An Empirical Study of the Equipercentile Assumption as a No-Treatment Expectation in Title I Evaluations

Alan C. Nowakowski

Western Michigan University

Follow this and additional works at: https://scholarworks.wmich.edu/dissertations

Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation


https://scholarworks.wmich.edu/dissertations/2571

This Dissertation-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Dissertations by an authorized administrator of ScholarWorks at WMU. For more information, please contact maira.bundza@wmich.edu.
AN EMPIRICAL STUDY OF THE EQUIPERCENTILE ASSUMPTION AS A NO-TREATMENT EXPECTATION IN TITLE I EVALUATIONS

by

Alan C. Nowakowski

A Dissertation
Submitted to the
Faculty of the Graduate College
in partial fulfillment of the
requirements for the
Degree of Doctor of Education
Department of Educational Leadership

Western Michigan University
Kalamazoo, Michigan
August 1981
AN EMPIRICAL STUDY OF THE EQUIPERCENTILE ASSUMPTION AS A NO-TREATMENT EXPECTATION IN TITLE I EVALUATIONS

Alan C. Nowakowski, Ed.D.
Western Michigan University, 1981

The focus of this study was the tenability of the underlying assumption of Model A, the norm-referenced model, of the Title I Evaluation and Reporting System (TIERS). Model A estimates the no-treatment expectation from normative data. The assumption made by the Model is that, in the absence of special treatment, student groups maintain a constant percentile position in relation to national norms.

The study examined the extent to which student groups maintain a constant percentile position over a six month period. The study used a sizable representative student subsample of the standardization group for the 1978 Metropolitan Achievement Test (MAT). The study design included replications of data analyses for four different grade levels on two MAT subject area test scores. Class means were used as the unit of analysis. The study sample contained approximately 250 classes per grade level.

Constancy of class percentile rank was examined for grades 2, 3, 4, and 5 on MAT reading and math tests using the normal curve equivalent (NCE) score metric. Analyses were conducted using NCE scores calculated according to TIERS guidelines. A second supplemental series of analyses were conducted for NCE scores derived from the percentile distributions of MAT class mean scale scores of the sample.

Study results indicated classes generally do not maintain a constant percentile position in relation to national norms. Class means
were found to vary considerably from fall to spring in relation to national norms. The degree to which class means maintained their position fall to spring was more variable in math than in reading and changes in class mean position fall to spring were slightly larger in math than in reading. These two patterns held in each of four grade levels examined in this study. No distinct between-grade-level patterns were apparent.

For the four grades combined, the 95% confidence interval for class change scores in reading was estimated at -6.3 to +9.3 NCEs; the 95% confidence interval for class change scores in math was estimated at -9.8 to +13.8 NCEs. Note that zero growth is not at the center of the interval. The degree to which classes maintained their position in relation to the group score distributions of the sample in the supplemental analysis was even more variable than the degree to which classes maintained their position in relation to the MAT test norms for individuals.

The no-treatment expectation under Model A is based on the assumption that student groups maintain a constant percentile position. Based on the results of this study, there is no reason to assume student groups maintain a constant percentile position as a no-treatment expectation. Users of the Model, given the results of this study, should not view Model A change scores as a meaningful indicator of the quality of their educational program.
ACKNOWLEDGEMENTS

Data for this study were obtained from The Psychological Corporation. Policies and procedures of The Psychological Corporation for processing my outside data request were efficient, direct, and manageable. Gratitude is extended to Dr. Michael Beck for the manner in which he processed my request and his responsiveness in answering questions that arose as I worked with the data.

Data were processed and analyzed at the WMU Computer Center. I wish to thank the Director of the Center, Mr. Meagher, for the high quality of service I received from his staff. Special thanks go to Russ Barr, Computer Programmer, and Demitra Collia, Program Consultant, for the competent assistance and help they provided throughout the many data processing steps of this study.

Supervision of the study was provided by my dissertation Review Committee, Mary Anne Bunda (chairperson), Bradley Huitema, and James Sanders. I feel fortunate to have worked under the guidance of three such skilled and rigorous researchers. Assistance and diligent support they provided are greatly appreciated. I am especially grateful to Mary Anne for the patient, firm, and skillful way in which she guided this dissertation from initial drafting of the proposal to submission of final copy.

This study itself represents the last step in a lengthy doctoral graduate experience begun at the Evaluation Center at Ohio State University then directed by Daniel Stufflebeam, and completed at the Evaluation Center at Western Michigan University presently directed
by Daniel Stufflebeam. I thank him for his friendship, his advocacy and support, the many opportunities, experiences and responsibilities he provided me, and for his example.

Finally, I wish to share the fulfillment of the completion of this program of study with my loving and supportive parents, our extended family, and with my loves, Jeri and Brian. Thanks.

Alan C. Nowakowski
INFORMATION TO USERS

This was produced from a copy of a document sent to us for microfilming. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help you understand markings or notations which may appear on this reproduction.

1. The sign or “target” for pages apparently lacking from the document photographed is “Missing Page(s)”. If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure you of complete continuity.

2. When an image on the film is obliterated with a round black mark it is an indication that the film inspector noticed either blurred copy because of movement during exposure, or duplicate copy. Unless we meant to delete copyrighted materials that should not have been filmed, you will find a good image of the page in the adjacent frame. If copyrighted materials were deleted you will find a target note listing the pages in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed the photographer has followed a definite method in “sectioning” the material. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For any illustrations that cannot be reproduced satisfactorily by xerography, photographic prints can be purchased at additional cost and tipped into your xerographic copy. Requests can be made to our Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases we have filmed the best available copy.

University Microfilms International
300 N. ZEEB RD., ANN ARBOR, MI 48106

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy. Problems encountered with this document have been identified here with a check mark ✓.

