations will be used to identify the districts in this study.

COL - CAT, out-of-level testing used
 CIM - CAT, improving MEAP scores
 CLS - CAT, high proportion of low SES students
 MOL - MAT, out-of-level testing used
 MIM - MAT, improving MEAP scores
 MLS - MAT, high proportion of low SES students

The six districts were selected from those that used either the CAT or MAT both for reporting Chapter 1 achievement data and for district-wide testing in Grades 2 through 6. A list of 120 districts which used the CAT and 119 which used the MAT for Chapter 1 evaluations in the 1982-83 school year was obtained. These were evaluated for each of the three variables in the design. From 12 to 22 districts were identified as ranking high on one of the variables and average or low on the other two. These districts were then contacted to determine whether or not they would have the data needed for the study. They were asked whether they did district-wide testing, which tests they used and at what grade levels. Those that did district-wide testing with either the CAT or MAT at three or more grade levels from Grades 2 through 6 were considered candidates for the study. From three to seven districts

in each cell were identified as having the needed data. Willingness to cooperate and ability to provide the data in a useable form were also evaluated. Where more than one district met the criteria for a cell and were equally able and willing to provide the desired data final selection was made by a random drawing. Reading and mathematics data were requested for Grades 2 through 6 for the years 1983, 1984 and 1985.

The final sample consisted of three suburban districts, COL, CIM and MIM, and three rural (largest town had 2,185 population) districts, CLS, MOL and MLS. These districts served from 7.25 (MIM) to 34.71 (MLS) percent of their students in Grades 2 through 6 in Chapter 1 programs. Four of the districts had Chapter 1 programs in both reading and mathematics. One, MIM, had a program in reading only and one, MOL, had programs in reading and language arts.

Collection and Processing of Data

Table 2 shows the total number of student scores obtained by subject, grade level and district. One district, MOL, had test data for Grade 6 only for 1985. For all other districts and grade levels three years of data were obtained.

Only one of the districts had computer facilities that enabled them to aggregate district-wide frequency distributions for submission. Two of the districts aggregated their data by building before submitting it. The

Table 2. Total Students Tested in 1983, 1984 and 1985

		Grade Level				
District		2	3	4	5	6
Reading						
COL		2559	2626	2625	2766	2901
CIM		887	808	821	777	775
CLS		484	486	494	477	522
MOL		178	166	208	204	66
MIM		225	247	251	244	275
MLS		107	93	107	92	111
Mathematics						
COL		2549	2625	2630	2766	2903
CIM		887	806	823	776	740
CLS		484	484	496	476	523
MOL		183	166	207	203	67
MIM		198	247	251	244	275
MLS		108	95	106	92	111

remaining three districts submitted individual pupil data. The data received from the local districts were entered into a Honeywell(GCOS) Computer, aggregated into district-wide frequency distributions where necessary and analyzed using SPSS^x Version 2.0.

Examination of the size of the districts participating in this study reveals that size of district is confounded with the test used. A review of the Michigan districts that used the CAT as opposed to those that used the MAT to measure Chapter 1 achievement in 1982-83 reveals that the 120 districts using the CAT reported data for more than twice as many students as the 119 districts using the MAT (CAT, 3,871; MAT, 1,662). The districts used in this study reflect the differences in size of

districts using the tests in Michigan. Surprisingly, in both cases the low SES districts were the smallest in the sample using their respective tests.

Development of Growth Curves Based on Local District Data

Growth curves were developed using local data for the six school districts. Local 10th, 30th and 50th percentile rank scores were identified for both reading and mathematics for each of the three years and for a composite of the three years' combined data.

Growth curves based on local data for the six school districts were compared with those for the national norms of the tests used by these districts. The question of whether to use the scores of all students tested in a district or only the scores of students who did not receive Chapter 1 services in developing local norms was considered. On the one hand inclusion of Chapter 1 students' scores in the local norms would tend to adjust out any Chapter 1 treatment effect. On the other hand omission of Chapter 1 students' scores would bias the distribution because nearly all the omitted scores would be taken from the low end of the distribution. The resulting local norms would be based on distributions with few or no values in the range where Chapter 1 students would be

expected to score. National norms for all tests currently used to evaluate Chapter 1 programs in Michigan are based on populations that include Chapter 1 students. It was decided to base the local norms on scores for all students tested in a district. This means that Chapter 1 students were included as they are in the national norms. The proportion of Chapter 1 students in the local norming population ranged from 7.25% to 34.71% with the median being 13.74%. The proportion in the national populations could not be determined though it is usually assumed to be comparable to the proportion of Chapter 1 students in the school population nationwide.

Effects of Local District Growth Patterns on the Validity of the Norm-referenced Model

Determination of No-treatment Gains

The local percentile ranks corresponding to the national 10th, 30th and 50th percentile ranks were identified and the expected posttest scores for each if the local percentile rank was maintained were determined. From this it was calculated what the apparent no-treatment gains would be under the model if growth occurred in accordance with local rather than national norms. With data for five grade levels gains for four spring-to-spring intervals or grades could be calculated.

At this point the data could be grouped into a

four-way design (2 tests x 3 district types x 3 percentile ranks x 4 grades). To determine whether the number of categories could be reduced four-way ANOVA's using the spring-to-spring gain scores as the dependent variables and the pooled grade by percentile rank interactions as the error terms were run separately for reading (Table 3) and for mathematics (Table 4). (SPSS^x subprogram MANOVA was used for the ANOVA runs.) A criterion of $\alpha = .05$ was used to determine whether there were significant effects for a variable. On the basis of the results of the ANOVA runs, as shown in Tables 3 and 4, it was determined that there were no significant main effects or interactions involving percentile rank. The data for the three

Table 3. Reading 4-Way ANOVA Summary

Source of Variation	SS	df	MS	F
Constant	20.06	1	20.06	1.77
Test	83.59	1	83.59	7.37*
District Type	12.48	2	6.24	.55
Grade	192.33	3	64.11	5.65**
Percentile Rank	.06	2	.03	.00
Test by District Type	20.23	2	10.12	.89
Test by Grade	139.33	3	46.44	4.09*
Test by Percentile Rank	15.72	2	7.86	.69
District Type by Grade	70.07	6	11.68	1.03
District Type by Percentile Rank	15.46	4	3.87	.34
Test by District Type by Grade	125.02	6	20.84	2.13
Test by District Type by Percentile Rank	39.08	4	9.77	.86
Residual	408.50	36	11.35	
Total	1141.93	72		

* Significant if $\alpha = .05$.

** Significant if $\alpha = .01$.

Table 4. Mathematics 4-Way ANOVA Summary

Source of Variation	SS	df	MS	F
Constant	3.32	1	3.32	.32
Test	1.50	1	1.50	.14
District Type	100.09	2	50.05	4.77*
Grade	210.65	3	70.22	6.69**
Percentile Rank	2.44	2	1.22	.12
Test by District Type	157.07	2	78.54	7.48**
Test by Grade	97.07	3	32.36	3.08*
Test by Percentile Rank	.16	2	.08	.01
District Type by Grade	141.08	6	23.51	2.24
District Type by Percentile Rank	94.60	4	23.65	2.25
Test by District Type by Grade	228.52	6	38.09	3.63**
Test by District Type by Percentile Rank	73.96	4	18.49	1.76
Residual	377.82	36	10.50	
Total	1488.28	72		

* Significant if $\alpha = .05$.

** Significant if $\alpha = .01$.

percentile ranks, therefore, were combined into a single category with a mean of 37.3 NCEs or at the 27th percentile rank. The resulting 2x3x4 design was used in the remaining mathematics analyses.

An additional reduction in categories was possible with the reading data. Since there were no significant effects associated with district type in reading the data for the three district types were combined. This resulted in a 2x4 design which was used in the reading analyses.

Probability of Type I Errors

The gains were then tested for significance to determine whether under simulated zero-gain conditions they

would result in a Type I error if the model were used to evaluate hypothetical groups of 10, 20 or 30 students.

The test statistic typically used in evaluating the effectiveness of Chapter 1 programs is a one-tailed t-test. For the purpose of comparing local and national norms the formula is:

$$t = \frac{\bar{X} - \mu}{(\sigma/\sqrt{n})}$$

where: \bar{X} = the score that would be observed on the basis of the local norms;

μ = the expected score on the basis of the national norms;

σ = the standard deviation of the national norming population, and

n = the number of students in the group being evaluated.

While the standard deviation for that portion of the norming population with a mean score at the 10th or 30th percentiles is not included in the technical data published for either of the tests used it is possible to calculate what it would be given the standard deviation for the total population and the following assumptions.

- 1) The scores are normally distributed.
- 2) Groups were established by setting a cut score and including all students scoring below it in the group.

The following procedure was used. First convert the desired percentile rank to its equivalent z-score. Assume that \underline{a} = the mean score of the distribution and that \underline{b} = the cut score that will result in a distribution with mean \underline{a} . (See Figure 1.)

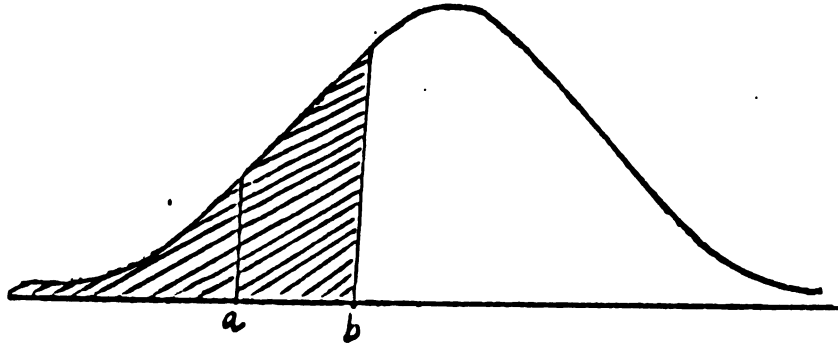


Figure 1. Distribution of Hypothetical Group

To find the variance for a population with mean \underline{a} it is first necessary to find the cut point \underline{b} . The mean z-score for the population with scores $x < b$ equals the ratio of the ordinate of the curve at $x = b$ to the density for $x < b$. These values were found by successive iteration first for the 10th and 30th percentile ranks and later for the 27th percentile. Since a population with a mean at the 50th percentile would include the entire distribution its standard deviation was assumed to be known.

The variance for the desired distribution is then given by the formula:

$$\text{Var } X_b = 1 + ab - a^2$$

Taking the square root gives the standard deviation in

z-scores. Multiplying by 21.06 (the standard deviation for NCE scores) converts the standard deviation to NCE's. The standard deviation for the 27th percentile rank was calculated to be 13.92 NCE's. That is, for the combined groups with a mean at the 27th percentile of 37.3 NCE's the standard deviation would be 13.92 NCE's.

The effect on the Type I error rate of using national norms when the local norms represent the true normal pattern of growth was tested for programs resulting in zero true gains with n's of 10, 20, and 30 and for $\alpha = .05$ and $.01$.

Probability of Detecting True Gains

The study also examined the effect on the power of the test to detect true gains under each of the above conditions when a program results in gains of 1.0, 4.0 and 7.0 NCE's.

In practical terms the effect of differences in power is most apparent in the size of the sample that would be needed to assure detecting real gains if they exist. The size sample that would be needed to detect true gains of 1.0, 4.0 and 7.0 NCEs with $\alpha = .05$ and $.01$ and with a probability of $p > .50$ was calculated. Since the t-value varies with the sample size successive iterations were used where necessary to match the sample size required with the t-value.

Comparison of Longitudinal Data with Expectations Based on Cross-sectional Norms

This study next examined the similarity between longitudinal and cross-sectional data for two of the six districts. Two districts were selected which were able to identify matched non-Chapter 1 scores from the data used earlier in the study. Those selected happened to be the two that used out-of-level testing(COL and MOL).

This portion of the study was limited to data from students not in Chapter 1 programs. The districts were asked to provide matched data for the three years, 1983, 1984 and 1985 for a sample of non-Chapter 1 students.

Students in the longitudinal sample belonged to one of three cohorts:

- Cohort 1 - students with test scores for Grades 2, 3 and 4;
- Cohort 2 - students with test scores for Grades 3, 4 and 5; and
- Cohort 3 - students with test scores for Grades 4, 5 and 6.

The matched data were used to compare the observed growth found in longitudinal data with that expected on the basis of the cross-sectional norms. Since the longitudinal study was limited to students with test data available for three years the samples were considerably more limited than the group on which the cross-sectional analyses were based. Elimination of students served by Chapter 1 programs further restricted the group. In

particular the number of low-achieving and highly mobile students were severely reduced in the longitudinal sample. These restrictions result in serious confounding that limits the interpretation that can be drawn from these data.

Changes in NCE scores (gains) from 1983 to 1984, 1984 to 1985 and 1983 to 1985 were computed and tested (using SPSS^x subprogram MANOVA) for significance. The purpose of this analysis was to determine whether the pattern of growth differs when longitudinal rather than cross-sectional data are used. This provided a further test of the validity of the equipercentile assumption.

CHAPTER V

RESULTS OF THE STUDY

The results of this study will be presented in the same three groupings as were used in Chapter 4.

Comparison of Growth Curves for the CAT and MAT

Table 5 lists the scale scores corresponding to the 10th, 30th and 50th percentile ranks for reading and mathematics from Kindergarten through Grade 12, for the CAT and the MAT. The growth curves for the CAT and MAT for reading and mathematics are shown in Figures 2-5.

There appear to be obvious differences in the growth curves, particularly for the reading tests. Both tests imply a higher rate of growth in the early grades and a tendency to level off at higher grade levels. They differ, however, in the proportion of total growth that is assumed to occur at the early grade levels. On the basis of the MAT scale, 53 percent of total growth in reading occurs by the end of Grade 2. Using the CAT norms and scale, 53 percent of total reading growth does not occur until the end of Grade 4. In mathematics a similar though less pronounced difference exists. The MAT scores indicate that 36 percent of total growth occurs by the end of Grade 2 while on the basis of the CAT scores it would appear that 36 percent of total growth is not attained

Table 5. CAT and MAT Scale Scores, Reading and Mathematics

Subj.	Gr.	CAT			MAT		
		10%-ile	30%-ile	50%-ile	10%-ile	30%-ile	50%-ile
Reading	K	185	214	235	271	326	363
	1	245	280	303	412	466	506
	2	290	332	360	524	576	620
	3	324	370	401	566	621	661
	4	357	406	443	602	655	695
	5	385	438	473	628	682	718
	6	405	460	500	632	693	737
	7	424	484	521	642	709	754
	8	447	509	553	645	725	775
	9	463	528	574	674	746	791
	10	481	548	596	704	773	815
	11	497	573	619	724	787	827
	12	505	582	633	746	806	846
Math	K	246	259	271	239	289	329
	1	278	299	311	327	378	418
	2	309	336	352	403	468	507
	3	343	374	394	454	521	567
	4	372	406	428	501	578	622
	5	395	436	463	546	624	664
	6	418	462	491	564	646	696
	7	425	482	517	597	681	731
	8	458	514	555	615	700	759
	9	469	533	579	637	717	772
	10	485	551	601	663	738	789
	11	499	566	612	682	757	800
	12	499	569	620	696	773	815