1. Glossy photographs or pages
2. Colored illustrations, paper or print
3. Photographs with dark background
4. Illustrations are poor copy
5. Pages with black marks, not original copy
6. Print shows through as there is text on both sides of page
7. Indistinct, broken or small print on several pages ✓
8. Print exceeds margin requirements
9. Tightly bound copy with print lost in spine
10. Computer printout pages with indistinct print
11. Page(s) ________ lacking when material received, and not available from school or author.
12. Page(s) ________ seem to be missing in numbering only as text follows.
13. Two pages numbered _________. Text follows.
14. Curling and wrinkled pages
15. Other

University
Microfilms
International

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
<table>
<thead>
<tr>
<th>TABLE OF CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS .............................................. ii</td>
</tr>
<tr>
<td>LIST OF TABLES ..................................................... vi</td>
</tr>
<tr>
<td>LIST OF FIGURES ................................................... viii</td>
</tr>
<tr>
<td>CHAPTER</td>
</tr>
<tr>
<td>I CONTEXT AND PURPOSE OF THE STUDY ......................... 1</td>
</tr>
<tr>
<td>II LITERATURE REVIEW ............................................. 8</td>
</tr>
<tr>
<td>The Title I Evaluation and Reporting System ................ 9</td>
</tr>
<tr>
<td>Model A .......................................................... 15</td>
</tr>
<tr>
<td>Validity of the Equipercentile Assumption .................... 17</td>
</tr>
<tr>
<td>Research on the Equipercentile Assumption ..................... 24</td>
</tr>
<tr>
<td>Summary Conclusion ............................................... 30</td>
</tr>
<tr>
<td>III PROCEDURES .................................................... 31</td>
</tr>
<tr>
<td>The Metropolitan Achievement Tests .............................. 32</td>
</tr>
<tr>
<td>MAT Standardization Sample ...................................... 33</td>
</tr>
<tr>
<td>Matched-Set Student Sample ..................................... 37</td>
</tr>
<tr>
<td>Assignment of Class Identification Numbers .................. 41</td>
</tr>
<tr>
<td>Comparability of Matched-Set Student Sample</td>
</tr>
<tr>
<td>To MAT Standardization Sample ................................ 43</td>
</tr>
<tr>
<td>Class Means Sample ............................................... 45</td>
</tr>
<tr>
<td>NCE' Scores ....................................................... 45</td>
</tr>
<tr>
<td>Analysis Procedures ............................................... 46</td>
</tr>
<tr>
<td>IV RESULTS ......................................................... 50</td>
</tr>
<tr>
<td>NCE Analysis ...................................................... 50</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
LIST OF TABLES

1  NCE Scores at Two Grades for Nearly-All Minority Schools and Nearly-All Majority Schools in Several Cities ........................................... 25

2  Percent of Pupils by School System Enrollment and System Type in National Population and Metropolitan Standardization Sample ........................................ 34

3  Percent of Pupils by Geographic Region in National Population and Metropolitan Standardization Sample ........................................... 34

4  Socioeconomic Status of National Population and Metropolitan Standardization Sample ........................................... 36

5  Ethnic Breakdown of Pupils in National Population and Metropolitan Standardization Sample ........................................... 36

6  Data Tape Record Layout ......................................................... 39

7  Test Levels Administered By Grade ........................................... 39

8  Fall Reading Scale Score Means and Standard Deviations for Matched-Set Student Sample Compared to Total MAT Standardization Sample ........................................... 44

9  Fall Math Scale Score Means and Standard Deviations for Matched-Set Student Sample Compared to Total MAT Standardization Sample ........................................... 44

10 Number of Class Means, Average Class Size, and Standard Deviation of Class Size By Grade ........................................... 45

11 Fall and Spring NCE Reading Means and Standard Deviations of Class Means By Grade ........................................... 52

12 Fall and Spring NCE Math Means and Standard Deviations of Class Means By Grade ........................................... 52

13 Reading NCE Deviation Score Mean Differences, Standard Deviations and Largest Negative and Positive Deviation Scores in Classroom Means By Grade ........................................... 53
LIST OF FIGURES

1 Area Under the Normal Curve Divided into NCEs, Percentiles, and Standard Deviation Units .......... 14
2 Regression Plot of Grade 2 Reading NCE Class Means Fall to Spring ........................................... 57
3 Regression Plot of Grade 3 Reading NCE Class Means Fall to Spring ........................................... 58
4 Regression Plot of Grade 4 Reading NCE Class Means Fall to Spring ........................................... 59
5 Regression Plot of Grade 5 Reading NCE Class Means Fall to Spring ........................................... 60
6 Regression Plot of Grades 2-5 Reading NCE Class Means Fall to Spring ........................................... 61
7 Regression Plot of Grade 2 Math NCE Class Means Fall to Spring ........................................... 62
8 Regression Plot of Grade 3 Math NCE Class Means Fall to Spring ........................................... 63
9 Regression Plot of Grade 4 Math NCE Class Means Fall to Spring ........................................... 64
10 Regression Plot of Grade 5 Math NCE Class Means Fall to Spring ........................................... 65
11 Regression Plot of Grades 2-5 Math NCE Class Means Fall to Spring ........................................... 66
12 Plot of Mean Deviation Scores By Grade Level for Math and Reading ........................................... 69
13 Distribution of NCE Deviation Scores of Grade 2 Reading Class Means ........................................... 70
14 Distribution of NCE Deviation Scores of Grade 3 Reading Class Means ........................................... 70
15 Distribution of NCE Deviation Scores of Grade 4 Reading Class Means ........................................... 71
16 Distribution of NCE Deviation Scores of Grade 5 Reading Class Means .......... 71
17 Distribution of NCE Deviation Scores of Grades 2-5 Reading Class Means .......... 72
18 Distribution of NCE Deviation Scores of Grade 2 Math Class Means .......... 72
19 Distribution of NCE Deviation Scores of Grade 3 Math Class Means .......... 73
20 Distribution of NCE Deviation Scores of Grade 4 Math Class Means .......... 73
21 Distribution of NCE Deviation Scores of Grade 5 Math Class Means .......... 74
22 Distribution of NCE Deviation Scores of Grades 2-5 Math Class Means .......... 74
23 Regression Plot of Grade 2 Reading NCE' Class Means Fall to Spring .......... 81
24 Regression Plot of Grade 3 Reading NCE' Class Means Fall to Spring .......... 82
25 Regression Plot of Grade 4 Reading NCE' Class Means Fall to Spring .......... 83
26 Regression Plot of Grade 5 Reading NCE' Class Means Fall to Spring .......... 84
27 Regression Plot of Grades 2-5 Reading NCE' Class Means Fall to Spring .......... 85
28 Regression Plot of Grade 2 Math NCE' Class Means Fall to Spring .......... 86
29 Regression Plot of Grade 3 Math NCE' Class Means Fall to Spring .......... 87
30 Regression Plot of Grade 4 Math NCE' Class Means Fall to Spring .......... 88
31 Regression Plot of Grade 5 Math NCE' Class Means Fall to Spring .......... 89
32 Regression Plot of Grades 2-5 Math NCE' Class Means Fall to Spring .......... 90

ix

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>Distribution of NCE' Deviation Scores of Grade 2 Reading Class Means</td>
<td>91</td>
</tr>
<tr>
<td>34</td>
<td>Distribution of NCE' Deviation Scores of Grade 3 Reading Class Means</td>
<td>92</td>
</tr>
<tr>
<td>35</td>
<td>Distribution of NCE' Deviation Scores of Grade 4 Reading Class Means</td>
<td>93</td>
</tr>
<tr>
<td>36</td>
<td>Distribution of NCE' Deviation Scores of Grade 5 Reading Class Means</td>
<td>94</td>
</tr>
<tr>
<td>37</td>
<td>Distribution of NCE' Deviation Scores of Grades 2-5 Reading Class Means</td>
<td>95</td>
</tr>
<tr>
<td>38</td>
<td>Distribution of NCE' Deviation Scores of Grade 2 Math Class Means</td>
<td>96</td>
</tr>
<tr>
<td>39</td>
<td>Distribution of NCE' Deviation Scores of Grade 3 Math Class Means</td>
<td>97</td>
</tr>
<tr>
<td>40</td>
<td>Distribution of NCE' Deviation Scores of Grade 4 Math Class Means</td>
<td>98</td>
</tr>
<tr>
<td>41</td>
<td>Distribution of NCE' Deviation Scores of Grade 5 Math Class Means</td>
<td>99</td>
</tr>
<tr>
<td>42</td>
<td>Distribution of NCE' Deviation Scores of Grades 2-5 Math Class Means</td>
<td>100</td>
</tr>
</tbody>
</table>
A new Title I evaluation and reporting system has recently been developed. The system was developed in response to a legislative mandate -- Public Law 93-380, Section 151. This mandate directed the U.S. Office of Education (USOE) to provide evaluation standards and models for evaluating Title I projects so that comparable data would be available nationwide (Wisler & Anderson, 1979).

To carry out this legislative mandate, USOE contracted with the RMC Research Corporation. The system developed by RMC includes three outcome evaluation models or designs -- Model A, the norm-referenced model; Model B, the control group model; and Model C, the special regression model. A common metric is used for reporting outcome results. Each of the models can be used with either criterion- or norm-referenced tests.

Nationally, Title I projects in basic skill learning areas are required to use one of the RMC models for assessing and reporting project impact. The choice of the model and the instrument(s) used to measure learning are locally determined, however. Procedures are specified for implementing each model and for translating test scores to a common metric for reporting and estimating project impact. The reporting metric is referred to as the Normal Curve Equivalent (NCE). The NCE is the link in the system which provides a basis for making outcome comparisons between projects and aggregating outcome results.
at state and national levels.

A number of issues and concerns have been raised in the literature regarding the soundness of the new Title I Evaluation and Reporting System (TIERS) developed by RMC. In the technical area concerns have been raised regarding: (1) threats to internal validity, i.e. can estimates of project impact be trusted, (2) possible noncomparability of treatment effects across models, and (3) validity questions that arise from the required NCE transformations and aggregation of results across different tests (Wisler & Anderson, 1979). Due to the newness of the system, however, a sound research base has not yet been established to answer many of the questions which have been raised.

The three models developed by RMC differ in their research rigor and ease of implementation. Model A, the norm-referenced model, is the least rigorous of the RMC models from a research perspective. It is also likely to be the model which receives the most widespread use (Linn, 1979). It is the easiest of the models to implement; it requires only the pre- and posttesting of project students. On the other hand, Model B, the control group model, requires the identification and testing for a local control group. And Model C, the special regression model, requires following a stringent student selection criterion.

The three models use a common definition of project impact. The definition of impact is expressed as follows (Tallmadge & Wood, 1978, p.6):

\[
\text{Project Impact} = \left\{ \frac{\text{Observed}}{\text{Post-treatment Performance}} \right\} - \left\{ \frac{\text{Expected}}{\text{No-Project Performance}} \right\}
\]
The models are similar in that they all use the mean or median posttest score as the observed post-treatment performance measure. The models differ in the manner in which the no-treatment expectation is estimated. In Models B and C, the no-treatment expectation is estimated from posttest scores of a local comparison group. Model A estimates the no-treatment expectation from normative data.