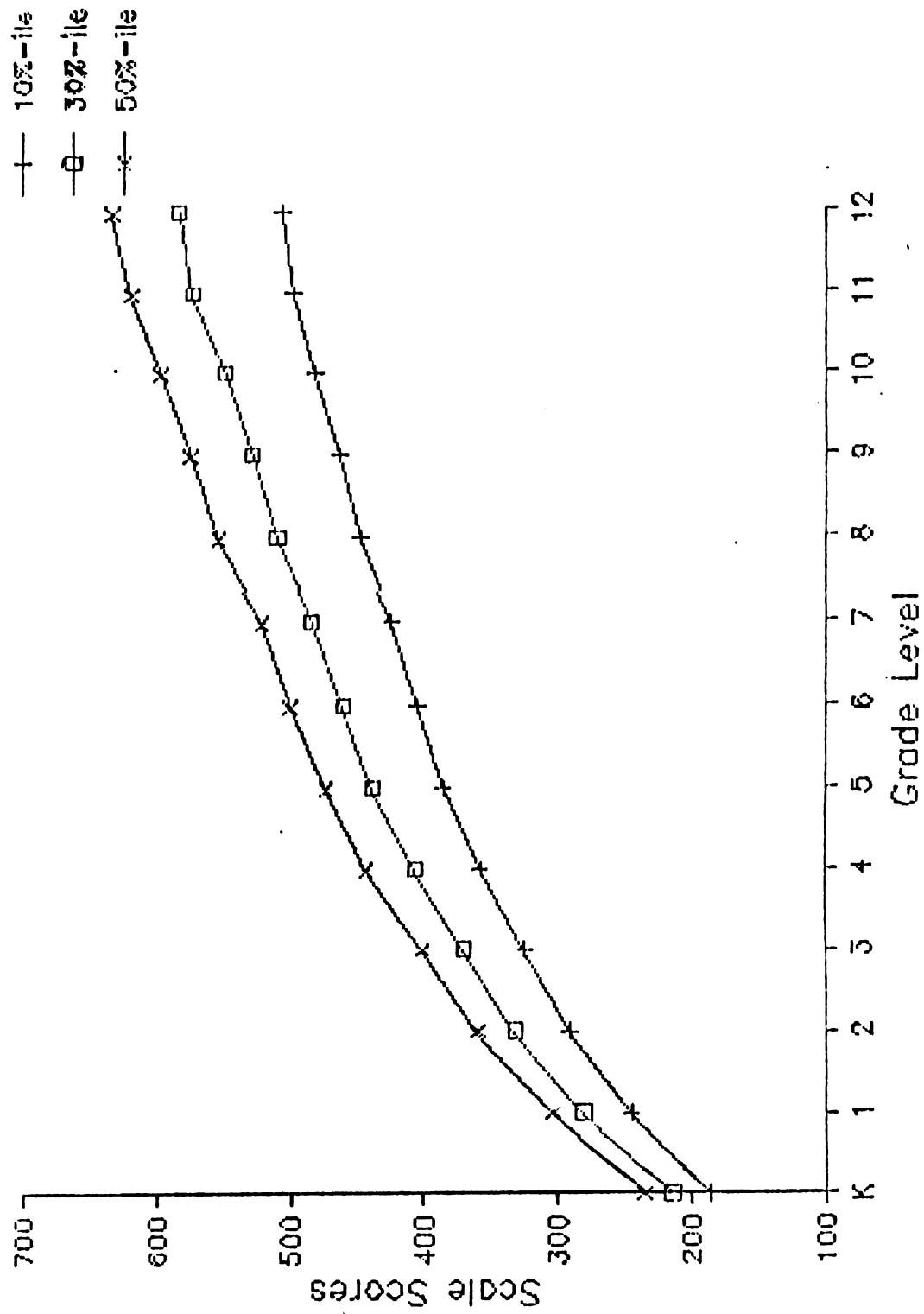


Figure 2. CAT Reading Growth Curves

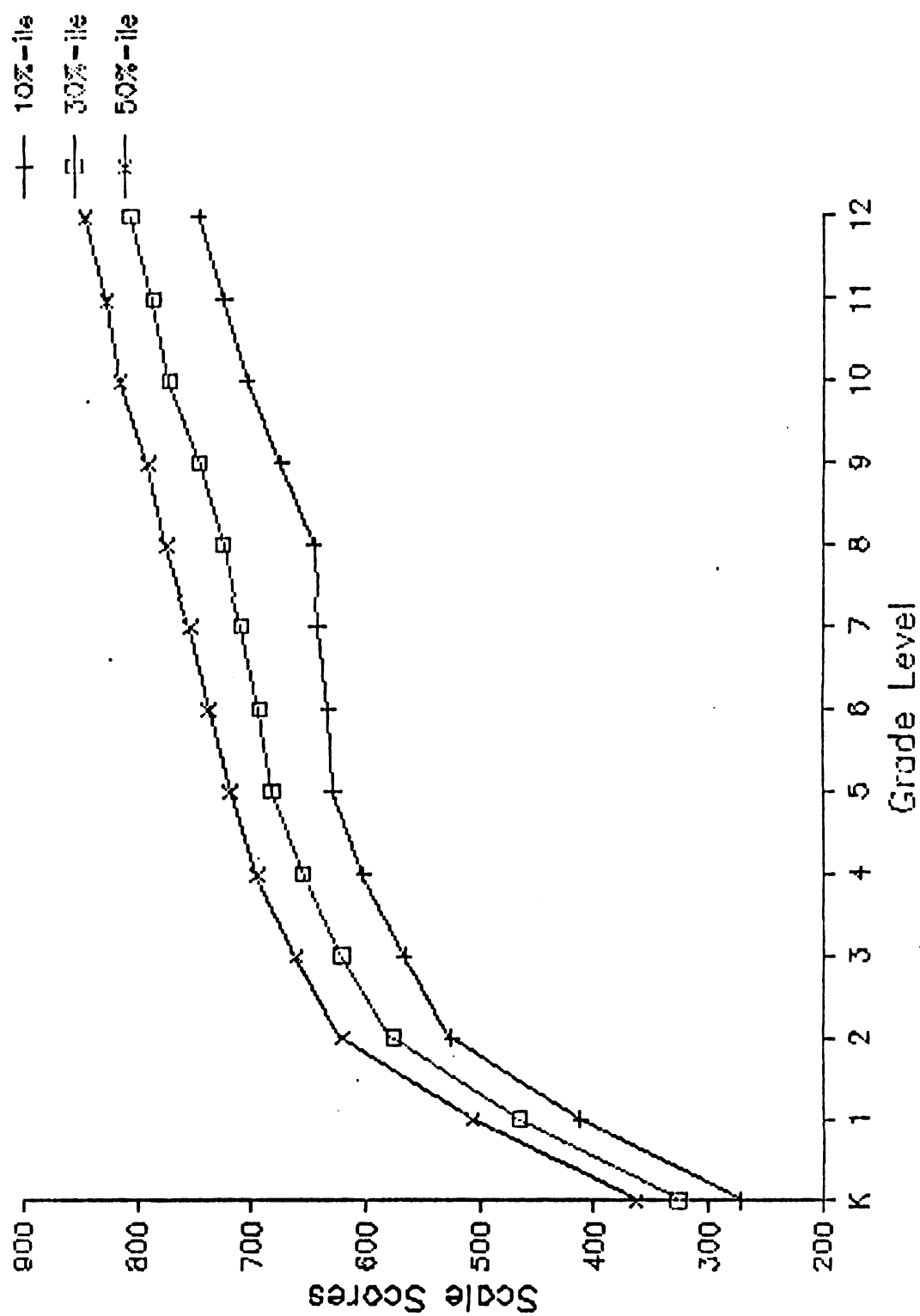


Figure 3. MAT Reading Growth Curves

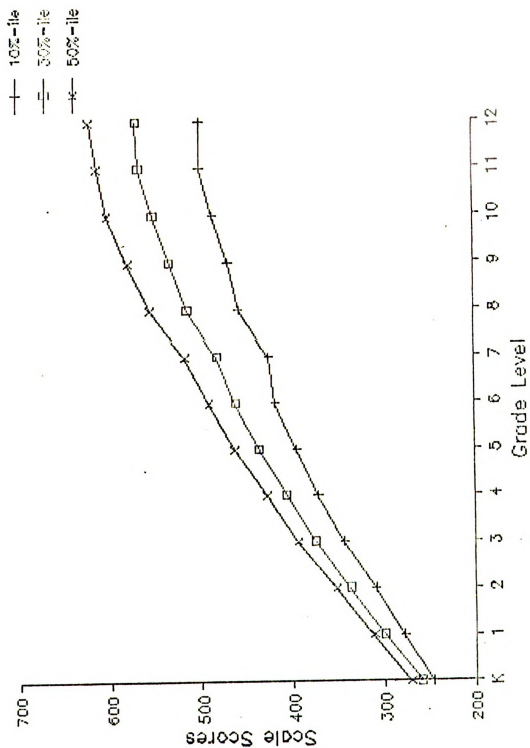


Figure 4. CAT Math Growth Curves

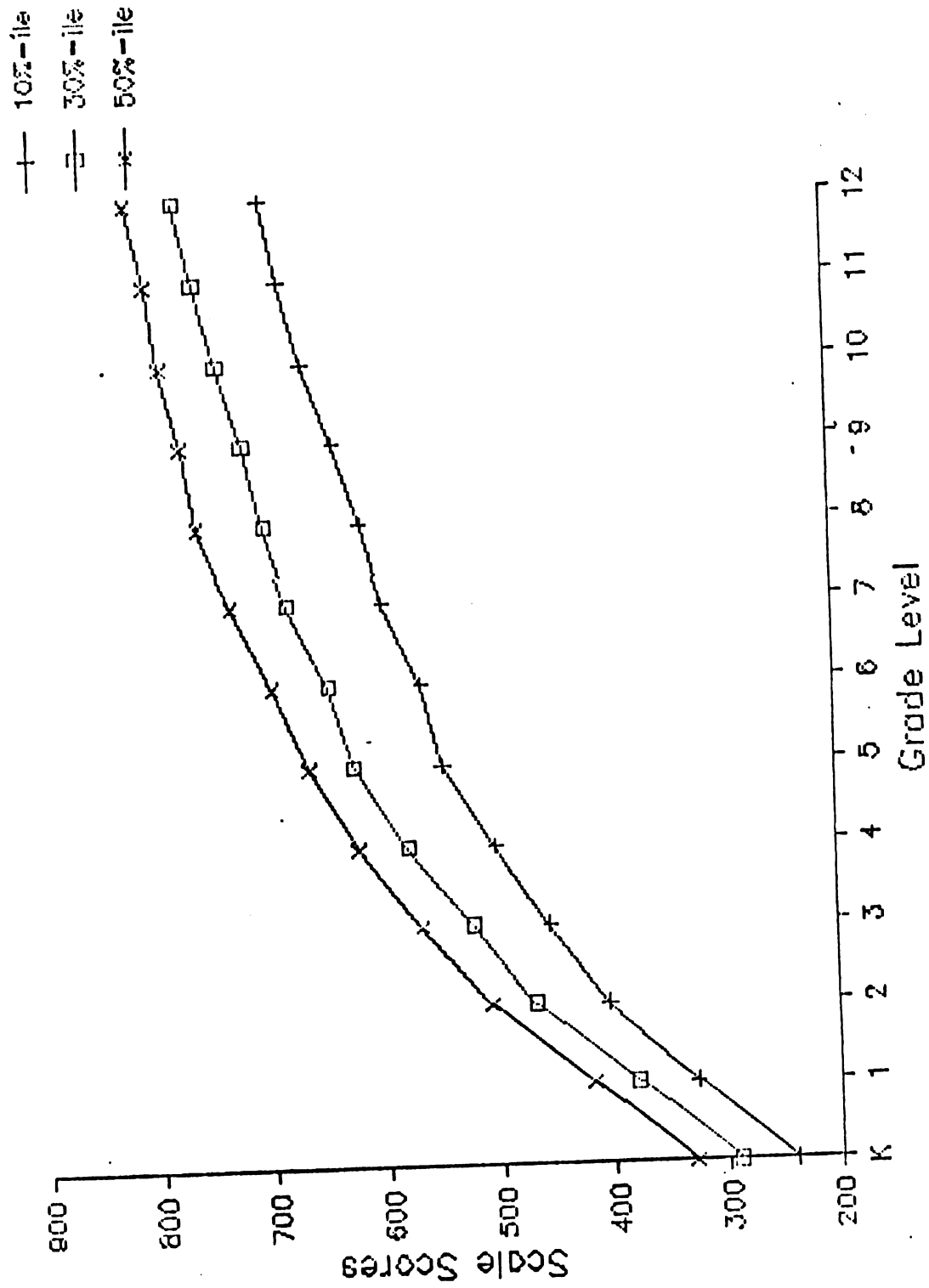


Figure 5. MAT Math Growth Curves

until the end of Grade 3.

Table 6 gives the scores adjusted to a common scale. Figures 6 and 7 show the plots of the two tests overlaid for reading and for mathematics.

With the scores on the same scale it is possible to compare the variation (from the 50th percentile to the 30th and 10th percentiles) as well as the configurations of the growth curves for the two tests. Setting the 50th percentile scores for grades 2 and 6 at the same values for both tests, reveals differences in configuration at the lower extremes (Kindergarten and Grade 1) of the curves for both reading and mathematics that are readily apparent. Differences in the mathematics curves are also apparent for Grades 8 through 12. For the portion of the curve from Grades 2 through 6, the grades of particular interest for this study, the configurations of the curves for the two tests appear quite similar at the 50th and 30th percentile levels. There appears to be a marked difference in the variation across percentiles for the two tests, especially from the 50th to the 10th percentile level. The difference between the 50th percentile scores and the 10th percentile scores is greater for the MAT than for the CAT both in absolute scale units and relative to the growth expected from one grade to the next.

Table 6. CAT and MAT Rescaled Scale Scores

Subj.	Gr.	CAT			MAT		
		10%-ile	30%-ile	50%-ile	10%-ile	30%-ile	50%-ile
Reading	K	350	391	421	3	97	161
	1	436	486	519	244	337	405
	2	500	560	600	436	525	600
	3	549	614	659	508	602	670
	4	596	666	719	569	660	728
	5	636	711	761	614	706	768
	6	664	743	800	621	725	800
	7	691	777	830	638	752	829
	8	724	813	876	643	779	865
	9	747	840	906	692	815	892
	10	773	869	937	744	862	933
	11	796	904	970	778	885	954
	12	807	917	990	815	918	986
Math	K	447	466	483	316	369	412
	1	494	524	541	410	463	506
	2	538	577	600	490	559	600
	3	587	632	660	544	615	663
	4	629	678	709	594	675	722
	5	662	721	760	641	724	766
	6	695	758	800	660	747	800
	7	705	787	837	695	784	837
	8	753	833	892	714	804	867
	9	768	860	927	738	822	880
	10	791	886	958	765	844	898
	11	812	908	974	785	865	910
	12	812	912	986	800	881	926

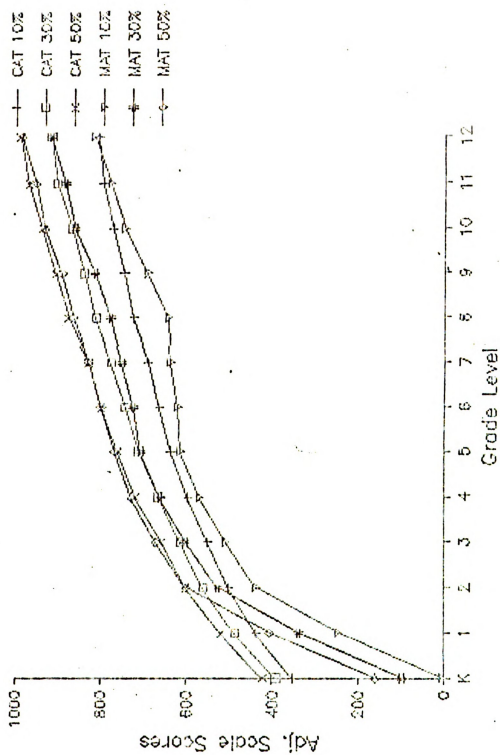


Figure 6. Rescaled Reading Growth Curves: CAT vs MAT

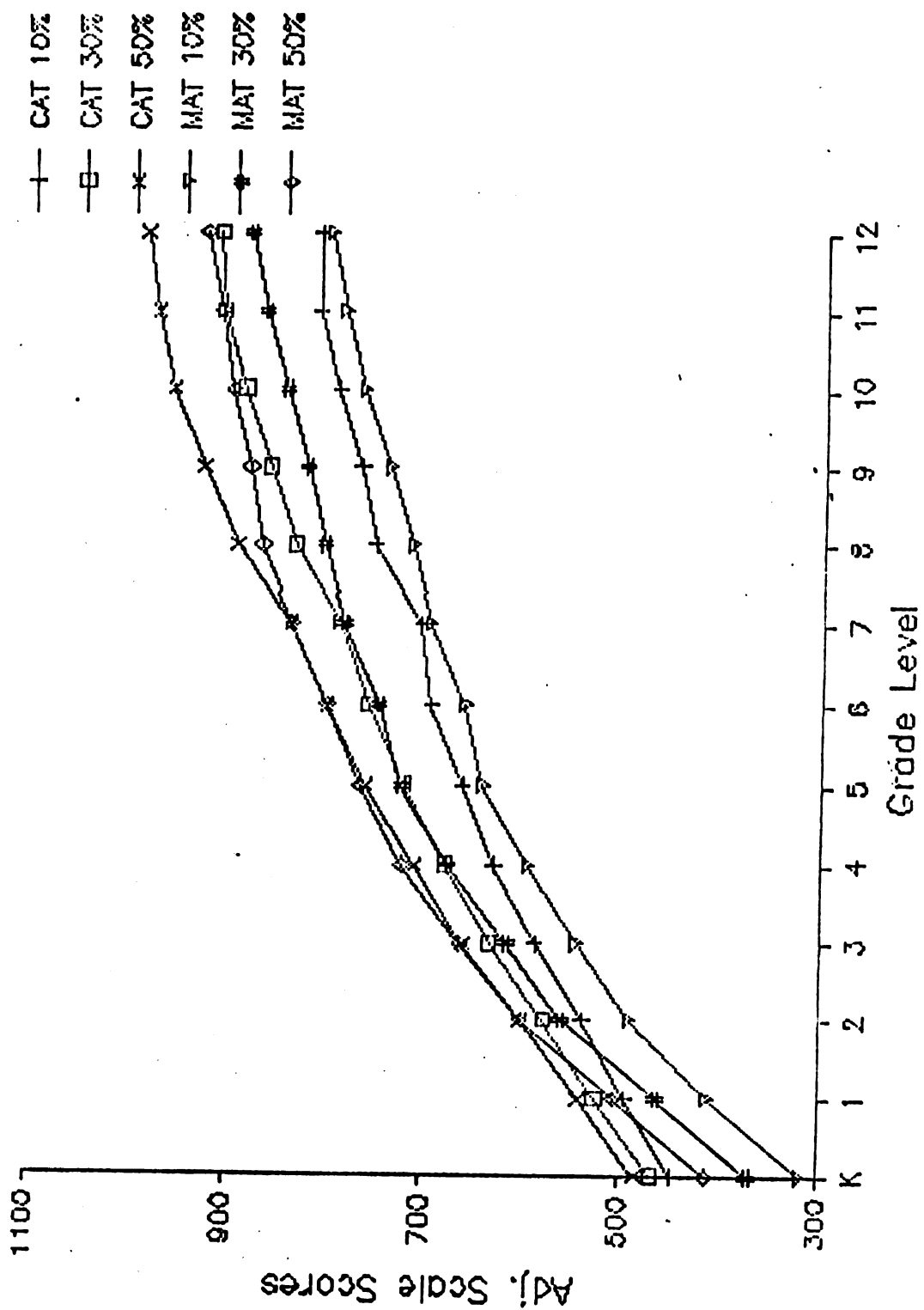


Figure 7. Rescaled Math Growth Curves: CAT vs MAT

Analysis of the growth curves for the CAT and MAT

Table 7 shows the results of the comparison of the growth curves for the two tests. The values for r_{xy} range from .9653 to .9917. Those for $\hat{\rho}$ vary from .9768 to .9941. These high values can be attributed to several factors. First, nearly all of the variation in the scores for any single curve is the result of yearly gains in achievement. Second, there is only one point on each curve at each grade level. And finally, the scores comprising each curve were obtained from national norms tables based on scores for several thousand students.

Table 7. Differences Between CAT and MAT Growth Curves

Subject /%-ile	r_{xy}	$\hat{\rho}$	χ_1	$P(\chi_1)$	\bar{d}	Se_d	t_{12}	$P_t(12)$

Reading								
50	.9711	.9806	8.3391	<.0001**	30.54	21.08	1.45	.173
30	.9720	.9813	9.0148	<.0001**	48.31	23.05	2.10	.058
10	.9653	.9768	8.1262	<.0001**	74.15	26.34	2.82	.016*
Mathematics								
50	.9831	.9884	6.4193	<.0001**	26.15	8.67	3.02	.011*
30	.9868	.9910	7.2590	<.0001**	30.00	7.62	3.94	.002**
10	.9917	.9941	5.0935	<.0001**	41.62	9.06	4.59	.001**

For an overall level of $\alpha = .05$ the levels for the three percentile ranks must be individually set at $\alpha = .017$. With regard to the configurations the probabilities of the growth curves being the same for the two tests at the three percentile ranks in reading and in

math are all less than .001. The growth curves for the CAT and MAT were found to be significantly different in shape. The test that the levels of the growth curves for reading and math for the two tests are the same showed significant differences at the 10th percentile in reading and at all levels in math ($p < .017$).

These results indicate that the hypothesis that the growth curves for the two tests are the same can be rejected.

Comparison of National and Local Norms

Table 8 shows the comparison of scale scores for the growth curves for the six local districts from Grade 2 through Grade 6 in reading and mathematics with the national norms. The most readily apparent difference between the local and national norms is that a number of the local growth curves are much higher than those based on the national norms. This is made clearer in Table 9 which shows the national percentile equivalents of the local percentile ranks.

In reading the scores for all three districts using the CAT are above the national norms at all percentile ranks. COL 10th percentile scores nearly match those for the national CAT 30th percentile while CLS scores exceed them. Both COL and CLS 30th percentile students achieved higher scores than the 50th percentile nationally. In several other cases the local 30th percentile scores approach those for the national 50th percentile. Figure 8

Table 8. National and Local Growth Curves

Reading									
X-file Rank*	Gr.	Scale Score Equivalents							
		CAT	COL	CIM	CLS	MAT	MOL	MIM	MLS
10	2	290	332	322	345	524	545	561	510
	3	324	365	351	376	566	582	607	546
	4	357	404	389	416	602	631	637	616
	5	385	434	427	441	628	658	660	627
	6	405	458	431	464	632	681	686	675
30	2	332	363	351	370	576	596	608	571
	3	370	399	393	404	621	629	647	630
	4	406	440	426	446	655	655	683	653
	5	438	476	460	476	682	697	693	688
	6	460	506	471	506	693	721	725	717
50	2	360	385	372	392	620	624	644	625
	3	401	423	412	432	661	645	670	664
	4	443	470	449	466	695	691	719	685
	5	473	505	481	496	718	723	718	703
	6	500	540	500	532	737	751	769	747

Mathematics									
10	2	309	326	338	337	403	464	464	425
	3	343	361	362	372	454	498	526	484
	4	372	394	389	410	501	553	578	511
	5	395	429	422	438	546	579	619	591
	6	418	452	442	468	564	604	637	639
30	2	336	347	356	354	468	501	507	478
	3	374	386	387	396	521	541	575	531
	4	406	423	414	429	578	588	626	569
	5	436	458	452	459	624	628	658	632
	6	462	484	473	500	646	676	704	678
50	2	352	355	373	365	507	525	541	507
	3	394	403	406	407	567	571	608	569
	4	428	441	432	443	622	618	666	617
	5	463	477	474	478	664	660	688	656
	6	491	509	495	524	696	721	741	722

*National percentile ranks for CAT and MAT;
local percentile ranks for all others.