Under Model A, project students are pre- and posttested. The group's mean pretest status in relation to a national norm is used as the estimate of the no-treatment expectation. It is assumed that the mean posttest score of the project students will be at the same percentile as their pretest mean in the absence of the project. That is, it is assumed that if the project group's status was at the 30th percentile on the pretest, it would also be at the 30th percentile on the posttest unless they participated in some special instructional intervention. If they participated in such an intervention and it had a positive effect, the project students would achieve a higher percentile status on the posttest than they did on the pretest (Tallmadge & Wood, 1978).

No empirical study of representative groups has been conducted on the extent to which group position in relation to the national norm may be expected to vary over time. Model A assumes there is no variance. This seems unlikely. Measurement error alone dictates there will be some variability. How much variance one might reasonably expect, however, is unknown. To the extent that groups actually vary in maintaining their position in relation to the norm, Model A's treatment estimates will be biased. Further, even slight variance in the
degree to which groups maintain their position in relation to the norm may be of concern if the size of the effects to be detected are small, or if instrument sensitivity to detect effects is low.

The purpose of this study is to provide information to help assess how tenable it is to assume that student groups maintain a constant percentile position as a no-treatment expectation. This is done by examining whether or not student group position changes within a norm group over a time period similar to that covered in Title I evaluations. If the assumption made under Model A is correct, the change in percentile rank for student groups within a norm group from one test period to another should be zero. If the assumption is incorrect, one would expect to get a distribution of percentile rank changes for student groups within the norm from one test period to another. A distribution of student group percentile rank changes would have implications for the use of the equipercentile assumption as a no-treatment expectation.

A match-set student subsample of the standardization groups for the 1978 Metropolitan Achievement Test (MAT) is used to examine the extent to which student groups actually maintain a constant percentile position over a six month period, fall to spring. Class means are used as the unit of analysis. Constancy of class percentile rank is examined for four different grade levels on MAT reading and math tests using the NCE score metric. The data base contains approximately 250 classes per grade level.

The MAT is a widely used standardized norm-referenced achievement test typical of the tests used in the field with Model A. MAT test
construction and norming procedures may be viewed as comparable to other commonly used standardized achievement tests. The MAT fall and spring standardizations were designed to have a large number of the same students participate in both testings. Because of this, the subsample of students used in this study can be viewed as highly comparable to a national sample of students at the grade levels investigated.

The intent in using Model A is to make inferences about project effects on a collection of students. Since Model A is used to assess project effects on student groups as opposed to investigating effects for individuals, a group unit of analysis is used in this study. The class was selected as the unit of analysis because it was a convenient group unit to use in terms of the size and composition of the MAT data base obtained for this study.

No size specifications are made in the User's Guide: ESEA Title I Evaluation and Reporting System, 1978, regarding the number of students needed to use Model A. At this time, "n" sizes of typical classrooms are likely to be viewed as acceptable for using Model A. In Michigan for example, the State Bilingual Program, which has adopted TIERS, instructs Bilingual projects that use the Model A design to provide individual school building evaluation reports for each school building having 10 or more students at a grade level in the program (Michigan Department of Education, 1980).

The design of this study includes four replications of data analysis in two subject areas in investigating the degree to which classes maintain a constant percentile position. The degree to which
classroom means maintain a constant percentile rank from fall to spring is examined for grades 2, 3, 4, and 5 on MAT reading and math test scores. These grade levels and subject areas are typical of the grade levels and subject areas at which many Title I projects are directed.

The NCE metric is the scale score used for analyzing the degree to which classes maintain a constant percentile position. NCEs are derived equal-interval normalized standard scale scores which have corresponding percentile rank values. Two kinds of NCE analyses were conducted as part of this study. The primary series was conducted on class mean NCE scores calculated according to procedures specified in TIERS implementation guidelines (Tallmadge & Wood, 1978). Using TIERS procedures for calculating NCE group means and calculating the amount of change in a class mean fall to spring provide an indication of how stable the mean is on the NCE scale for individuals. This is so because test norms for individuals are used as the underlying referent points in TIERS calculation procedures.

One can assume that the major reason the TIERS NCE conversion procedures were set up as they are is that most norm-referenced tests provide only norms for individuals. However, the argument can be made that since TIERS evaluations are interested in estimating group treatment effects, as opposed to individual treatment effects, more appropriate referent points would be group score distributions instead of individual score distributions. In other words, it would be more appropriate to look directly at how stable group scores are in direct relation to each other instead of how stable a group score remains on
a scale for individuals.

As a supplemental analysis in this study, the decision was made to investigate how stable group scores remain in direct relation to each other. NCE scores for reporting on these analyses are referred to as NCE'. Class NCE' values were derived from actual fall and spring class percentile distributions of MAT scale score class means of the sample. The difference in a class's fall and spring NCE' values was then used as an indicator of how much the class's scale score mean had shifted its position fall to spring in relation to the other classes in the sample.

Several different statistical techniques were used to describe the extent to which classes maintain a constant position using NCE and NCE' scores. Variance estimates of score distributions, i.e. standard deviations of NCE deviation scores (spring NCE class mean minus fall NCE class mean) and standard error of regression estimates using fall NCE class mean to predict spring NCE class mean, serve as statistical indicators for summarizing the degree to which classes maintain a constant position. Analysis of variance of class mean NCE deviation scores was used to investigate whether stability of means appears to vary by grade level or subject area test scores.
CHAPTER II

LITERATURE REVIEW

This chapter is a review of information in the literature pertaining to Model A. The major advantage of Model A, as pointed out by Horst, Tallmadge, and Wood (1975), is that it eliminates the need for a local comparison group. As a result, Model A can be implemented at much less cost and effort than conventional designs that require a local comparison group. Furthermore, in situations where a local comparison group is not feasible or practical, Model A can be used. The major disadvantage of Model A, as pointed out by Kaskowitz and Norwood (1977), pertains to the questionable validity of the equipercentile assumption as a no-treatment expectation. Many theoretical and methodological arguments can be made as to why the assumption may not be valid. Information pertaining to these issues is presented in this chapter.

This chapter is divided into several subsections. The first two parts present descriptive information regarding the Title I Evaluation and Reporting System (TIERS) in general and Model A in particular. This is followed by a discussion of different issues that may jeopardize the validity of the equipercentile assumption. These issues are grouped in terms of use of national norms as a comparison base, normal growth expectations, and technical concerns pertaining to test development and assessment procedures. The final section of this chapter reviews research findings regarding the degree to which groups
maintain a constant percentile position. Only a relatively small number of studies have been conducted due to the recency in the development of Model A procedures.

The Title I Evaluation and Reporting System


(d) The Commissioner shall provide to State educational agencies, models for evaluations of all programs conducted under this title [Title I], for their use in carrying out their functions under section 143(a), which shall include uniform procedures and criteria to be utilized by local educational agencies, as well as by the State agency in the evaluation of such programs.

and

(f) The models developed by the Commissioner shall specify objective criteria which shall be utilized in the evaluation of all programs and shall outline techniques (such as longitudinal studies of children involved in such programs) and methodology (such as the use of tests which yield comparable results) for producing data which are comparable on a statewide and nationwide basis. (Tallmadge & Wood, 1978, p.1)

These legislative requirements for the evaluation of Title I of the Elementary and Secondary Education Act (ESEA) are probably more detailed than Congress has ever specified for the evaluation of any educational program (Wisler & Anderson, 1979).

In tracing the development of how the legislation came into existence, several writers (Cross, 1979; Wiley, 1979; Wisler & Anderson, 1979) emphasize the lack of data and information available to Congress.
on which to base decisions for continuing or modifying the program. When Congress began considering the extension of ESEA Title I in January of 1973, some $10 billion had already been appropriated for that program over an eight-year period. There was little to point to regarding how well or poorly Title I was meeting its objectives. It was not that information did not exist regarding individual programs in school districts, but that those data were not collected in such a way that information could be compared among school districts and aggregated for Congressional review (Cross, 1979). A good part of the problem, it was felt, was the wide variation in local evaluation procedures and reports and in state summarization and reporting. This variation made it extremely difficult for one to make sense of the information available to the U.S. Office of Education (USOE) (Gamel, Tallmadge, Wood, & Binkley, 1975; Wargo, Tallmadge, Michaels, Lipe, & Morris, 1972). It also needs to be noted, that several early attempts were made to circumvent this problem by providing information through the use of other procedures, such as review teams, audits, surveys, and centralized national studies. None of these proved to be a very effective solution to the problem (Wisler & Anderson, 1979). It would seem that both factors, the wide variation of procedures and practice in the field and the failure of previous USOE attempts to produce desired information, played a major role in setting the stage for the passage of the highly specific evaluation requirements.

The legislative mandate eventually led USOE to contract with the RMC Research Corporation. The major result of this contract has been the specification of:
1. Evaluation models that focus on achievement data as objective criteria for evaluating project effectiveness.

2. Reporting systems that convert these data into comparable form using uniform procedures, (Wiley, 1979, p.43).

Overview of TIERS

The developed system is referred to as the Title I Evaluation and Reporting System (TIERS). In its present form, it focuses on measuring the impact of Title I projects on academic achievement. Mandated implementation of the system became effective in the 1979-80 school year. Its use is required of all Title I projects serving basic skill areas -- reading, language arts, and mathematics (Tallmadge & Wood, 1978).

Five years have gone into development and preparation for implementation of TIERS in the field. A detailed description of the system and implementation procedures is provided in User's Guide: ESEA Title I Evaluation and Reporting System (Tallmadge & Wood, 1978). Numerous other technical papers and pamphlets regarding various aspects of the system have been prepared by the staff of the RMC Corporation. All these materials are available through the U.S. Office of Education.

The TIERS system specifies the use of three alternative evaluation models or designs. These models are labeled: Model A, the norm-referenced model; Model B, the control group model; and Model C, the special regression model. Each model may be used with either norm-referenced achievement tests (Models A1, B1, and C1) or achievement tests for which normative data are not available (Models A2, B2, and C2).
(Tallmadge & Wood, 1978). However, whenever non-normed tests are used to measure project impact, a nationally normed test must also be administered to provide a common frame of reference for assessing these results in relation to other projects.

The three models use a common definition of project impact in measuring the achievement effects of Title I projects. The gain attributed to the project is defined as the difference between the Title I students' performance on a posttest and an estimate of what their performance would have been had they not participated in the Title I project. This definition is frequently expressed as observed post-treatment performance minus the expected no-treatment performance equals project impact (Tallmadge & Wood, 1978).

The models are similar in that they all use the mean or median posttest score as the observed post-treatment performance measure. The models differ in the manner in which the no-treatment expectation is estimated. The norm-referenced model generates a no-treatment expectation from the assumption that, without the project, the treatment group would maintain its position relative to a national or local norm group from pretest to posttest. The control group model uses the posttest scores of the control group as the no-treatment expectation. In the special regression model, the no-treatment expectation is derived from the relationship between the pretest and posttest scores of students scoring higher than Title I students on the pretest measure (Tallmadge & Wood, 1978).