Table 9. Comparison of National and Local Percentile Ranks

Reading

Local		National Percentile Rank Equivalents					
%-ile	Gr.	COL	CIM	CLS	MOL	MIM	MLS
10	2	30	22	38	18	24	6
	3	26	20	33	15	24	5
	4	28	21	34	20	22	14
	5	28	25	32	19	20	10
	6	29	17	32	25	27	23
30	2	51	42	56	38	44	28
	3	49	44	52	34	43	34
	4	48	41	52	30	44	29
	5	52	42	52	38	36	33
	6	53	35	53	42	44	40
50	2	67	58	71	52	62	52
	3	65	58	70	42	56	52
	4	67	54	64	49	62	45
	5	70	56	65	53	50	42
	6	70	50	66	57	66	55

Mathematics

10	2	20	32	30	28	28	15
	3	20	21	29	23	32	18
	4	22	19	34	22	30	12
	5	25	21	31	16	28	19
	6	25	20	34	18	27	28
30	2	42	56	51	47	50	35
	3	41	42	51	38	54	34
	4	44	37	49	34	52	27
	5	45	41	46	32	47	34
	6	44	37	55	42	53	43
50	2	55	73	64	60	68	50
	3	58	61	62	52	68	51
	4	59	52	61	48	70	48
	5	60	58	61	48	60	46
	6	61	52	71	60	68	60

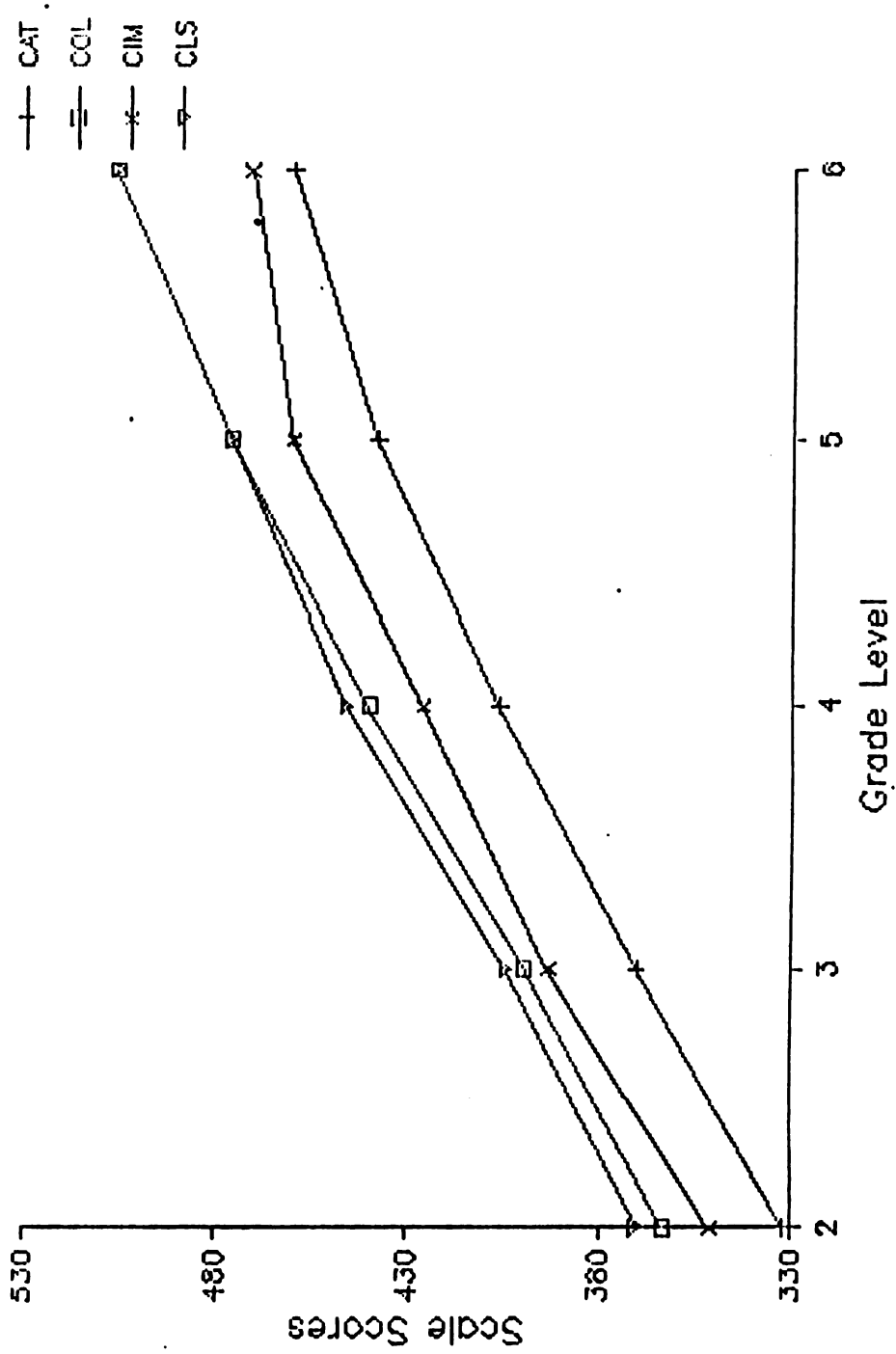


Figure 8. National and Local Growth Curves: CAT - Reading

shows the 30th percentile reading growth curves for the districts using the CAT.

The reading scores for two (MOL and MIM) of the three districts using the MAT are above the national norms for the 10th and 30th percentile ranks. MIM scores are above the national norms for the 50th percentile as well. Grade 6 scores for all three districts are higher than the national norms for all percentile ranks. Figure 9 shows the 30th percentile reading growth curves for the districts using the MAT.

In mathematics the scores for all three districts using the CAT are also above the national norms at all percentile ranks. CLS 10th percentile scores exceed those for the national CAT 30th percentile and CLS 30th percentile scores exceed those for the national 50th percentile. Figure 10 shows the 30th percentile mathematics growth curves for the districts using the CAT.

Mathematics scores for all three districts using the MAT are above the national norms at the 10th and 30th percentiles. For MIM they also exceed the national norms for the 50th percentile rank. MIM 10th percentile scores in mathematics are close to the national MAT 30th percentile norms and MIM 30th percentile scores are close to the national 50th percentile norms. Figure 11 shows the 30th percentile mathematics growth curves for the districts using the MAT.

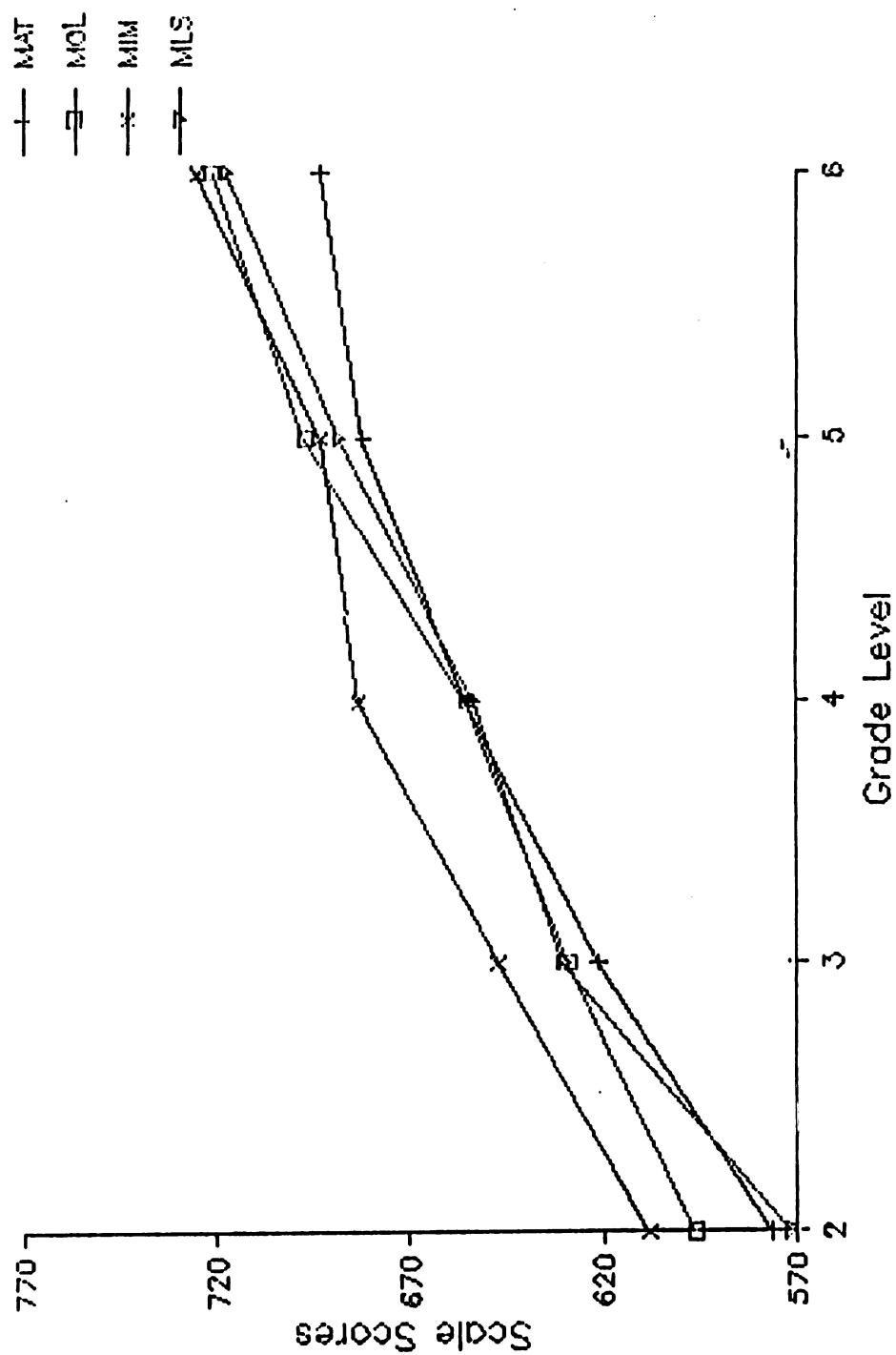


Figure 9. National and Local Growth Curves: MAT - Reading

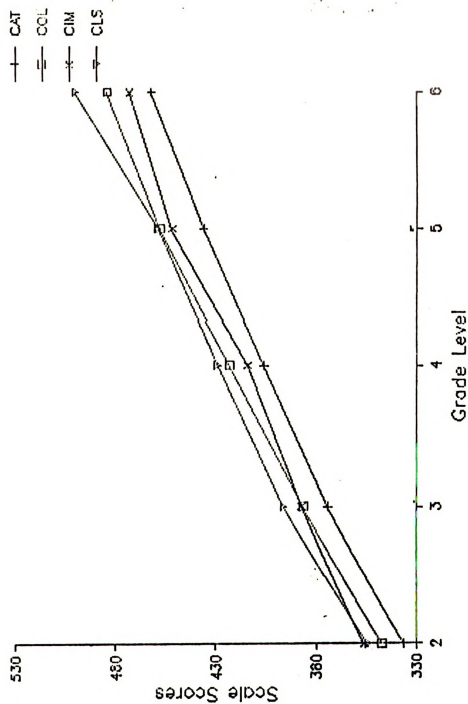


Figure 10. National and Local Growth Curves: CAT - Math

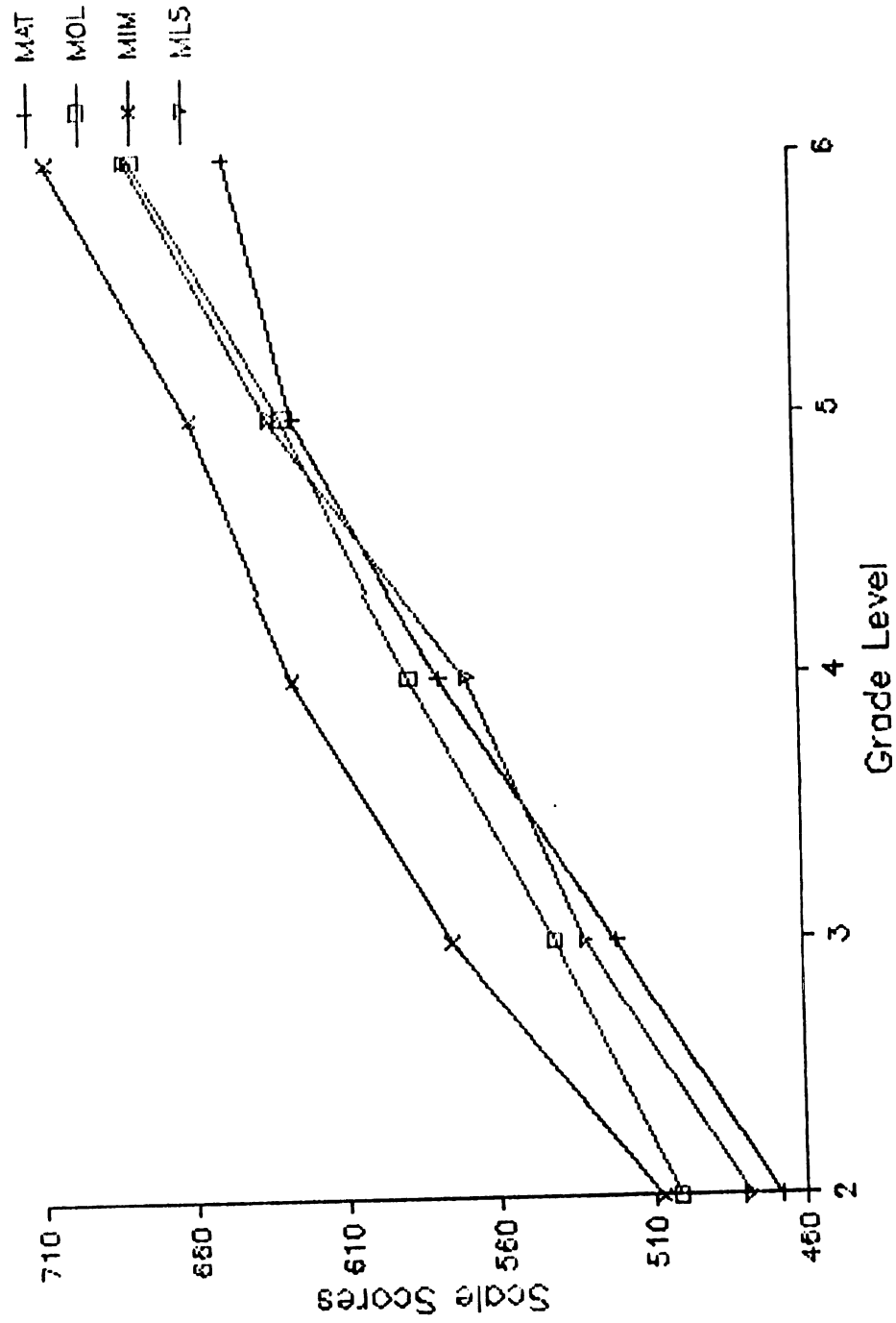


Figure 11. National and Local Growth Curves - MAT - Math

For graphs showing the growth curves for each of the six districts for the three percentile ranks see Appendix A.

Stability of Local Norms

Since three years' data were collected from each of the districts it was possible to examine the amount of yearly variation in the local growth curves. Table 10 shows the average range in the scale scores across percentile ranks for the three years' data. There is clearly more year-to-year variation in the norms for the districts using the MAT than for those using the CAT. Part but not all of this difference in variation can be accounted for by the difference in the scales on which the scale scores for the two tests are based. In terms of NCE's the yearly variation in scale scores for districts using the CAT amounts to 2-5 NCE's. The variation for districts using the MAT amounts to differences of 6-12 NCE's from year to year. See Appendix B for tables showing the yearly data for each of the six districts.

Effect of Local Growth Patterns on Gains Measured by the Norm-referenced Model

Effects on Expected No-treatment Growth

Table 11 shows the apparent spring-to-spring no-treatment gains calculated under the model by grade and

**Table 10. Size of District and Stability of Local Norms
(Average Yearly Range of Scale Scores*)**

Reading							
District	Size**	Grade Level				Overall	
		2	3	4	5	6	Average
COL	898.47	8.00	3.67	8.33	10.00	8.00	7.60
CIM	271.20	8.67	17.00	7.67	7.33	14.33	11.00
CLS	164.20	6.33	8.67	9.33	15.00	15.33	10.93
MOL	63.00	48.33	51.67	19.33	30.67	N/A	37.50
MIM	82.80	18.67	27.00	19.33	11.67	38.00	22.93
MLS	34.00	35.67	57.00	84.33	15.00	8.67	40.13

Mathematics							
COL	898.20	4.33	2.33	4.00	6.33	5.33	4.47
CIM	268.80	10.00	8.67	8.00	8.33	9.33	8.87
CLS	164.20	9.00	7.33	7.00	14.00	16.00	10.67
MOL	63.25	40.67	25.00	14.67	24.67	N/A	26.25
MIM	81.00	29.00	34.33	27.67	40.00	41.00	34.40
MLS	34.13	34.00	30.33	60.33	28.00	32.33	37.00

 * Range of scale scores for each grade and percentile rank
 averaged across percentile ranks.

** Avg. number of students per grade per year

Table 11. Significance of Expected Zero Gain Scores
(t Based on n=30, SD=13.92)

Reading					
Grade	CAT		MAT		
	NCE Gain	t(29)	NCE Gain	t(29)	
3	-2.28	-.8971	-2.03	-.7988	
4	.49	.1928	2.25	.8853	
5	.82	.3227	.62	.2440	
6	-1.23	-.4840	5.58	2.1956*	

Mathematics					
	COL		MOL		
	NCE Gain	t(29)	NCE Gain	t(29)	
3	1.30	.5115	-4.67	-1.8375*	
4	.00	.0000	-9.23	-3.6318**	
5	1.83	.7201	-.37	-.1456	
6	-.20	-.0787	1.78	.7004	

	CIM		MIM		
	NCE Gain	t(29)	NCE Gain	t(29)	
3	-5.33	-2.0972*	2.54	.9994	
4	-3.67	-1.4441	.60	.2361	
5	2.90	1.1411	-1.63	-.6414	
6	-3.09	-1.2158	.53	.2085	

	CLS		MLS		
	NCE Gain	t(29)	NCE Gain	t(29)	
3	1.59	.6256	1.00	.3935	
4	-.44	-.1731	-3.81	-1.4992	
5	-2.56	-1.0073	3.47	1.3654	
6	3.36	1.3221	8.94	3.5177**	

* Significant if $\alpha = .05$					
** Significant if $\alpha = .01$					

test for reading and by grade and district for mathematics. In reading the only significant differences found in the expected growth rates for the local and national norms were between tests at Grade 6. Assuming that they grew in reading achievement at the local district rates with no special treatment Grade 6 students tested with the MAT would be expected to show an average gain of nearly 6 NCE's if evaluated with the norm-referenced model. Given the differences between the local and national norms the resulting probability of making a Type I error with $\alpha = .05$ is greater than .6. Figure 12 shows the expected reading gains by grade for each test. The difference between the expected gains at Grade 6 shows up clearly.

In mathematics significant differences in the growth rates were found for MLS at Grade 6, for MOL and CIM at Grade 3 and for MOL at Grade 4. Assuming that they grew in mathematics achievement at their local district rates both MOL and CIM Grade 3 students would show a loss of approximately 5 NCE's relative to the national norms. MOL students in Grade 4 would show an even greater loss(-9 NCE's) relative to the national norms. Given these differences between the local and national norms the probability of making a Type I error drops to less than .01. There is an accompanying loss of power so that for local programs with 30 or fewer students even true gains of as much as 4 NCE's would be

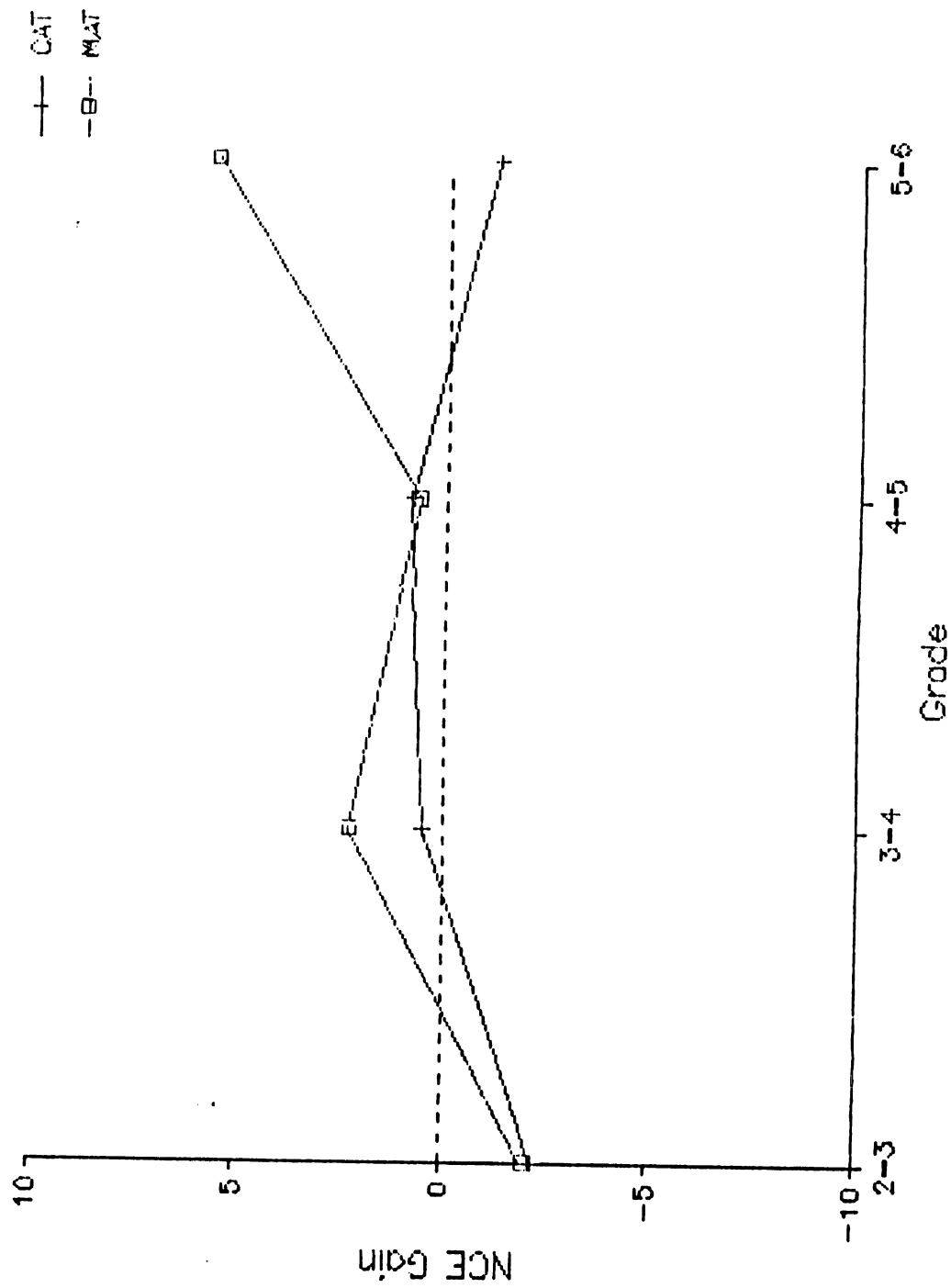


Figure 12. Expected Zero Gain Scores: Reading

detected with a probability of less than .05.

In contrast MLS Grade 6 mathematics students with no special treatment would be expected to show an annual gain of nearly 9 NCE's. With this difference between the local and national norms a Type I error would be expected to occur with a probability greater than .9.

The finding of significant differences in mathematics was not linked either to test used or district type. Figure 13 shows the expected mathematics gains by grade for the districts using the CAT. Figure 14 shows the same for those districts using the MAT. It would appear differences in expected no-treatment gains in mathematics are related to variables specific to the individual districts but not measured in this study. Figures 15, 16 and 17 show the expected gains for the districts using out-of-level testing, those with improving MEAP scores and those with a high proportion of low SES students respectively.

Effects on Likelihood of Type I Errors

From Table 11 it can be seen that there are two conditions where significant differences between the local and national norms would result in an increased likelihood of Type I errors. For Grade 6 reading programs evaluated using the MAT, the differences between the local and national norms result in an expected

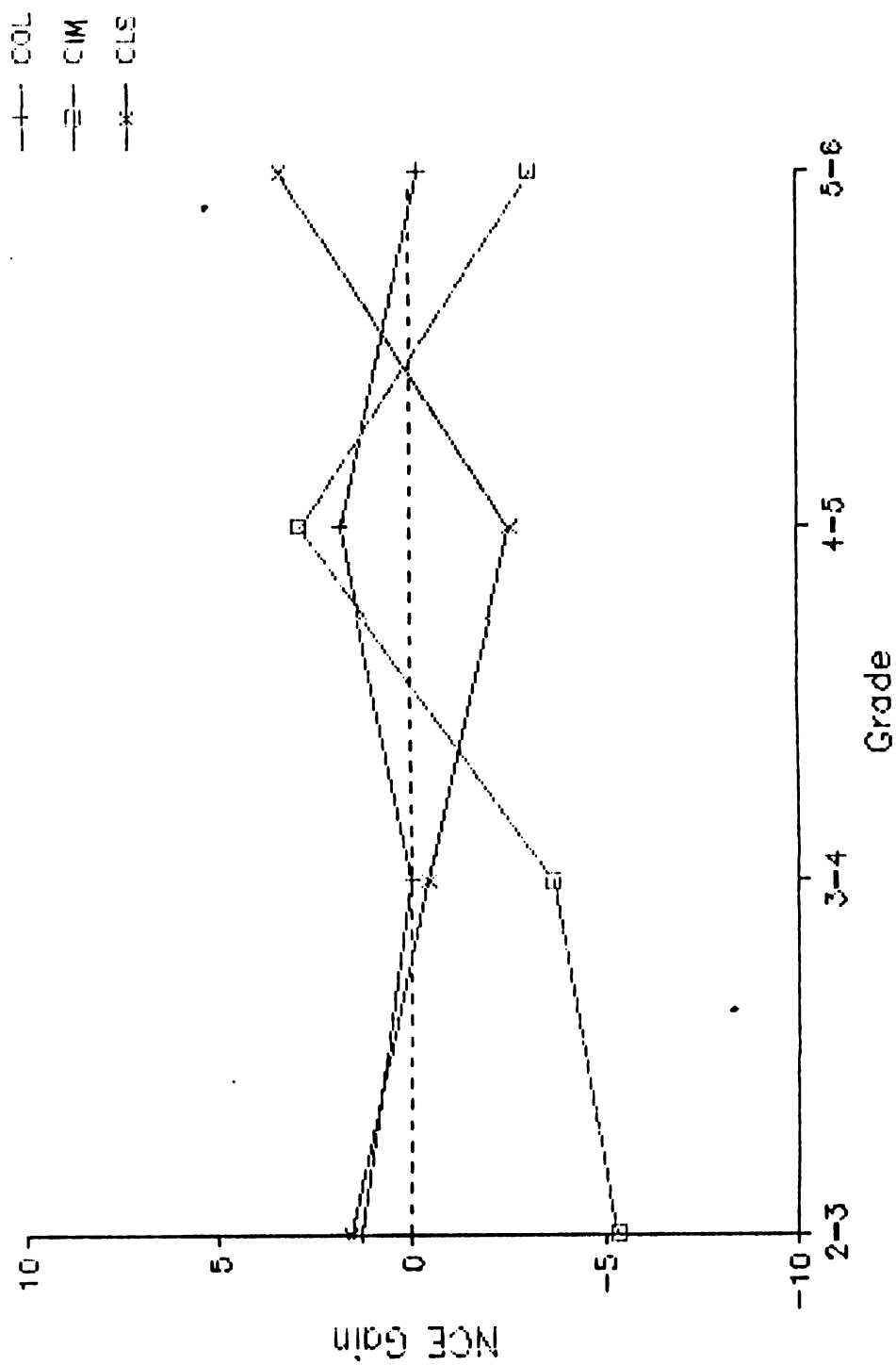


Figure 13. Expected Zero Gain Scores in Math: CAT

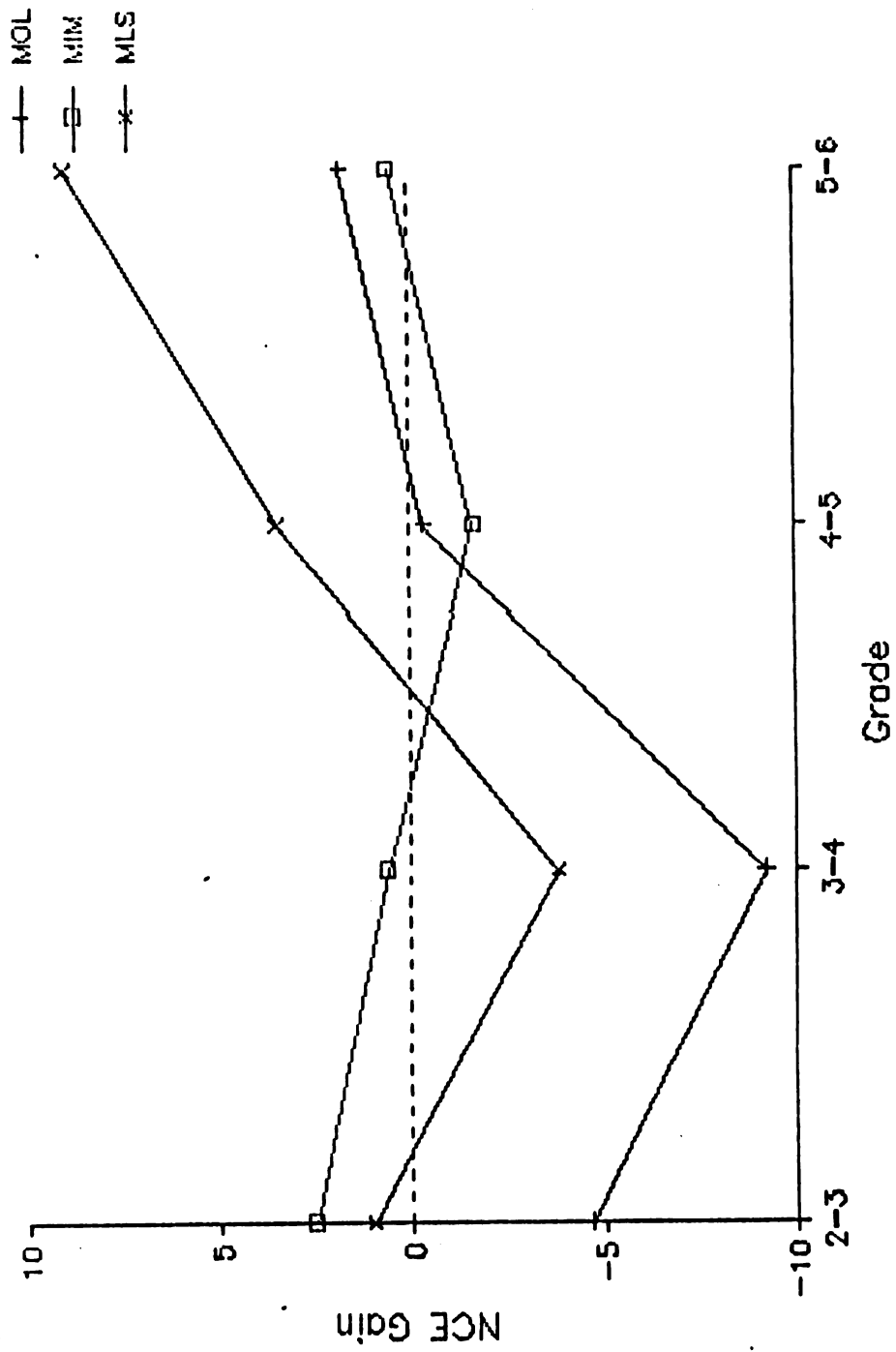


Figure 14. Expected Zero Gain Scores in Math: MAT

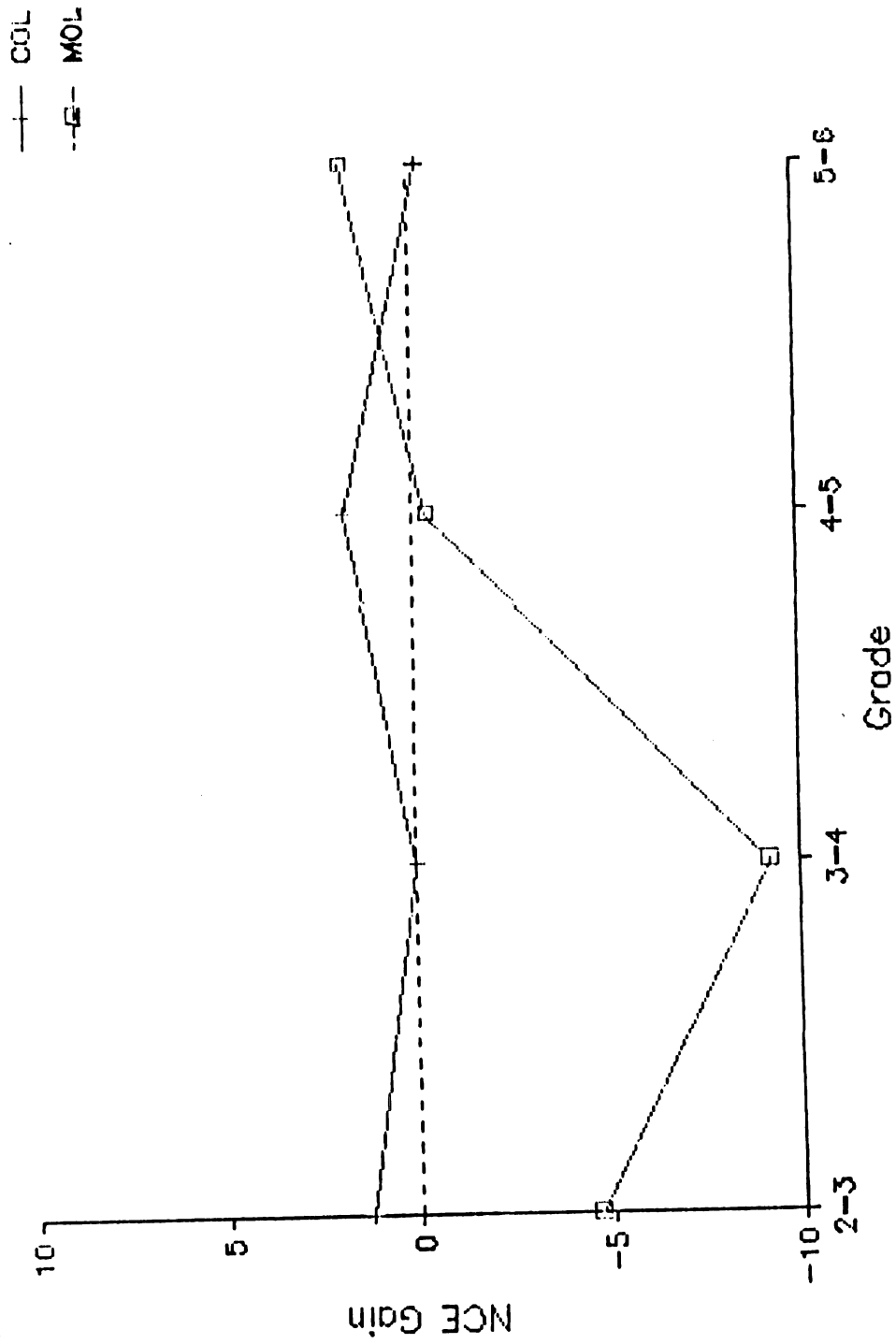


Figure 15. Expected Zero Gain Scores in Math: Out-of-Level Districts

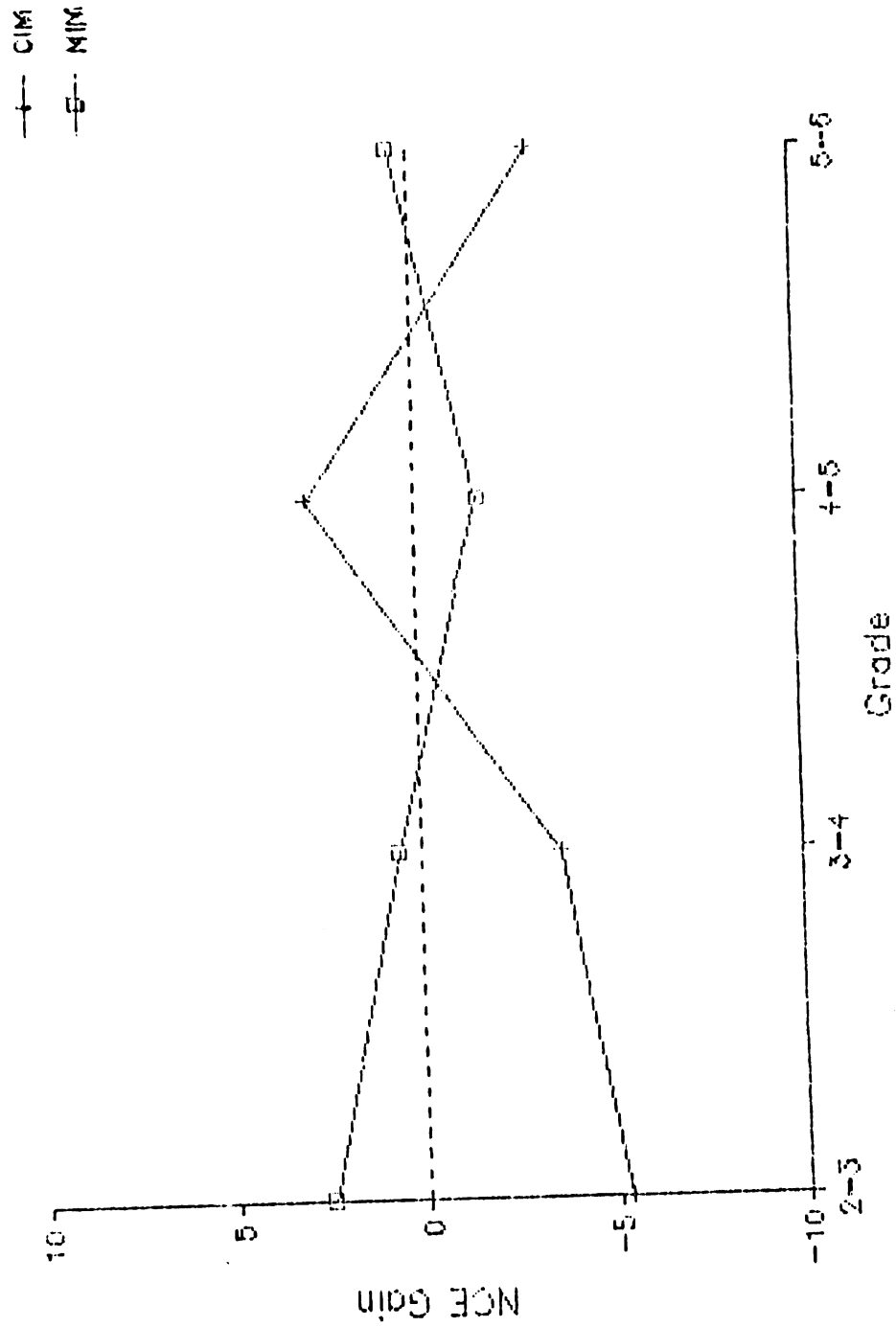


Figure 16. Expected Zero Gain Scores in Math: Improving MEAP Districts

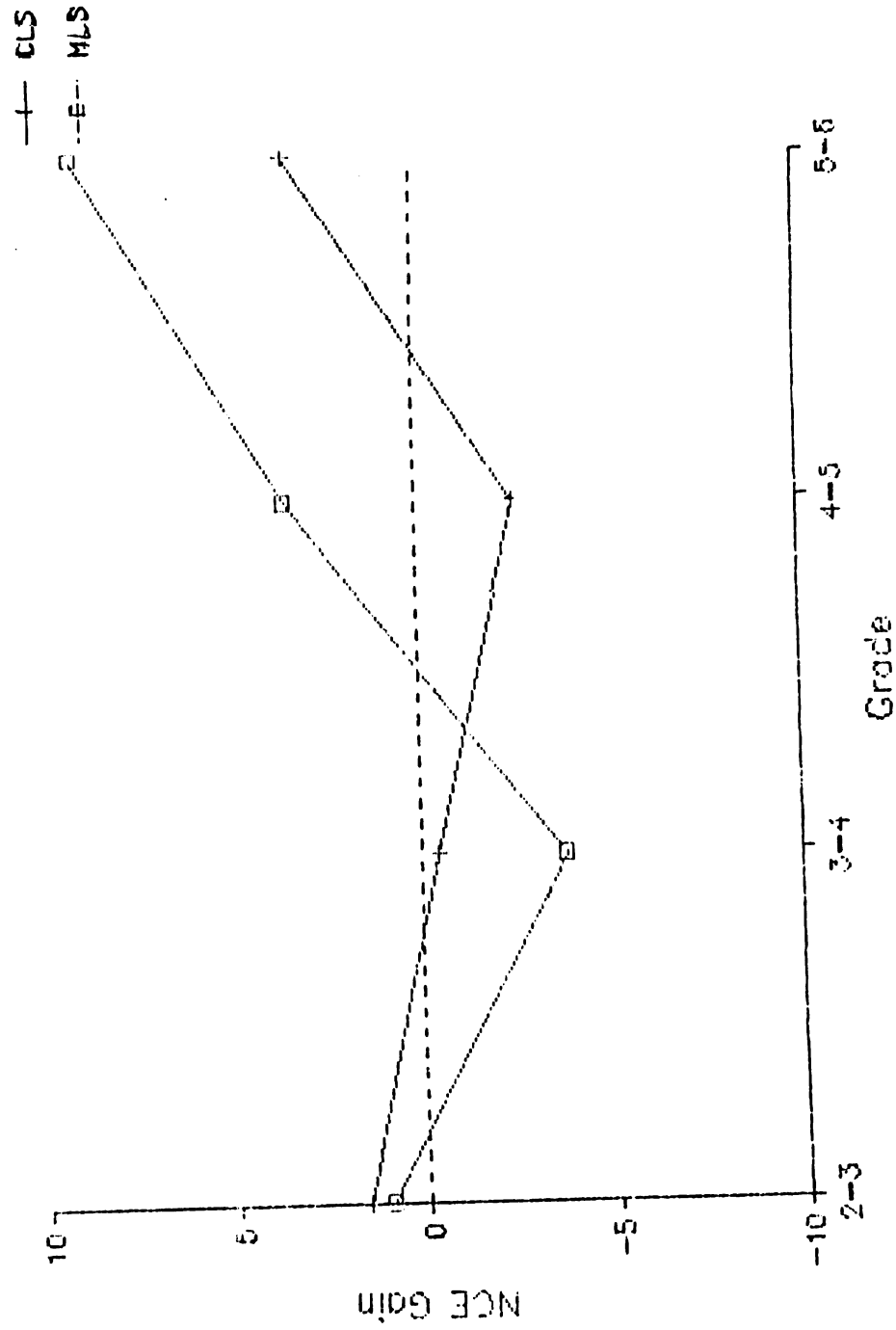


Figure 17. Expected Zero Gain Scores in Math: Low SES Districts

apparent gain of 5.58 NCE's. Given this difference and assuming that the measured mean gains would have a $t_{(29)}$ distribution, the probability of a measured gain high enough to result in a Type I error where $n = 30$ and $\alpha = .05$ is $p > .60$.