Regardless of the model used, estimates of project impact are quantified in terms of Normal Curve Equivalents (NCEs) in relation to
a national norm sample. NCEs serve as the TIERs reporting metric and provide the basis for making project comparisons and aggregating results at local and national levels.

**Normal Curve Equivalents**

Normal curve equivalents provide a common metric for reporting estimates of project impact. Procedures are specified for each model for estimating project impact in terms of NCE units with respect to a national norm. When nationally normed tests are used to measure project impact, NCEs can be derived directly from the percentile rank corresponding to each possible test score. When non-normed tests are used to assess project impact, a nationally normed test must also be administered. Normative data on the normed test can be linked to observed results on the non-normed test to measure project impact in NCE units.

The NCE scale is a normalized standard scale which has a mean of 50 and a standard deviation of 21.06. Unlike percentiles, NCE units are equal-interval, thus making it legitimate to perform math computations. The NCE scale matches the percentile scale at values of 1, 50, and 99. The relationship between NCEs, percentiles, and standard deviation units is portrayed in Figure 1. Notice that while the NCE scale points are equally spaced, percentiles are widely spaced at the extremes of the distribution and tightly clustered near the center. For example, the distance between 50 and 60 is the same as the distance between 80 and 90 on the NCE scale. Whereas, the distance between 50 and 60 on the percentile scale is only about half the size of the distance between 80 and 90 on the percentile scale.
FIGURE 1

Area Under the Normal Curve Divided into NCEs, Percentiles, and Standard Deviation Units
Model A

In Model A, normative data are used to substitute for data from a control group. The derivation of treatment effects under Model A as given by Tallmadge and Wood (1976) is as follows:

... It is assumed that, without the Title I treatment, the status of the group at posttest time would be the same as it was at pretest time. Thus the group's pretest percentile becomes the expected no-treatment posttest performance. The observed post-treatment performance is simply the percentile rank corresponding to the group's mean posttest score. If the group's posttest status is higher than the no-treatment expectation, the assumption is made that the improvement resulted from participation in the Title I project (p.4).

Model A can be implemented with tests having national norms, with tests having local norms but no national norms, or with criterion-referenced or other non-normed tests. Whenever tests without national norms are used to measure project impact, a test having national norms must also be administered at pretest or posttest periods. NCE treatment estimates are derived through use of the percentile norms of the nationally normed test using equipercentile equating procedures (Tallmadge & Wood, 1978).

There are two basic implementation rules for using Model A listed in User's Guide: ESEA Title I Evaluation and Reporting System. They are:

1. Project participants must be selected before they are pretested. Using the same set of scores for both selection and pretest purposes invalidates the model.

2. Both pre- and posttesting must be accomplished at or near the midpoint(s) of the time interval(s) during which the norm
groups were actually tested. (Tallmadge & Wood, 1978, p.7)

Standardization data for developing test norms are generally collected over a several week period. Widely used commercially published achievement tests usually provide test norms for fall and spring. Tallmadge and Wood (1978) indicate the midpoint of the norming interval should be considered as the empirical norming date and that pre- and post-testing of project students should be done within two weeks of empirical norming dates.

When tests with national norms are used to measure project impact, the User's Guide: ESEA Title I Evaluation and Reporting System, 1978, lists several allowable variations for obtaining the group no-treatment expectation and the observed posttest-performance in NCE units. Variations listed in the Guide for in-level testing are given below.

**Variation 1:** If the test publisher provides raw-score-to-NCE conversion tables, convert each pretest raw score and posttest raw score to an NCE using the tables for the appropriate grade level and time of year. Calculate the average pretest NCE and posttest NCE. Subtract the pretest NCE (the no-treatment expectation) from the posttest NCE (the observed posttest performance) to determine the NCE gain.

**Variation 2:** If the test publisher provides raw-score-to-percentile conversion tables, first convert the pretest and posttest raw scores to percentiles. Second, convert the percentiles to NCEs. Then proceed as above to calculate the average pretest NCE and posttest NCE and calculate the gain.

**Variation 3:** If the test publisher provides standard-score-to-percentile conversion tables, first convert each pretest raw score and posttest raw score to a standard score. Second, convert each standard score to its percentile equivalent and then convert each percentile to an NCE. Average the pretest NCEs, average the posttest NCEs, and determine the difference between the two averages to obtain the NCE gain. (p.51)
When out-of-level testing takes place, the test publisher's expanded scale score (i.e. the standard score scale spanning all test levels) is averaged and then the conversions to NCE units are made. The specific procedural steps for doing so can be found in the technical paper entitled Out-of-Level Testing (Roberts, 1978). Out-of-level testing refers to testing students with a level of the test other than the test level recommended for a particular grade level by the test publisher. The purpose of out-of-level testing is to test students with the level of the test that is most appropriate for the skill level at which they are functioning. TIERS evaluation procedures allow for out-of-level testing for tests which have a standard score scale spanning all test levels.

Validity of the Equipercentile Assumption

Under Model A, project students are pre- and posttested. The group's mean pretest status in relation to a national norm is used as the estimate of the no-treatment expectation. It is assumed that the mean posttest score of project students will be at the same percentile as their pretest mean in the absence of the project. A wide range of issues and concerns exist which may jeopardize the validity of this assumption for estimating project effects. For purposes of discussion, they are presented under three broad categories of concern -- national norms as a comparison base, normal growth expectation, and technical concerns pertaining to test development and assessment procedures.
National Norms as a Comparison Base

In the conventional comparison group design, a local comparison group is used to provide the basis for estimating the treatment effect. That is, procedures are used to select a local group that is similar to the treatment group, random selection procedures generally seen as superior. With the exception of the treatment, both groups then undergo a similar set of conditions. One is then in the position to argue that differences on the posttest are due to the treatment and not due to other rival explanations such as those pointed out by Campbell and Stanley (1966), e.g. selection, maturation, history, mortality.