For MLS Grade 6 mathematics programs the result of the differences between the local and national norms is an expected apparent gain of 8.94 NCE's. Given this difference the probability of a measured gain high enough to result in a Type I error given $n = 30$ and $\alpha = .05$ is greater than .90.

Effects on Power to Detect True Gains

It can also be seen that significant differences from zero will affect the power of the model to detect true gains when they are present. A true gain of 1 NCE will rarely be detected ($p < .15$) with groups of 30 or fewer students. However, if the pattern of local growth is such that apparent positive gains are expected without true additional gains then the likelihood of detecting a true 1 NCE gain as a positive program effect is greatly increased. Table 12 shows the expected observed gains and the probabilities of detecting true gains of 1, 4 and 7 NCE gains with $n=30$ and α levels of .05 and .01 for the local data in Table 11. Only in those cases where Type I errors are likely is there a high probability of detecting a true 1 NCE gain. The probability of detecting

Table 12. Effect on Power of Local Growth Patterns

Reading

Tst/		0 NCE	+1 NCE	<u>p, α =</u>		+4 NCE	<u>p, α =</u>		+7 NCE	<u>p, α =</u>	
Dst	Gr	Gain	Gain	.05	.01	Gain	.05	.01	Gain	.05	.01
=====											
CAT	3	-2.28	-1.28	<.05	<.01	1.72	.2	<.05	4.72	.6	.3
	4	.49	1.49	.1	<.05	4.49	.5	.2	7.49	.9	.7
	5	.82	1.82	.2	<.05	4.82	.6	.3	7.82	.9	.7
	6	-1.23	-.23	<.05	<.01	2.77	.3	.1	5.77	.7	.4

MAT	3	-2.03	-1.03	<.05	<.01	1.97	.2	.1	4.97	.6	.3
	4	2.25	3.25	.3	.1	6.25	.8	.5	9.25	>.95	.9
	5	.62	1.62	.2	<.05	4.62	.5	.3	7.62	.9	.7
	6	5.58	6.58	.8	.6	9.58	>.95	.9	12.58	>.99	>.99
=====											
Mathematics											
=====											
COL	3	1.30	2.30	.2	.1	5.30	.6	.4	8.30	.9	.8
	4	.00	1.00	.1	<.05	4.00	.5	.2	7.00	.8	.6
	5	1.83	2.83	.3	.1	5.83	.7	.4	8.83	>.95	.8
	6	-.20	.80	.1	<.05	3.80	.4	.2	6.80	.8	.6

MOL	3	-4.67	-3.67	<.01	<.01	-.67	<.05	<.01	2.33	.2	.1
	4	-9.23	-8.23	<.01	<.01	-5.23	<.01	<.01	-2.23	<.01	<.01
	5	-.37	.63	.1	<.05	3.63	.4	.2	6.63	.8	.6
	6	1.78	2.78	.3	.1	5.78	.7	.4	8.78	>.95	.8

CIM	3	-5.33	-4.33	<.01	<.01	-1.33	<.05	<.01	1.67	.2	<.05
	4	-3.67	-2.67	<.01	<.01	.33	.1	<.05	3.33	.4	.1
	5	2.90	3.90	.4	.2	6.90	.8	.6	9.90	>.95	.9
	6	-3.09	-2.09	<.01	<.01	.91	.1	<.05	3.91	.4	.2

MIM	3	2.54	3.54	.4	.2	6.54	.8	.5	9.54	>.95	.9
	4	.60	1.60	.2	<.05	4.60	.5	.3	7.60	.9	.7
	5	-1.63	-.63	<.05	<.01	2.37	.2	.1	5.37	.7	.4
	6	.53	1.53	.1	<.05	4.53	.5	.3	7.53	.9	.7

CLS	3	1.59	2.59	.3	.1	5.59	.7	.4	8.59	.9	.8
	4	-.44	.56	.1	<.05	3.56	.4	.2	6.56	.8	.5
	5	-2.56	-1.56	<.05	<.01	1.44	.1	<.05	4.44	.5	.2
	6	3.36	4.36	.5	.2	7.36	.9	.7	10.36	>.95	.9

MLS	3	1.00	2.00	.2	.1	5.00	.6	.3	8.00	.9	.8
	4	-3.81	-2.81	<.01	<.01	.19	.1	<.05	3.19	.3	.1
	5	3.47	4.47	.5	.2	7.47	.9	.7	10.47	>.95	.9
	6	8.94	9.94	>.95	.9	12.94	>.99	>.99	15.94	>.99	>.99
=====											

a 1 NCE gain in Grade 6 reading with the MAT is .8 if $\alpha = .05$ and .6 for $\alpha = .01$. A 1 NCE gain made by MLS students in Grade 6 would be detected with $p > .95$ when $\alpha = .05$ and with $p = .9$ for $\alpha = .01$.

Even the likelihood of detecting a true 4 NCE gain is small with a group of 10 to 20 students. However, a Chapter 1 program that produces a 4 NCE gain on top of an expected 4 NCE gain from the regular program has a high probability of being detected. With an n of 30 the probability of detecting a true 4 NCE gain is 0.5 if $\alpha = .05$ and .2 when $\alpha = .01$. In those cases where a Type I error is likely, a true 4 NCE gain will be detected most of the time. The probability of detecting a 4 NCE gain with the MAT in Grade 6 reading is $> .95$ if $\alpha = .05$ and .9 for $\alpha = .01$. A true 4 NCE gain in mathematics by Grade 6 MLS students will be detected with $p > .99$ with an α level of either .05 or .01. Where the expected no-treatment gain is negative the probability of detecting a true 4 NCE gain is small. In mathematics for MOL and CIM Grade 3 students the probability is $< .05$ when $\alpha = .05$ and $< .01$ if $\alpha = .01$. For MOL Grade 4 mathematics the probability is $< .01$ for an α level of either .05 or .01.

A 7 NCE gain, in contrast, has a high probability of being detected for a group of 20 or 30 students and is likely to be identified even for as few as 10 students provided the expected rate of growth for the group is at

least as great as that for the national norm group. Where the expected no-treatment gain is negative even a true 7 NCE gain in Chapter 1 students' achievement is unlikely to be detected. For MOL Grade 4 mathematics students the probability of detecting such a gain remains $<.01$ for either $\alpha = .05$ or $\alpha = .01$. For MOL and CIM Grade 3 mathematics a true 7 NCE gain would be detected only with $p = .2$ if $\alpha = .05$ and with $p = .1$ and $p < .05$ respectively if $\alpha = .01$.

Effects on Size of Sample Needed to Detect True Gains

Table 13 shows the effect of local growth patterns on the size of the sample needed to detect true gains of 1.0, 4.0 and 7.0 NCE's with $\alpha = .05$ and $.01$. If the expected no-treatment gain plus the true gain results in an expected observed gain that is negative the positive program gains will not be detected no matter how large the sample involved. Where the expected observed gain in NCE's is positive but less than 1.0 the size of the sample needed is very large. However, where the expected observed gain is large (7.0 NCE's or more) the sample needed may be less than 10.

Comparison of Longitudinal Data with Cross-Sectional Norms

Table 14 shows the mean NCE scores for each of the three cohorts (Cohort 1 - students tested in Grades 2, 3 and 4; Cohort 2 - students tested in Grades 3, 4 and 5;

Table 13. Effect on Sample Size of Local Growth Patterns

Reading											
Tst/	0 NCE	+1 NCE	$\frac{n}{\sigma} =$		+4 NCE	$\frac{n}{\sigma} =$		+7 NCE	$\frac{n}{\sigma} =$		
Dst	Gr	Gain	Gain	.05 .01	Gain	.05 .01	Gain	.05 .01	Gain	.05 .01	
CAT	3	-2.28	-1.28	---	---	1.72	178	356	4.72	25	50
	4	.49	1.49	238	474	4.49	28	55	7.49	11	22
	5	.82	1.82	159	318	4.82	24	48	7.82	11	20
	6	-1.23	-.23	---	---	2.77	70	139	5.77	18	35
MAT	3	-2.03	-1.03	---	---	1.97	138	271	4.97	23	46
	4	2.25	3.25	51	103	6.25	15	30	9.25	8	16
	5	.62	1.62	201	401	4.62	27	52	7.62	11	21
	6	5.58	6.58	14	28	9.58	8	15	12.58	6	10

Mathematics

COL	3	1.30	2.30	101	199	5.30	21	40	8.30	10	19
	4	.00	1.00	528	1052	4.00	35	69	7.00	13	25
	5	1.83	2.83	67	134	5.83	17	34	8.83	9	17
	6	-.20	.80	824	1644	3.80	38	76	6.80	13	26
MOL	3	-4.67	-3.67	---	---	-.67	---	---	2.33	98	194
	4	-9.23	-8.23	---	---	-5.23	---	---	-2.23	---	---
	5	-.37	.63	1329	2650	3.63	42	83	6.63	14	27
	6	1.78	2.78	70	138	5.78	18	35	8.78	9	17
CIM	3	-5.33	-4.33	---	---	-1.33	---	---	1.67	189	377
	4	-3.67	-2.67	---	---	.33	4844	9660	3.33	49	98
	5	2.90	3.90	36	72	6.90	13	25	9.90	7	14
	6	-3.09	-2.09	---	---	.91	637	1270	3.91	36	72
MIM	3	2.54	3.54	44	88	6.54	14	28	9.54	8	15
	4	.60	1.60	206	411	4.60	27	53	7.60	11	21
	5	-1.63	-.63	---	---	2.37	95	187	5.37	20	40
	6	.53	1.53	225	449	4.53	28	54	7.53	11	22
CLS	3	1.59	2.59	81	158	5.59	19	37	8.59	9	17
	4	-.44	.56	1682	3354	3.56	43	87	6.56	14	28
	5	-2.56	-1.56	---	---	1.44	254	507	4.44	28	57
	6	3.36	4.36	29	58	7.36	12	23	10.36	7	13
MLS	3	1.00	2.00	133	263	5.00	23	45	8.00	10	20
	4	-3.81	-2.81	---	---	.19	14613	29140	3.19	53	107
	5	3.47	4.47	28	56	7.47	11	22	10.47	7	13
	6	8.94	9.94	7	14	12.94	5	10	15.94	4	8

Table 14. Longitudinal Data

Cohort	N	Mean Scores				Gains	
		1983	1984	1985	1983-84	1984-85	1983-85

- COL - Reading							
1	72	51.59	54.78	54.35	3.19	-.43	2.76
2	201	57.00	57.55	58.80	.55	1.25	1.80
3	191	56.38	57.75	59.04	1.37	1.29	2.66
Total	464	55.91	57.20	58.21	1.29	1.01	2.30**
- Mathematics							
1	72	55.48	59.38	60.31	3.90*	.93	4.83*
2	201	60.55	64.74	65.65	4.19**	.91	5.10**
3	191	60.17	64.67	68.44	4.50**	3.77**	8.27**
Total	464	59.61	63.88	65.97	4.27**	2.09*	6.36**

- MOL - Reading							
1	29	66.81	57.23	64.52	-9.58**	7.29*	-2.29
2	28	59.99	59.33	60.75	-.66	1.42	.76
3	33	57.68	62.54	64.11	4.86	1.57	6.43*
Total	90	61.34	59.83	63.20	-1.51	3.37	1.86
- Mathematic							
1	29	63.58	62.72	61.89	-.86	-.83	-1.69
2	28	58.53	60.77	55.13	2.24	-5.64*	-3.40
3	33	55.17	58.72	60.02	3.55	1.30	4.85*
Total	90	58.93	60.65	59.10	1.72	-1.55	.17
=====							

* Significant if $\alpha = .05$ ** Significant if $\alpha = .01$

and Cohort 3 - students tested in Grades 4, 5 and 6) for the two districts over the three-year period for which data were collected.

For COL there were small positive gains(1.3 and 1.0 NCE's) in reading for each of the two years. Individually these gains were not significant but they were both positive and when taken together resulted in a significant overall two-year gain. In mathematics significant gains of approximately 4 NCE's were found for all cohorts the first year and for Cohort 3 the second year as well. The mathematics gains were all positive resulting in significant overall two-year gains for all cohorts and the total sample.

For MOL the results were mixed. In reading Cohort 1 showed a significant loss the first year followed by a significant gain the second year. The resulting two year loss was not significant. This may be a reflection of the year-to-year variation that was apparent in the yearly data particularly for the smaller districts. (See Table 11 and Appendix B.) MOL Cohort 3 showed nonsignificant positive gains for the two years individually that added up to a significant total gain. In mathematics the results showed no significant overall change.

CHAPTER VI

DISCUSSION

This study set out to answer certain questions related to the validity of the norm-referenced model and its usefulness in evaluating local programs. In this chapter the results of the study and their implications relative to these questions will be discussed.

Assumptions the Norm-referenced Model Makes about Natural Growth

An examination of the norm-referenced model and its assumptions found that they do yield a method for depicting graphically the patterns of growth implied by the model and a test's norms. The model is based on the equipercentile assumption that without intervention students' percentile rank standings would remain constant from year to year. It further assumes that percentile ranks can be converted to NCE's and to a test's scale scores. When given percentile ranks were converted into corresponding scale scores at each grade level it was possible to depict graphically expected growth curves for the two tests examined in this study, the CAT and the MAT. The resulting growth curves for both tests in reading and mathematics are curvilinear with the most rapid growth in

the lower grades and decreasing rates of growth in the higher grades. This indicates that the skills measured by these tests are learned primarily in the lower grades.

Comparison of Growth Curves for the CAT and MAT

Comparison of the graphs indicates the growth curves for the CAT and MAT are not the same. They differ significantly in configuration for both reading and mathematics. Significant differences in level found for mathematics at all percentile ranks and for reading at the 10th percentile indicate there is greater spread from the 50th to the 10th percentile for the MAT than for the CAT. Differences in both the rate and proportion of total growth expected at certain grade levels imply that either the scales are not both equal interval or the tests are measuring different skills. Either makes the practices of aggregating data and comparing programs across tests invalid. In addition, if the scale scores for a test are not equal interval the procedure for converting out-of-level test scores to in-level NCE's is also invalid for that test. It is beyond the scope of this study to determine whether the scale scores for these two tests are actually equal interval. Possibly the application of Item Response Theory methods to data from these tests would provide the answer.

Robustness of the Norm-referenced Model with Respect to
Variations Found in Local Patterns of Growth

When the growth curves implied by the norm-referenced model were examined using local data the most apparent difference found between the local and national growth curves was in the levels of the corresponding percentile ranks. The local growth curves in most cases appear to match the national curves but for a higher percentile rank. To the extent that they match the national curves in shape and are only displaced vertically the estimates of expected posttest scores and hence the gains will still be unbiased when national norms are used. Since the analysis of expected no-treatment gains found no significant differences due to percentile rank the hypothesis that the percentile ranks are parallel for the grade levels examined cannot be rejected. This indicates the norm-referenced model is robust with respect to differences in percentile rank.

No systematic bias was found for any of the district types. The variables identified here do not appear to have produced common effects across tests on the achievement of students in the selected districts. Use of out-of-level testing with low achieving students cannot be rejected as inappropriate for use with this model. The effects of an improving total school program are also not a sufficient basis for rejecting use of national norms in

evaluating student gains. Even a high proportion of low SES students does not result in differing patterns of growth at all grade levels. This indicates the model is robust with respect to the variables identified.

Where significant differences were found they were specific to particular grade levels. The test by grade level interactions found in the local data suggest that the differences between the tests are specific to particular test levels. In reading the local norms for the MAT are significantly different from the national norms at Grade 6 only. The local norms are higher than the MAT national norms which underestimate student achievement.

The finding of significant bias in reading for the MAT at Grade 6 indicates that for these Michigan districts use of the MAT Grade 6 norms does not provide a valid expectation of no-treatment reading achievement. Specifically the MAT norms underestimate the reading achievement of their Grade 6 students by more than 5 NCE's. Use of the norm-referenced model in this situation would overestimate any program effect.

In mathematics the significant differences found were specific to given districts and to one or two grades in those districts. MLS Grade 6 local norms were higher than the MAT national norms. Local norms for CIM Grade 3 and MOL Grades 3 and 4 were lower than the national norms for their respective tests. These differences would result in apparent gains of 9 NCE's for MLS at Grade 6, of

-5 NCE's for CIM and MOL at Grade 3 and of -9 NCE's for MOL at Grade 4 when the norm-referenced model with national norms is used to evaluate student achievement that by local norms would show a zero NCE gain. The district by grade level interaction suggests that any significant differences are the result of local district variables. It can be hypothesized that there are local conditions or practices not identified in this study that do affect the validity of the model when used for local evaluation. One possibility may be differences in the degree to which the local curriculum matches the test content at each grade level.

All of the districts in this study that used the CAT are larger than any of the districts using the MAT. Therefore, any effects of district size are confounded with the effects of the test used. It is impossible to tell from the data obtained what if any part size played in these findings.

The only instances where Type I errors are significantly more likely to occur using the model with national norms are those where the no-treatment expectation yields positive gain scores. From the data reported here this occurs in Grade 6 reading evaluated with the MAT and in Grade 6 math in district MLS.

Any bias in the no-treatment expectation will affect the power of the model to detect true gains that do occur. Local districts with 30 or fewer Chapter 1 students at a

grade level were found unlikely to detect small gains of 1 to 4 NCE's. These gains are of the magnitude usually found in statewide and national aggregations of achievement data. Gains of 7 NCE's which would generally be considered to be educationally significant would be detected even with small groups. An exception would be in instances of significant negative no-treatment expected gains, as in the cases of Grade 3 math in districts MOL and CIM and Grade 4 math in district MOL. In these cases even gains as large as 7 NCE's are unlikely to be detected.

Stability of Local Growth Patterns from Year to Year

Districts that used the CAT showed less year-to-year variation and, therefore, more stability in their growth patterns than did those that used the MAT. Stability of the yearly norms is one effect that might be hypothesized to be related to district size. In this study it is impossible to separate the effects of district size on stability from the effects of the test used. If the difference in stability of the yearly curves for the two tests is the result of differences in size the combining of three years' data should provide composite growth curves for the MAT districts comparable in reliability to the yearly curves of the smaller CAT districts. If a year-to-year variation equivalent to less than 7 NCE's is

considered acceptable then the data for the CAT districts is within acceptable limits.

Adequacy of Cross-sectional Norms for Estimating
Longitudinal Patterns of Growth

The examination of the longitudinal data provides mixed signals regarding the use of cross-sectional norms to evaluate programs where observed achievement is measured longitudinally. The MOL mathematics data support the model. So does the MOL reading data overall though it indicates that, for local projects at least, allowance must be made for cohort differences. The COL reading data supports use of the model for a single year's evaluation but indicates it is invalid for evaluations over periods of two or more years. On the basis of the COL longitudinal math data, however, the model would have to be rejected. The positive gains found in the COL math data represent gains significantly above what would have been expected on the basis of either the national or the local district norms. Though Chapter 1 students were excluded from the longitudinal sample it is possible that local district programs or staff provided extra help to students in non-Chapter 1 buildings. In any case it is clear that data from the longitudinal sample do not necessarily reflect either national or district norms.

The students in the longitudinal sample were limited to those tested annually in the same school district over a three-year period. It may be that these students in

districts such as COL improve relative to the total district population over time. It may even be that the entire population is gradually improving over time. This study cannot determine whether either of these possibilities holds. The higher local norms found in several of the districts suggests that for some districts the entire population may be improving.

Implications of These Findings

In summary these outcomes indicate that the norm-referenced model is robust with respect to percentile rank. It is also robust with respect to the district variables identified in this study in reading. Due to the small sample size with only one district per cell it is impossible from the data collected to separate the test by district type interaction found in mathematics from the effects of other district variables. The hypothesis that the differences found in mathematics are due to district variables other than those identified in this study cannot be rejected.

The growth curves developed from the national norms for the two tests indicate that there are significant differences between them. They imply that either the tests are measuring different skills at certain grade levels or the scale for at least one of the tests is not equal interval. Overall it appears that the MAT is more likely to detect gains at Grade 6 than is the CAT. In

mathematics it was impossible to determine whether the significant differences found are the result of differences between the national norms at certain levels of the two tests or of local variables, such as differences in the curriculum, which affected achievement relative to the national norms in an individual district. In any case further research on the impact of the differences between tests on data aggregated across tests at the state and national level is needed. National norms may be appropriate for evaluating nationally aggregated data for a given test but not be a valid basis for aggregating data across tests.

The effect of district size also needs to be studied further. In this study district size and test used were confounded so that it was not possible to separate their effects in the data. The relationship between number of students and stability of local norms is especially important. Although possible effects of test used cannot be completely discounted it appears from the data that districts with fewer than one hundred students per grade would be well advised to pool more than one year's data if they want to use local norms.