In Model A, the local comparison group is substituted with students in the norm who scored at the same percentile level as project students on the pretest, and the no-treatment expectation is that these students maintain a constant percentile position. By definition, there are dissimilarities between any local group and the cross-section of students comprising a national norm. These dissimilarities are possible rival alternatives to the treatment for students maintaining or changing their position relative to the norm. From the way norms are developed, there is no way to determine the characteristics of the treatments or students that obtained particular scores in the norm group -- the characteristics of those students to which the project students are being compared. Thus one has little basis for judging comparability or ruling out alternative hypotheses.

Normal Growth Expectation

With respect to the growth area of concern, a similar set of
problems exist. In the conventional comparison group design, no major assumptions regarding growth are needed. The progress of the comparison group from pre- to posttest is used as the normally expected or no-treatment growth expectation. Model A rests on the major assumption that the normal expectation is for groups to maintain a constant percentile position relative to the norm. Corollaries of this assumption are that groups at different positions in the norm progress at the same rate so that relative position is maintained, and that this is true across grade levels, content areas, varying educational conditions, background conditions, etc. Research has shown many different factors to be related to growth, and logically one would assume that there is no constant growth rate for all individuals and groups. Whether or not different factors affecting growth serve, in effect, to cancel each other out so that groups maintain a relatively constant percentile position, is not known.

At present, we know relatively little about student group growth patterns in relation to test norms. Some of what we do know is not supportive of the equipercentile assumption. For example, a number of studies suggest that the achievement pattern of students fitting compensatory education categories (e.g. Coleman, Campbell, Hobson, McPortland, Mood, Weinfeld, & York, 1966; Mayeske & Beaton, 1975; Armor, Conry-Oseguera, Cox, King, McDonnel, Pascal, Pauly, & Zellman, 1976; Kaskowitz & Norwood, 1977) appears to be that they tend to lose ground over time, relative to the norm population.

Additionally, there is nothing to indicate how much groups may vary or how constantly a position is maintained relative to the norm
in the way test norms are developed or published. Standardized achievement tests are developed to discriminate among individuals. Test norms which are published are for individuals, not groups, and they are cross-sectional, not longitudinal. Test norms provide information about a cross-section of students in a particular grade at a particular time of the school year. They present no information about individuals or subsets of individuals within the norm from one point in the school year to another or from one school year to another. That is, common norming procedures do not track the same individuals from fall to spring or from one grade level to another.

Technical Concerns: Test Development and Assessment Procedures

A number of threats to the constancy of percentile rank can be grouped under the category of technical concerns. Six concern areas which pose technical threats follow.

Norm sampling procedures. The advantage of standardized tests is that they are accompanied by norms. If one accepts the validity of using these tests for program evaluation, the norms need to be accurate. One factor that can affect their accuracy is sampling error (Tallmadge & Horst, 1978). To the degree that the samples on which norms are based are non-representative of the population, the norms will be inaccurate.

Test level articulation. Test developers provide a standard score scale that has the function of providing scores that are continuous across levels of the test. The scale links the different levels of the test, i.e. raw scores on different levels of the test are equated
on the standard score scales and are supposed to represent the same level of achievement. The articulation process (smoothing and fitting empirical data to equate scores across forms and levels of a test) is a source of measurement error. These errors tend to be small but add up in the extremes of the distribution (Murray, Arter, & Faddis, 1979). This type of error is important because different test levels are frequently used for pre- and posttesting. Also, a considerable amount of out-of-level testing takes place in Title I evaluations.

Out-of-level testing is recommended when it is expected that students will get extremely high or low scores on the level recommended for a particular grade. Such scores are considered unreliable and not useful. A lower or higher level of the test is then given instead and the students' scores on the out-of-level test are related back to the norm of interest through means of the scale score.

Murray et al. (1979) cite a number of studies that indicate articulation errors can create problems in program evaluation when out-of-level testing is used and students score in the extremes of the distribution.

Cross-sectional norms. Test norms that are provided by test publishers are cross-sectional and not longitudinal. Different student samples are used to provide the norms at each grade and within a grade for fall and spring. The result is that even though each grade represents the national population at that grade, the students in each grade are different. Fall and spring standardization groups within a grade are also different. There may be some overlap, however, between the fall and spring norm samples, depending on test publishers procedures.
The amount of growth reflected in the norms then, may be affected by using different students to establish the norms instead of tracking the same students (Kaskowitz & Norwood, 1977; Linn, 1979).

**Testing procedures.** A number of errors can influence the process in terms of when and how students are tested. Three examples refer to norm dates, test levels, and treatment period. Regarding norm dates, test publishers usually provide empirical norms for fall and spring. If one tests outside of the dates at which the norms were established, one encounters possible interpolation and/or extrapolation errors (Bridgeman, 1978). Regarding test levels, if students are tested at a level that is either too easy or too hard (floor and ceiling effects), their true status may be under- or overestimated (Faddis & Estes, 1978; Horst et al., 1975). And regarding treatment period, a number of studies indicate that different test administration designs (fall-fall, spring-spring, fall-spring) routinely give different measures of impact (Devito & Long, 1977; Faddis & Estes, 1978; Holthouse, Stofflet, & Tokar, 1976; Pelavin & David, 1977).