The longitudinal data raise serious questions about the validity of using cross-sectional norms with longitudinal data. The sample used differed from the district population in two important ways. Students were included in the longitudinal sample only if they 1) were tested in

the same district on three separate dates over a span of two years and 2) had not received Chapter 1 services. Scores used in the cross-sectional norms include both Chapter 1 as well as non-Chapter 1 students who were tested on a single test date. It might be expected that merely limiting the sample to students who remain in the same school district for one or two years would result in gains relative to a national norming population that includes students who move more frequently.

Areas Needing Additional Research

This study examined only two of the dozen or so standardized achievement tests in wide use by school districts. Other standardized tests need to be examined to determine what differences exist in the patterns of growth which the norm-referenced model implies for them. If gain scores are to be aggregated across tests, there is a need to determine whether there are differences in content and scale which will bias the results. The question of the validity of aggregating achievement data across tests needs to be further investigated.

Data from a sample of only six districts and longitudinal data from only two districts were analyzed in this study. Still, it identified an important number of instances in which the model doesn't work. Until the variables which are the source of the bias observed are identified the effort needed to determine whether or not the

model will work in the case of a given local program is more than would be involved in implementing a better method of evaluation in the first place.

Other district variables than the ones identified here which may affect the validity of the model in evaluating mathematics programs need to be identified. Local curriculum variables and the match between the test used and the curriculum taught probably are more crucial in mathematics than in reading. Reading skills are more independent of the curriculum and are not all learned in school.

The effect of district size needs to be more thoroughly studied. What effect does it have on the impact of local variables such as those identified in this study? In particular to what extent are the differences observed between the tests in this study a result of differences in the size of the districts using each?

The question of the validity of cross-sectional norms for measuring longitudinal data is particularly important. The results may reflect differences in achievement of students who remain in the same district for at least two years as opposed to the achievement of all students present for a single test administration. If this is the case they indicate an important difference between cross-sectional and longitudinal data as measures of expected achievement. Another possible hypothesis is that they reflect differences in the achievement of non-Chapter 1

students, particularly those attending schools not eligible for Chapter 1, as opposed to the average achievement of all students in a district. If this is the case they would not affect the validity of using cross-sectional norms to evaluate the achievement of Chapter 1 students. It is also possible that true local effects on student achievement may become apparent only when longitudinal data collected over a number of years is analyzed.

CHAPTER VII

SUMMARY AND CONCLUSIONS

Although there have been numerous studies of the norm-referenced model there is still disagreement about its validity. Due to its wide usage there has remained a need for additional research to provide a clearer understanding of the model. This study examined the validity of the model in the following ways.

1. It approached the norm-referenced model as implying a model of growth in measured achievement over successive school years.
2. It compared patterns of growth implied by the norm-referenced model for two widely used standardized tests, the CAT and the MAT.
3. It examined patterns of growth in six school districts selected as exemplifying factors thought likely to affect the validity of the norm-referenced model. Local patterns of growth were compared with those implied by the national norms for the test used. District growth patterns were analyzed and used to estimate gains that would be observed under no-treatment conditions.
4. It evaluated the statistical effects of using local rather than national norms in testing the effectiveness of local programs. Likelihood of Type I errors and effects on power to detect true gains were

analyzed.

5. It compared the results obtained from longitudinal data with those expected on the basis of cross-sectional norms.

Conclusions

The patterns of growth implied by the norm-referenced model and the norms of the standardized tests studied are curvilinear with the highest rate of growth in the early grades and a decreasing rate in the higher grades. There are significant differences in the growth curves implied by the norms of the CAT and MAT. These affect the results obtained with the model differently at different grade levels.

The model is robust with respect to differences in percentile rank. In reading it is also robust with respect to the district variables identified in this study. The mathematics results are more complex but it appears probable that the differences observed were due to local variables other than those identified in the design of the study.

Results of the longitudinal analyses are inconclusive. Data from one district allow the hypothesis that the cross-sectional norms provide a valid expectation for longitudinal data to be rejected. The second district's data do not. Clearly in some cases local non-Chapter 1 data may not match the results of cross-sectional data for

the national or local district populations.

From the analysis of the statistical effects of local patterns of growth it is apparent that they affect the probability of Type I errors and the power of the model to detect true gains. When the local expectation is significantly higher than that based on the national norms the probability of Type I errors is significantly increased. Small true gains are unlikely to be detected at the local level unless the local rate of growth is greater than in the national norms. Educationally significant gains of 7 NCE's or more will usually be detected at the local level except in cases where the local rate of growth is significantly less than that in the national norms.

The norm-referenced model evaluates students' total educational experience including both their regular school program and any special programs such as Chapter 1. While this study involved only a small sample of districts it does identify several areas that should definitely be considered when interpreting local data and that may cause bias at the national level. It is, however, possible that the norm-referenced model yields valid aggregate estimates at the national level even though the estimates of effectiveness for specific programs are biased. Much work is still needed to identify the significant variables that affect the growth of local populations of students. Further research is needed to develop local models of growth

in reading and mathematics achievement as measured by
standardized tests.

APPENDIX A

GRAPHS OF LOCAL GROWTH CURVES

APPENDIX A

GRAPHS OF LOCAL GROWTH CURVES

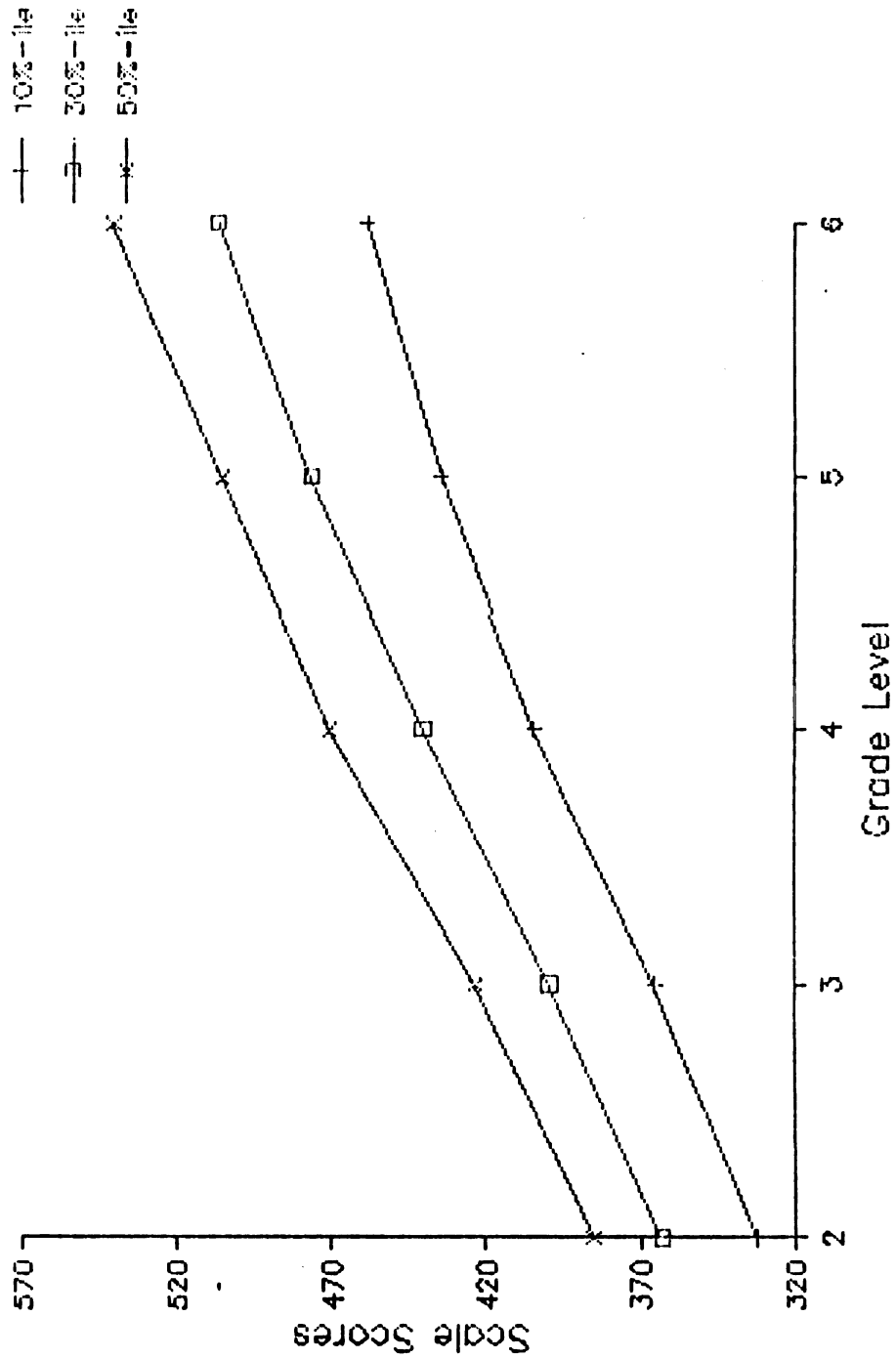


Figure 18. COL Growth Curves: Reading

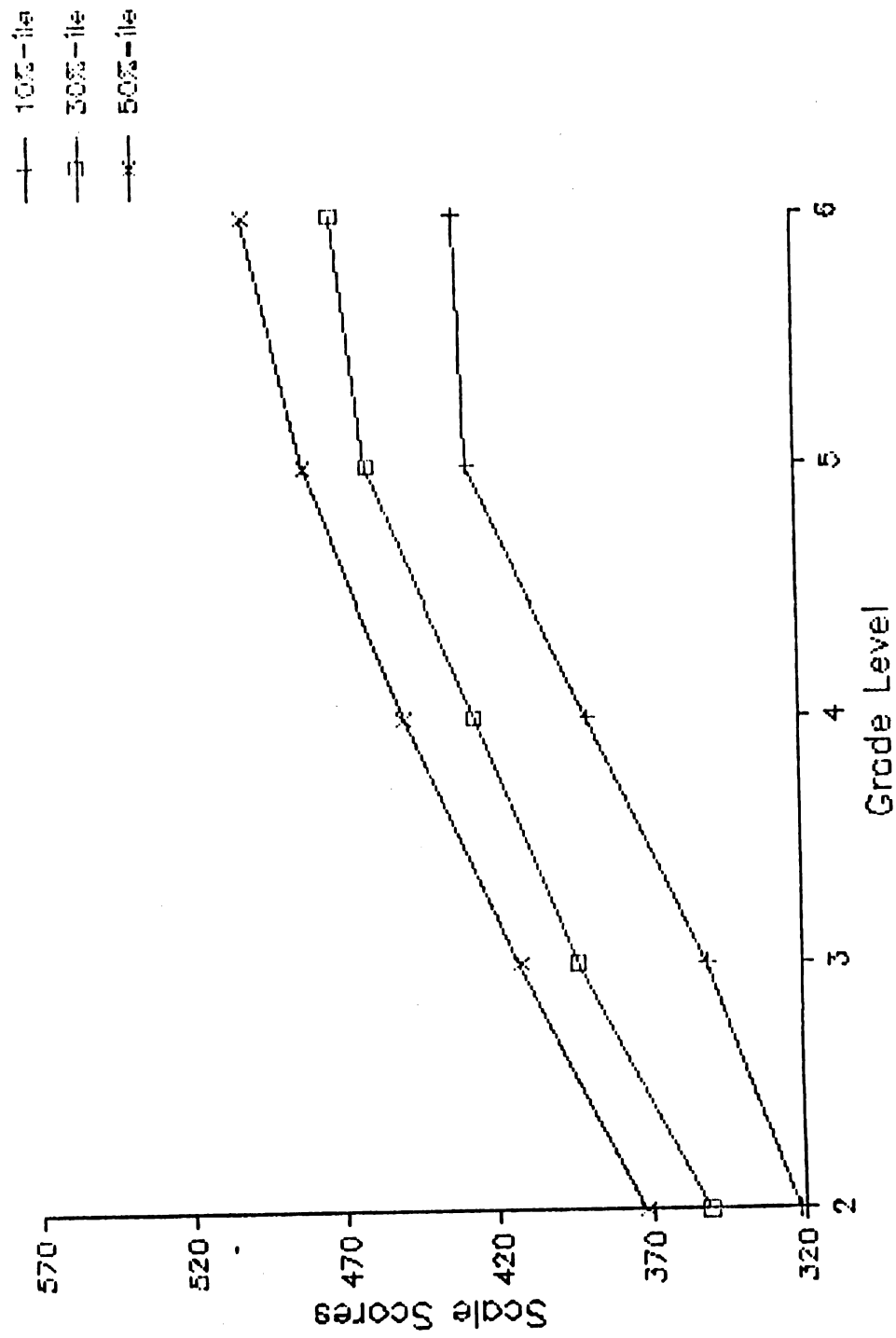


Figure 19. CIM Growth Curves: Reading

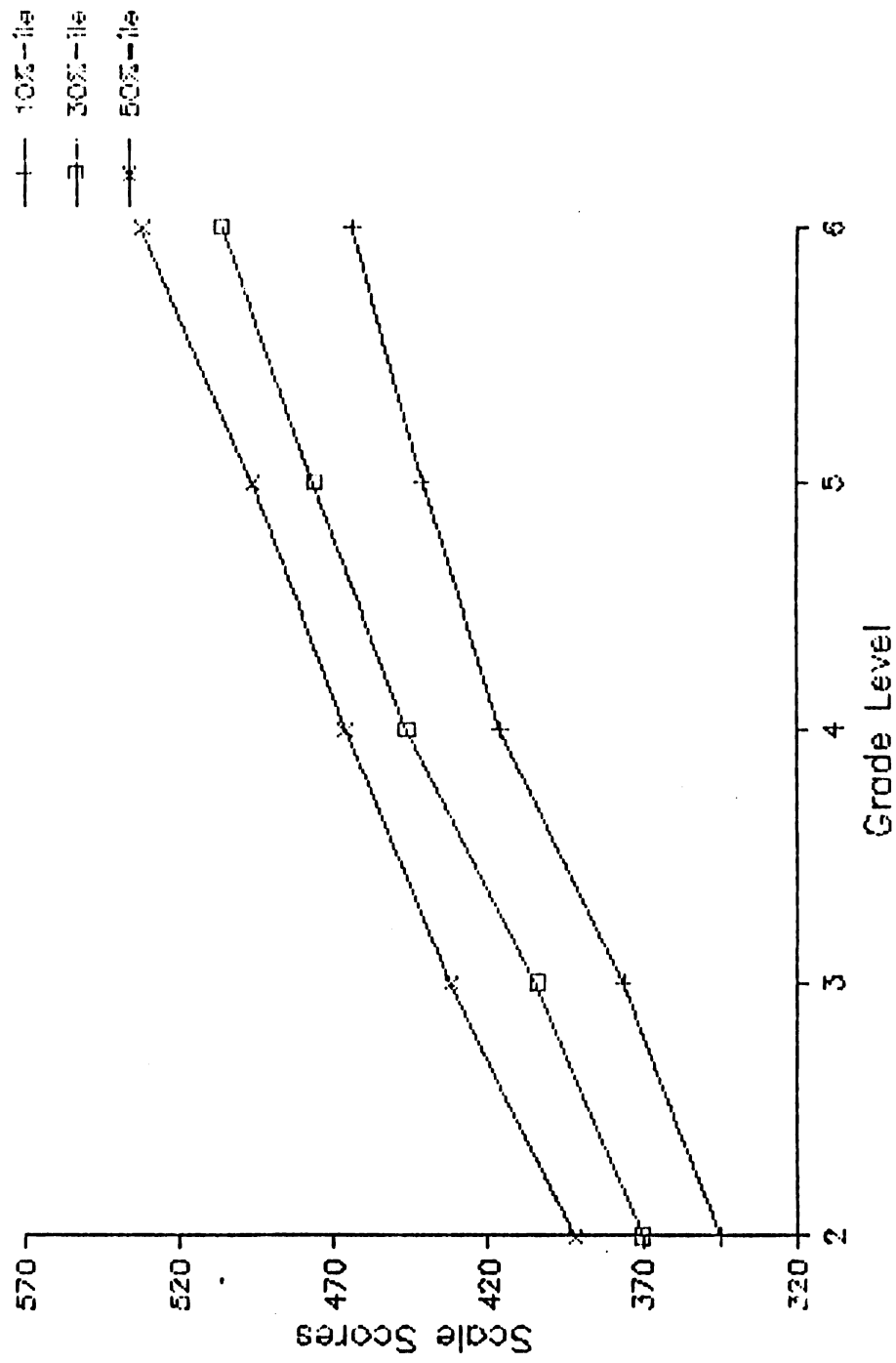


Figure 20. CLS Growth Curves: Reading

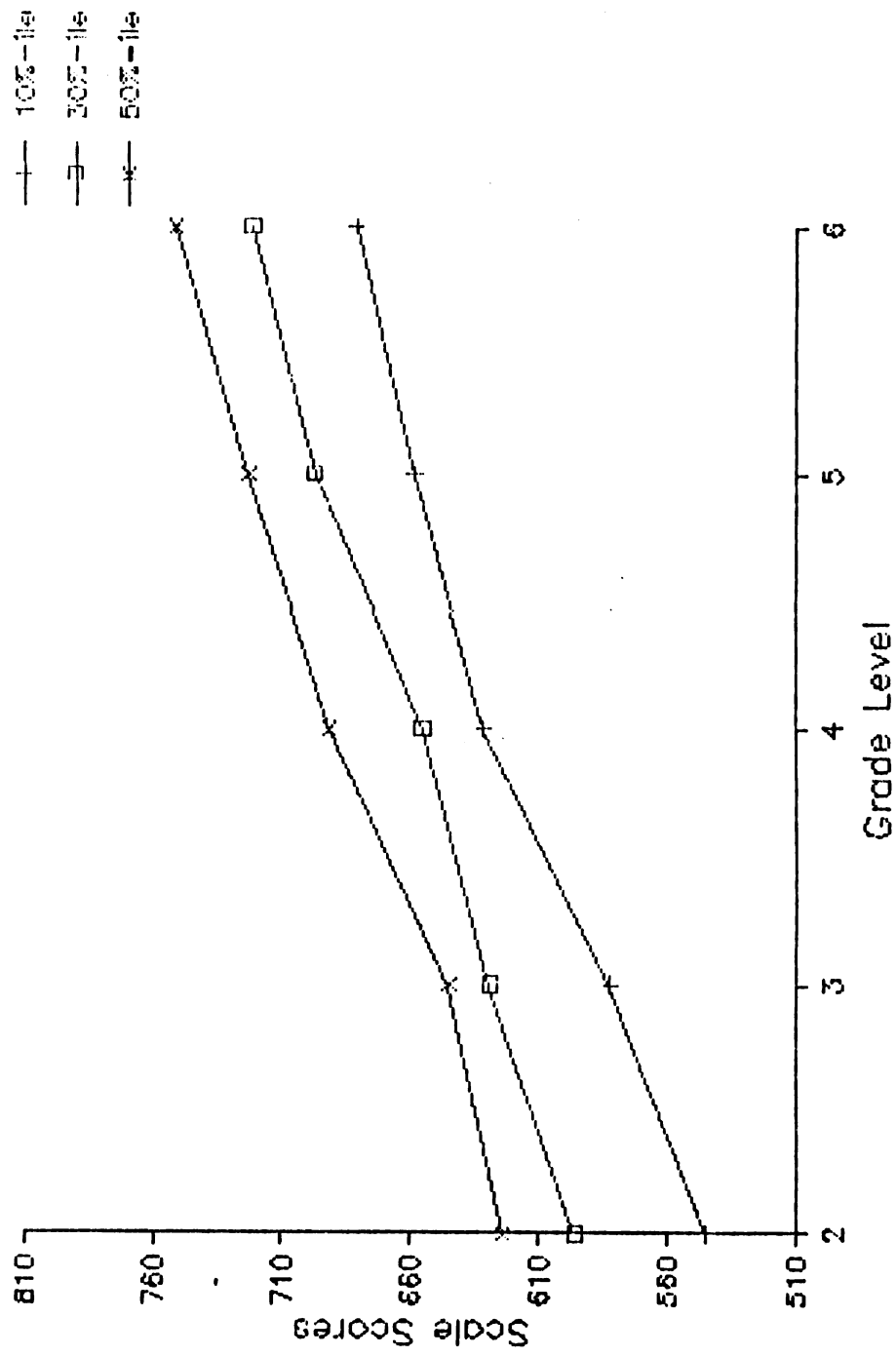


Figure 21. MOL Growth Curves: Reading

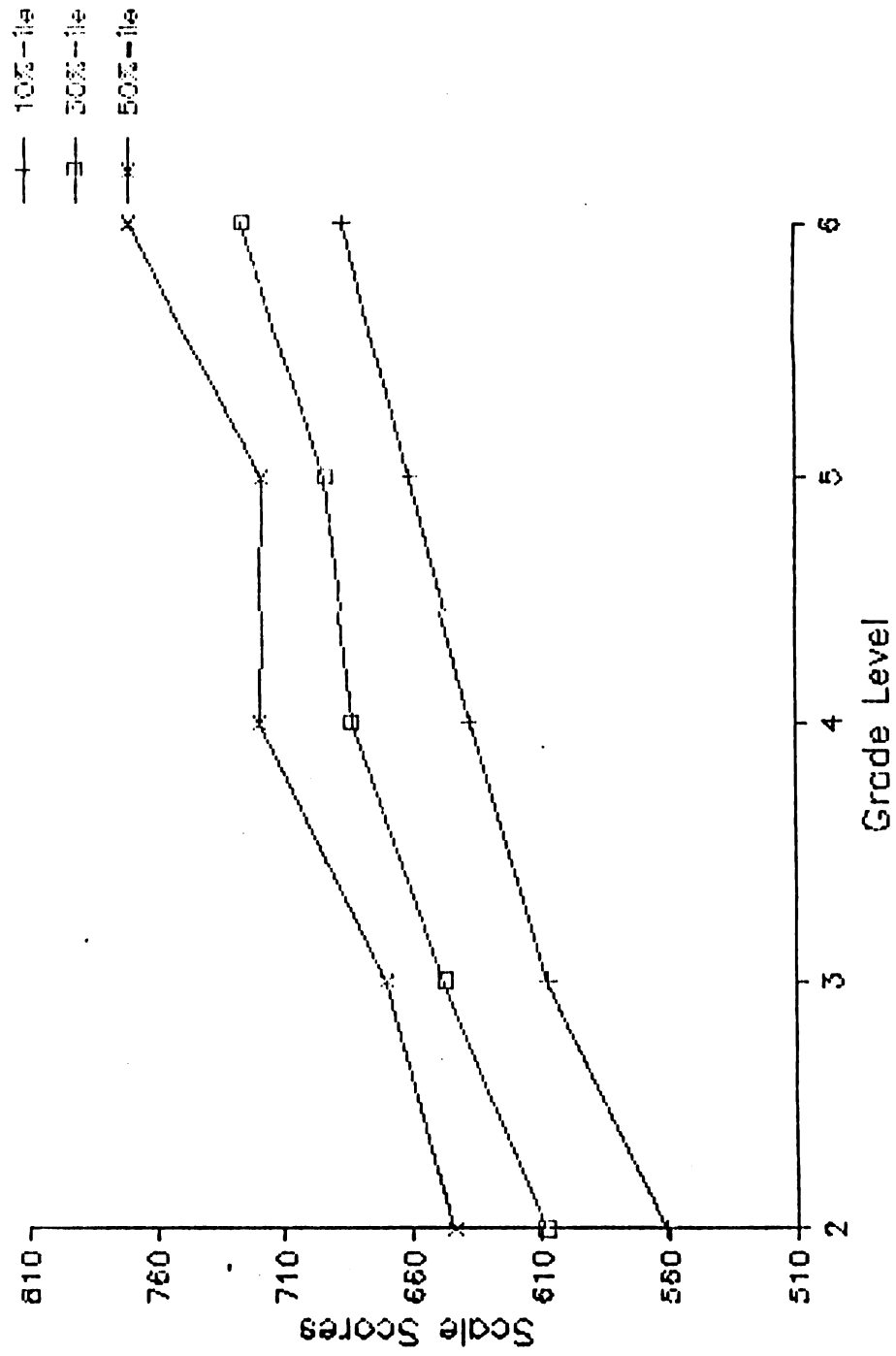


Figure 22. MIM Growth Curves: Reading

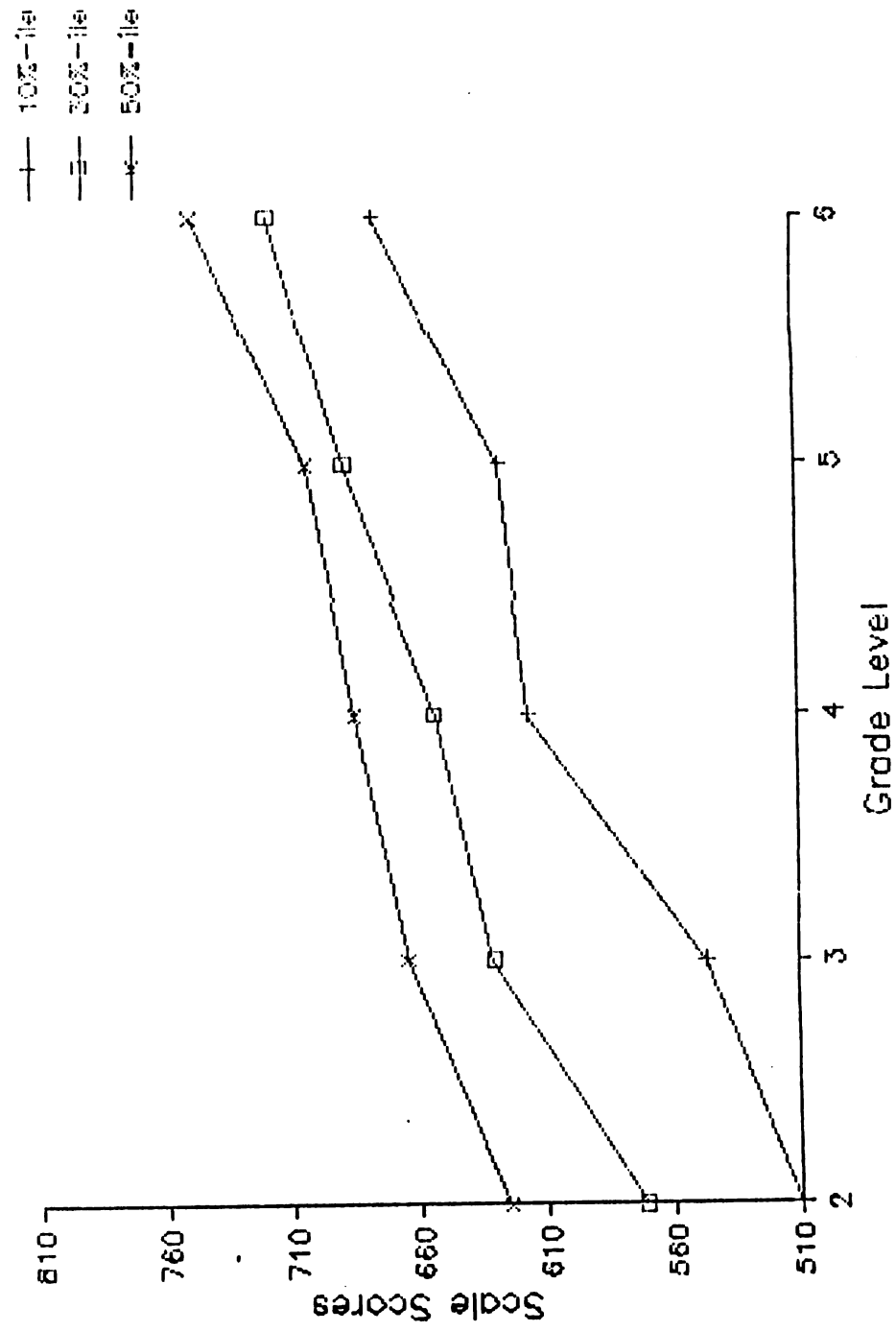


Figure 23. MLS Growth Curves: Reading

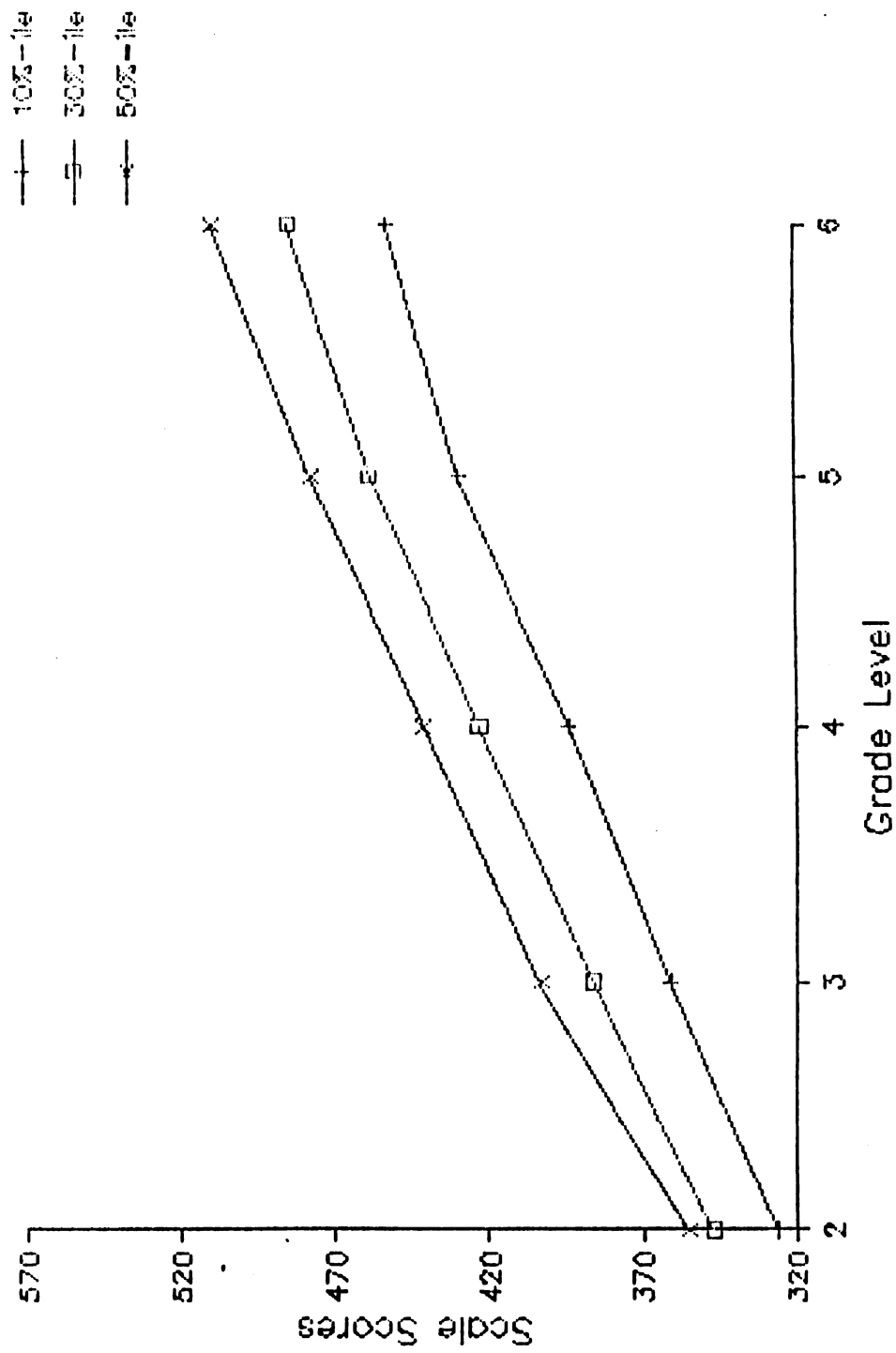


Figure 24. COL Growth Curves: Mathematics

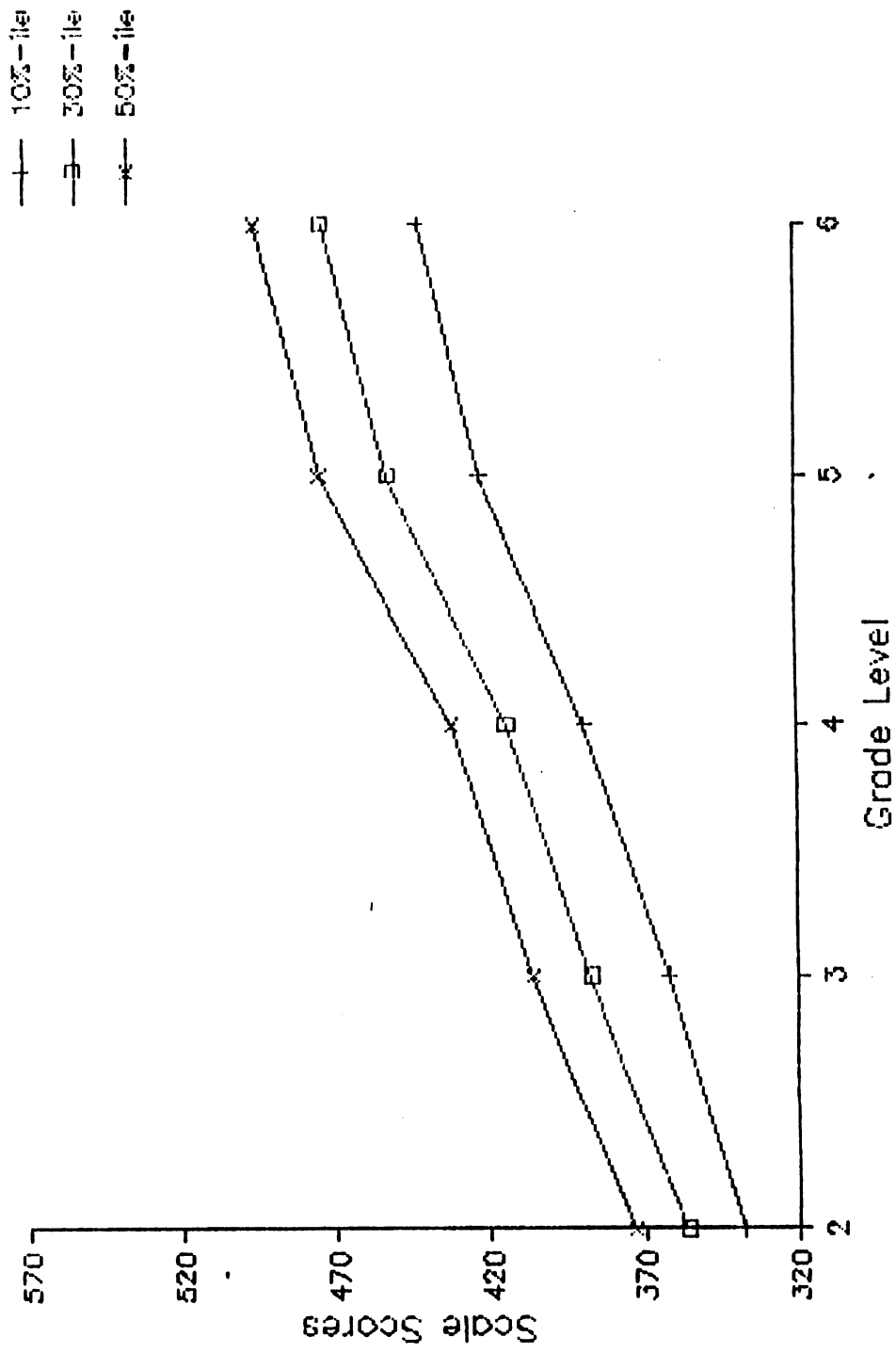


Figure 25. CIM Growth Curves: Mathematics

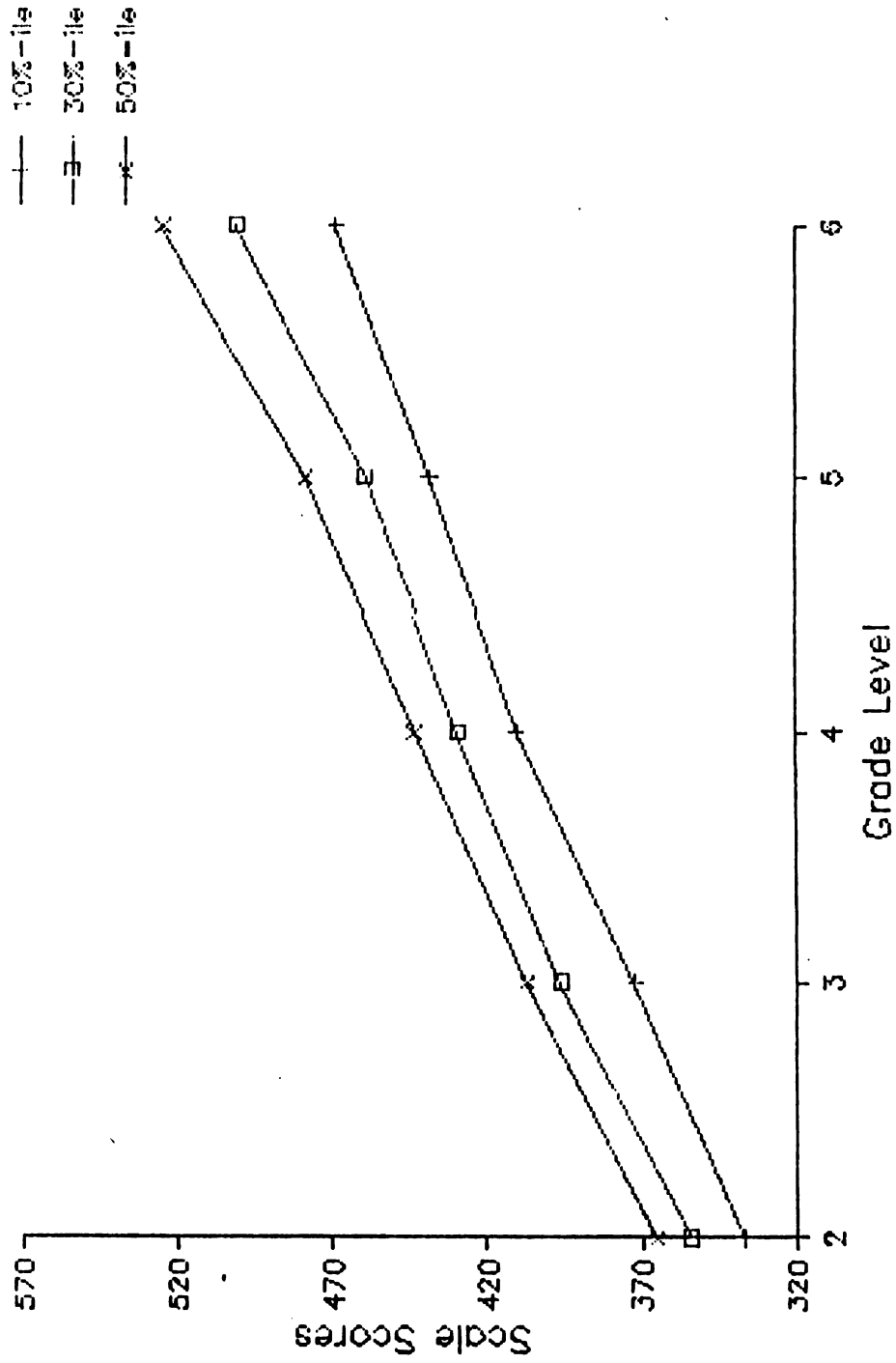


Figure 26. CLS Growth Curves: Mathematics

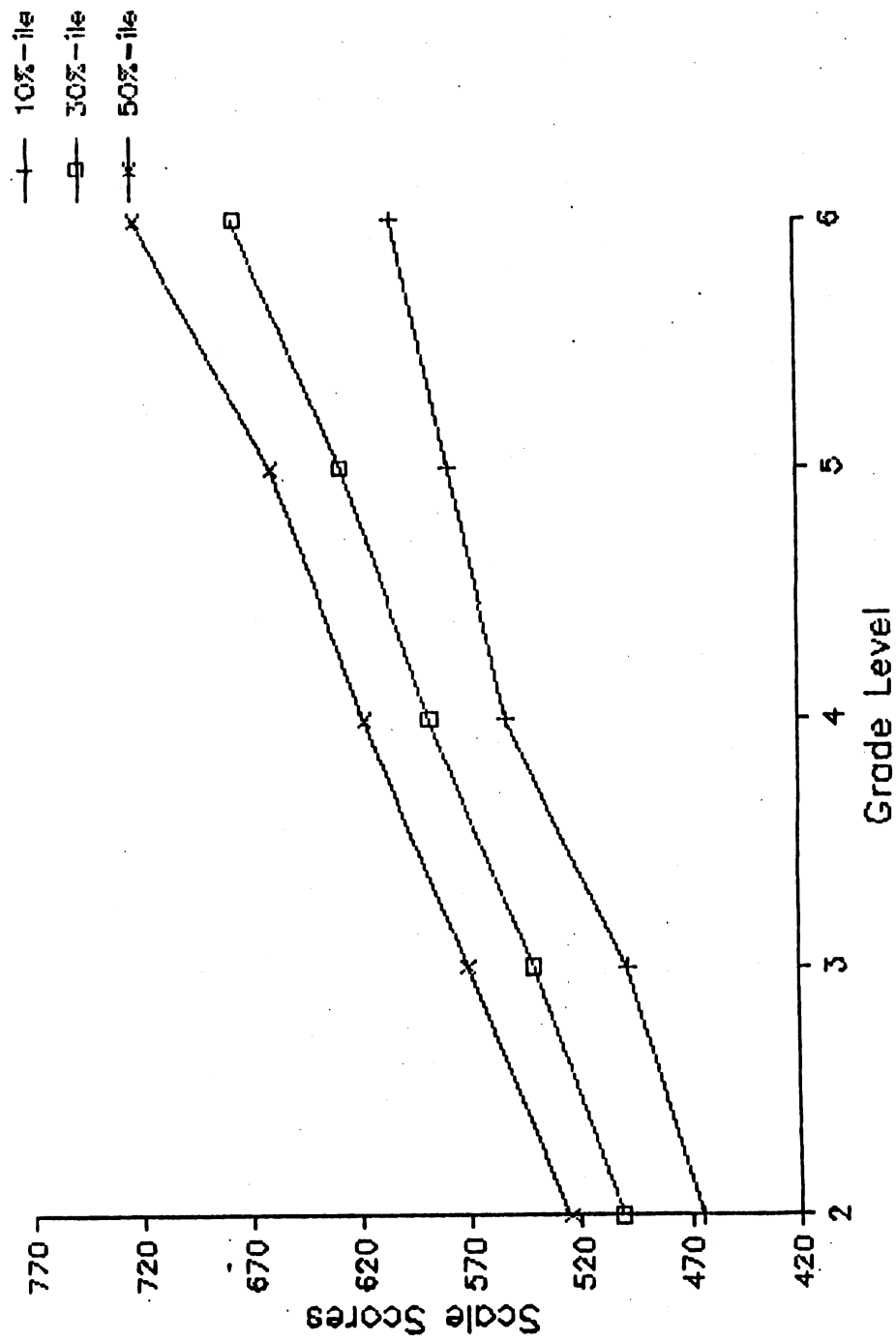


Figure 27. MOL Growth Curves: Mathematics

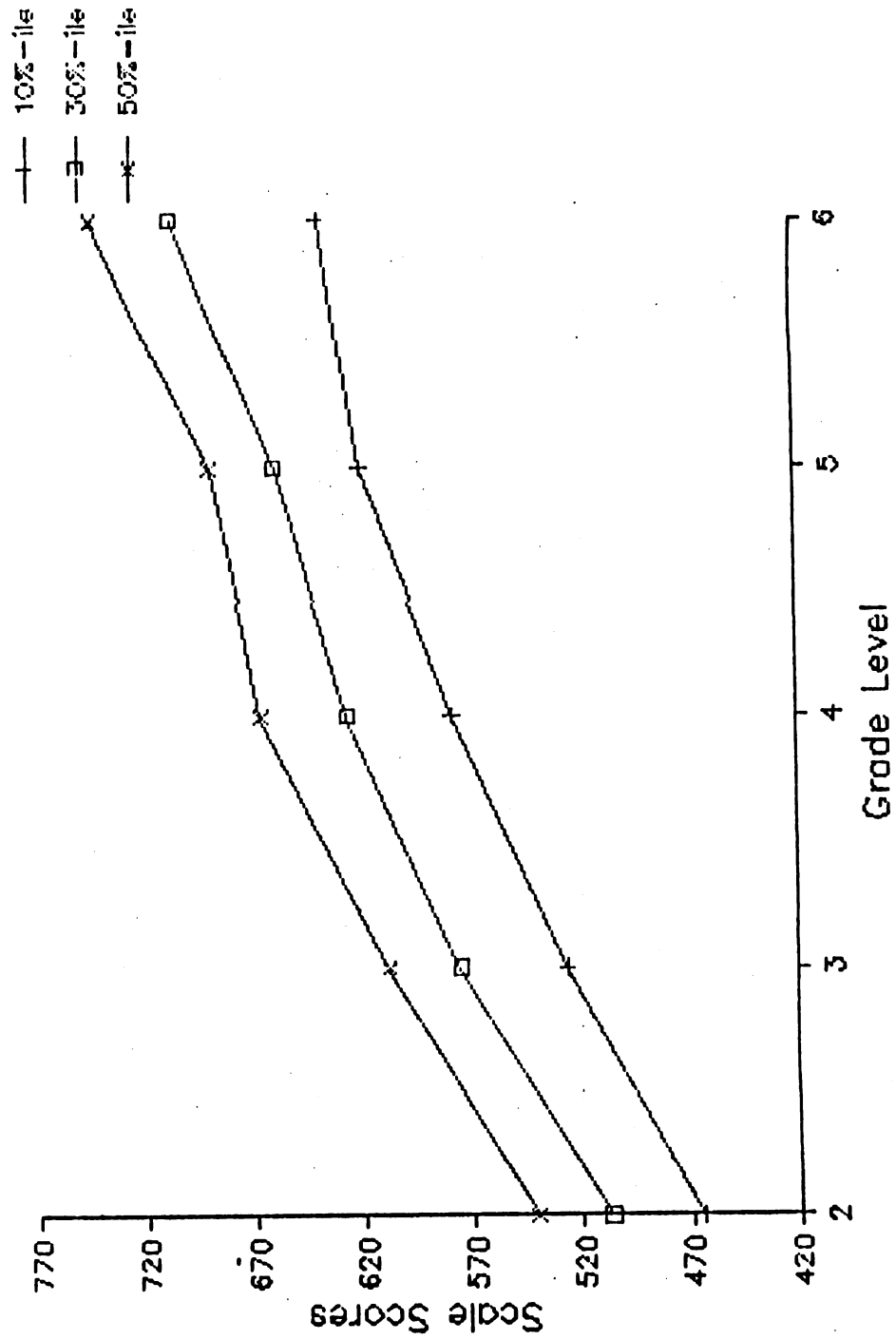


Figure 28. MIM Growth Curves: Mathematics

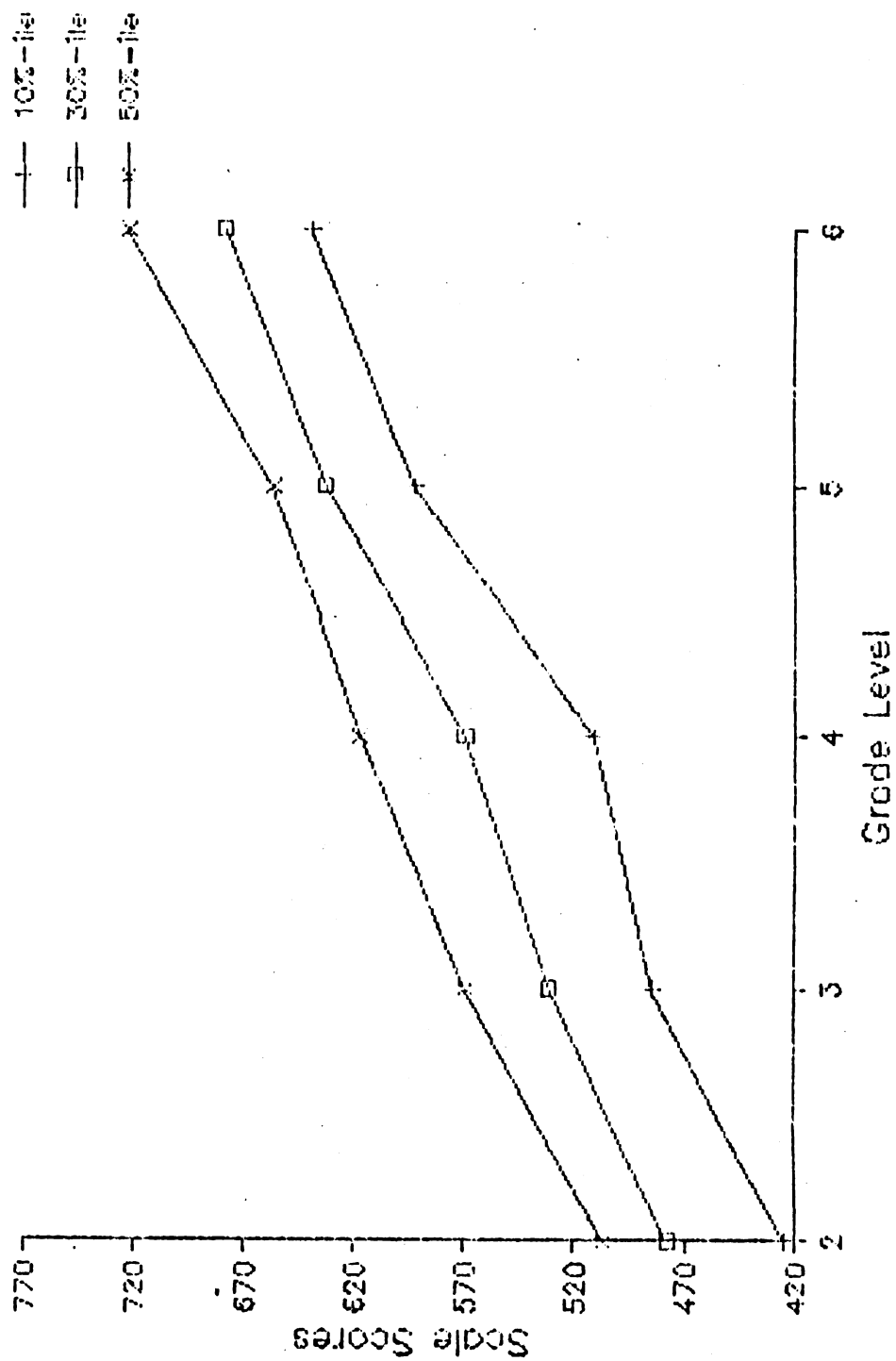


Figure 29. MLS Growth Curves: Mathematics

APPENDIX B

YEARLY DISTRICT DATA

APPENDIX B

YEARLY DISTRICT DATA

Table 15. Yearly District Data: CAT - Reading

Test/ District	Norms/ Year	Grade Level				
		2	3	4	5	6
CAT - 10% National		290	324	357	385	405
COL	Composite	332	365	404	434	458
	1983	326	364	400	424	458
	1984	338	365	404	434	460
	1985	330	365	404	435	458
CIM	Composite	322	351	389	427	431
	1983	312	346	378	425	427
	1984	322	355	393	422	438
	1985	326	362	393	427	438
CLS	Composite	345	376	416	441	464
	1983	349	370	419	441	455
	1984	342	383	410	456	460
	1985	348	370	419	440	482
CAT - 30% National		332	370	406	438	460
COL	Composite	363	399	440	476	506
	1983	358	394	435	472	502
	1984	365	402	444	480	510
	1985	363	402	444	479	506
CIM	Composite	351	393	426	460	471
	1983	348	380	424	454	464
	1984	351	395	430	463	471
	1985	354	395	430	460	481
CLS	Composite	370	404	446	476	506
	1983	373	402	448	476	498
	1984	373	408	440	482	506
	1985	366	408	450	464	510
CAT - 50% National		360	401	443	473	500
COL	Composite	385	423	470	505	540
	1983	383	423	460	496	534
	1984	388	425	472	507	548
	1985	385	425	470	507	544
CIM	Composite	372	412	449	481	500
	1983	369	403	449	476	493
	1984	375	417	449	484	493
	1985	372	423	451	484	508
CLS	Composite	392	432	466	496	532
	1983	392	425	469	496	528
	1984	392	432	460	504	535
	1985	387	432	469	493	532

Table 16. Yearly District Data: MAT - Reading

Test/ District	Norms/ Year	Grade Level				
		2	3	4	5	6
MAT - 10% National		524	566	602	628	632
MOL	Composite	545	582	631	658	681
	1983	584	621	631	654	
	1984	516	566	625	673	
	1985	560	558	631	628	681
MIM	Composite	561	607	637	660	686
	1983	568	608	626	649	701
	1984	546	625	646	649	690
	1985	561	597	639	668	669
MLS	Composite	510	546	616	627	675
	1983	489	521	640	613	678
	1984	488	603	456	637	680
	1985	534	540	618	600	675
MAT - 30% National		576	621	655	682	693
MOL	Composite	596	629	655	697	721
	1983	620	643	651	704	
	1984	581	630	652	711	
	1985	596	598	665	686	721
MIM	Composite	608	647	683	693	725
	1983	616	649	669	691	753
	1984	604	657	691	690	725
	1985	612	636	689	701	708
MLS	Composite	571	630	653	688	717
	1983	552	615	678	688	724
	1984	560	657	650	690	712
	1985	603	606	647	688	714
MAT - 50% National		620	661	695	718	737
MOL	Composite	624	645	691	723	751
	1983	646	676	709	732	
	1984	608	643	671	736	
	1985	620	629	703	714	751
MIM	Composite	644	670	719	718	769
	1983	646	672	711	718	786
	1984	630	685	727	716	755
	1985	652	653	727	721	749
MLS	Composite	625	664	685	703	747
	1983	615	658	698	709	753
	1984	617	680	660	706	744
	1985	625	642	695	703	753

Table 17. Yearly District Data: CAT - Mathematics

Test/ District	Norms/ Year	Grade Level				
		2	3	4	5	6
CAT - 10% National		309	343	372	395	418
COL	Composite	326	361	394	429	452
	1983	326	359	392	423	450
	1984	328	362	398	430	452
	1985	326	362	394	432	454
CIM	Composite	338	362	389	422	442
	1983	333	356	388	417	440
	1984	342	368	388	422	446
	1985	343	366	392	428	444
CLS	Composite	337	372	410	438	468
	1983	337	371	410	425	463
	1984	342	379	408	447	462
	1985	337	371	414	436	481
CAT - 30% National		336	374	406	436	462
COL	Composite	347	386	423	458	484
	1983	343	384	420	454	480
	1984	348	386	424	458	484
	1985	350	386	420	459	486
CIM	Composite	356	387	414	452	473
	1983	353	384	409	451	467
	1984	364	386	421	448	473
	1985	364	391	416	457	479
CLS	Composite	354	396	429	459	500
	1983	351	392	431	455	494
	1984	360	399	422	468	497
	1985	350	392	433	453	509
CAT - 50% National		352	394	428	463	491
COL	Composite	355	403	441	477	509
	1983	353	403	440	475	503
	1984	357	401	442	480	509
	1985	355	403	441	477	509
CIM	Composite	373	406	432	474	495
	1983	367	402	432	474	490
	1984	376	404	439	469	495
	1985	376	409	431	474	500
CLS	Composite	365	407	443	478	524
	1983	364	402	445	475	522
	1984	372	409	441	480	519
	1985	360	407	445	478	533

Table 18. Yearly District Data: MAT - Mathematics

Test/ District	Norms/ Year	Grade Level				
		2	3	4	5	6
MAT - 10% National		403	454	501	546	564
MOL	Composite	464	498	553	579	604
	1983	496	497	550	586	
	1984	436	509	555	604	
	1985	472	497	537	569	604
MIM	Composite	464	526	578	619	637
	1983	460	532	563	609	676
	1984	459	526	588	609	631
	1985	464	526	578	619	637
MLS	Composite	425	484	511	591	639
	1983	418	487	554	609	631
	1984	399	469	486	542	646
	1985	435	486	524	571	639
MAT - 30% National		468	521	578	624	646
MOL	Composite	501	541	588	628	676
	1983	517	541	583	636	
	1984	483	552	592	636	
	1985	507	531	574	624	676
MIM	Composite	507	575	626	658	704
	1983	492	588	618	656	723
	1984	516	588	626	640	698
	1985	533	546	646	688	682
MLS	Composite	478	531	569	632	678
	1983	486	524	600	632	660
	1984	460	562	539	635	696
	1985	493	515	569	632	696
MAT - 50% National		507	567	622	664	696
MOL	Composite	525	571	618	660	721
	1983	545	593	626	656	
	1984	517	588	618	683	
	1985	541	551	618	660	721
MIM	Composite	541	608	666	688	741
	1983	517	628	651	683	754
	1984	541	608	666	668	741
	1985	558	573	681	730	717
MLS	Composite	507	569	617	656	722
	1983	519	569	643	660	695
	1984	486	584	591	653	731
	1985	515	558	608	667	741

REFERENCES

REFERENCES

- Bryk, R. S., Strenio, J. R., & Weisberg, H. I. A method for estimating treatment effects when individuals are growing. Journal of Educational Statistics, 1980, 5, 5-4.
- Bryk, A. S., & Weisberg, H. I. Value-added analysis: A dynamic approach to the estimation of treatment effects. Journal of Educational Statistics, 1976, 1, 127-155.
- California Achievement Tests, Norms Tables. Monterey, CA: CTB/McGraw-Hill, 1978.
- David, J. L., & Pelavin, S. H. Evaluating compensatory education: Over what period of time should achievement be measured? Journal of Educational Measurement, 1978, 15, 91-99.
- Hiscox, S. B., & Owen, T. R. Behind the basic assumption of Model A. Paper presented at the annual meeting of the American Educational Research Association. Toronto: 1978.
- Horst, D. P., & Tallmadge, G. K. A Procedural Guide for Validating Achievement Gains in Educational Projects, Monograph Series on Evaluation in Education, No. 2. Washington, D. C.: Office of Education, 1976.
- Langer, P., Kalk, J. M., & Searles, D. T. Age of achievement and trends in achievement: A comparison of blacks and caucasians. American Educational Research Journal, 1984, 21, 61-78.
- Linn, R. L. Validity of inferences based on the proposed Title I evaluation models. Educational Evaluation and Policy Analysis, 1979, 1, 2, 23-32.
- Linn, R. L. Discussion: Regression toward the mean and the interval between test administrations. New Directions for Testing and Measurement, 1980, 8, 83-89.
- Linn, R. L. Measuring pretest-posttest performance changes. In R. A. Berk (ed.) Educational Evaluation Methodology: The State of the Art. Baltimore: Johns Hopkins University Press, 1981.
- Lord, F. M. Significance test for the hypothesis that two variables measure the same trait except for errors of measurement. Psychometrika, 1957, 22, 207-220.

Murray, S. L., Arter, J., & Faddis, B. Title I technical issues as threats to internal validity of experimental and quasi-experimental designs. Paper presented at the annual meeting of the American Educational Research Association. San Francisco: 1979.

Olejnik, S. F. Data analysis strategies for quasi-experimental studies where differential group and individual growth rates are assumed. Unpublished dissertation. East Lansing: Michigan State University, 1977.

Porter, A. C., Schmidt, W., Floden, R., & Freeman, D. J. Practical significance in program evaluation. American Educational Research Journal, 1978, 15, 529-539.

Powers, S., Slaughter, H., & Helmick, C. A Test of the Equipercentile Hypothesis of the TIERS Norm-referenced Model. Tucson, Arizona: Tucson Unified School District, 1982.

Powers, S., Slaughter, H. & Helmick, C. A test of the equipercentile hypothesis of the TIERS norm-referenced model. Journal of Educational Measurement, 1983 20, 3, 299-302.

Prescott, G. A., Balow, I. H., Hogan, T. P. & Farr, R. C. Metropolitan Achievement Tests, Teacher's Manual for Administering and Interpreting. Cleveland, OH: The Psychological Corporation, 1978.

Reisner, E. R., Alkin, M. C., Boruch, R. F., Linn, R. L., & Millman, J. Assessment of the TITLE I Evaluation and Reporting System. Washington, D. C.: U.S. Department of Education, 1982.

Roberts, R. O. H. Regression toward the mean and the regression-effect bias. New Directions for Testing and Measurement, 1980, 8, 59-82.

Tallmadge, G. K. Cautions to evaluators. In Wargo, M. J., & Green, D. R. (Eds.) Achievement Testing of Disadvantaged and Minority Students. New York: CTB McGraw-Hill, 1977.

Tallmadge, G. K. An empirical assessment of norm-referenced evaluation methodology. Journal of Educational Measurement, 1982, 19, 2, 97-112.

Tallmadge, G. K. Rumors regarding the death of the equipercentile assumption may have been greatly exaggerated. Journal of Educational Measurement, 1985, 22, 1, 33-39.

Tallmadge, G. K., & Horst, D. P. A Procedural Guide for Validating Achievement Gains in Educational Projects. Washington, D. C.: U.S. Government Printing Office, 1976.

Tallmadge, G. K., & Wood, C. T. User's Guide: ESEA Title I Evaluation and Reporting System. Washington, D. C.: U.S. Office of Education, 1976.

Tallmadge, G. K., & Wood, C. T. User's Guide: ESEA Title I Evaluation and Reporting System (rev.). Washington, D. C.: U.S. Office of Education, 1978.

Thurstone, L. L. A method of scaling psychological and educational tests. Journal of Educational Psychology, 1925, 16, 7, 433-451.

Trochim, W. M. K. Methodologically based discrepancies in compensatory education evaluations. Evaluation Review, 1982, 6, 4, 443-480.

Van Hove, E., Coleman, J. S., Rabben, K., & Karweit, N. Schools' performance: New York, Los Angeles, Chicago, Philadelphia, Detroit, Baltimore. Unpublished manuscript, Baltimore, 1970.

MICHIGAN STATE UNIV. LIBRARIES



31293000796858