**Standardized growth expectation.** The amount of growth needed to maintain position in the norm group varies across grades, tests, percentile status, and subject areas (Kaskowitz & Norwood, 1977; Norwood & Stearns, 1976; Stenner, Hunter, Bland, & Cooper, 1978). This has been determined through use of an index referred to as the "standarized growth expectation" (SGE). The SGE is a ratio of the amount of score change needed from one interval to another to maintain percentile position divided by the standard deviation of scores. A common pattern seen across tests is that achievement growth becomes more difficult.
to detect in the upper grades because learning decelerates at the same time that student scores are becoming more variable -- standard deviations increase (Stenner et al., 1978). Therefore, normal growth is a decreasing fraction of the standard deviation of scores. Said another way, the tests are less sensitive for measuring growth in upper grades than lower grades. The effect of variable SGEs on stability of position in the norm is not known.

Regression effects. The equipercentile assumption involves the prediction of posttest scores from pretest scores. Regression effects could be involved that affect constancy of percentile status. One type of regression can occur when a treatment group is selected on the basis of low scores which are subsequently used as a pretest measure. In general, the pretest scores will contain larger than normal amounts of negative error which will cause any treatment effects to appear larger than they really are (Hiscox & Owens, 1978). This type of regression is substantially controlled in Model A by implementing the guideline calling for separate pretest and selection measures.

Using separate pretest and selection measures will control a significant amount of the regression problem. It is not a complete solution, however. The reason for this is that correlations exist between the selection measure and the pre- and posttest. As argued by Linn (1979), since it is known that correlation between tests given closer together tend to be higher than those between tests given further apart in time, it is reasonable to expect that the selection measure will correlate higher with the pretest than with the posttest. Consequently, there will be more regression toward the mean on the
posttest than on the pretest, which will result in an overestimate of the treatment effect.

Research on the Equipercentile Assumption

Assuming constancy of percentile rank to estimate no-treatment effects has not been an accepted procedure in the past. Relatively little is known regarding the validity of this procedure. Research in this area has been stimulated by Model A and is only in the early stages of development. Much of what has taken place has involved retroactive analysis of existing data banks and has been hampered by a number of problems such as use of cross-sectional norms, test floor and ceiling effects, confounding with treatment variables, and the use of small non-representative groups.

Linn (1979), for example, analyzed data from VanHove, Coleman, Rabben, and Karweit (1970) and Kaskowitz and Norwood (1977) in light of the equipercentile assumption. VanHove et al. had summarized achievement test results for two grade levels in New York, Los Angeles, Chicago, Philadelphia, Detroit, and Baltimore. Test results were presented for grade 6 and either grade 3 or grade 4 on one of three standardized achievement test batteries. Unweighted average percentile ranks were provided for schools categorized by percentage of minority students in the school. Global results reported by VanHove et al. were converted by Linn to NCE scores. The results reported by Linn are reproduced in Table 1. As Linn notes, except for city A, the later-grade NCE is lower than the earlier-grade NCE for the nearly-all minority schools in all cities. The later-grade NCEs are also
TABLE 1

NCE Scores at Two Grades for Nearly-All Minority Schools and Nearly-All Majority Schools in Several Cities\textsuperscript{a}

<table>
<thead>
<tr>
<th>City</th>
<th>Test\textsuperscript{b}</th>
<th>Nearly-All Minority</th>
<th>Nearly-All Majority</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Earlier Grade</td>
<td>Later Grade</td>
<td>Diff.</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>30.7</td>
<td>35.1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>33.0</td>
<td>29.9</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>29.9</td>
<td>26.3</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>33.0</td>
<td>25.3</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>34.4</td>
<td>33.7</td>
</tr>
<tr>
<td>E</td>
<td>2</td>
<td>40.1</td>
<td>33.7</td>
</tr>
<tr>
<td>F</td>
<td>3</td>
<td>23.0</td>
<td>20.4</td>
</tr>
</tbody>
</table>

Note. From "Validity of Inferences Based on the Proposed Title I Evaluation Models" by R.L. Linn, Educational Evaluation and Policy Analysis, 1979, Volume 1, Number 2, p. 27.

\textsuperscript{a}Based on average percentile ranks over parts of the same test battery reported by VanHove et al.

\textsuperscript{b}The tests are: (1) Iowa Tests of Basic Skills, (2) Metropolitan Achievement Tests, and (3) Stanford Achievement Tests.

lower than the earlier grade NCEs in five of the seven instances for nearly-all majority schools. The decline for the nearly-all majority schools is generally less than that for the nearly-all minority schools, however. The unweighted average of the difference in NCEs for the nearly-all minority schools is -2.8 and ranges from -7.7 to +4.4. A major problem regarding this data base, however, is that it is cross-sectional, i.e. the same students are not used in the early and late grade samples, rather than longitudinal.
Linn also computed NCE fall-to-spring NCE differences for a longitudinal sample comprised of a subsample of second through eighth graders who were in both spring and fall test samples for norming the 1970 edition of the Metropolitan Achievement Test. The data had been previously reported on by Kaskowitz and Norwood (1977). Linn found NCE differences ranging from -1.2 to 5.5 NCEs on Total Math score and from -.2 to 2.9 NCEs on Total Reading for a grade as a whole. These data indicate that cross-sectional norms may not yield the same information that longitudinal norms would provide.

In Coleman et al. (1966) test data are provided on the reading and math components of the Sequential Test of Education Progress for samples of third, sixth, ninth, and twelfth graders throughout the country. Coleman et al. developed standard scores (with a mean of 50 and standard deviation of 10) and compared the achievement of ethnic and regional subgroups using the white, metropolitan, Northeast subgroup as a comparison base. Regarding stability, data are provided on the relative achievement gains or losses of 19 subgroups. Two-thirds of the subgroups moved at least .1 standard deviation units from one grade to another. The number of subgroups moving .2 or more standard deviation units varied by grade range and topic. The data argue against the equipercentile assumption, but again, it's of a cross-sectional nature and may further be confounded by the use of the highest-scoring group as the comparison base.

Hiscox and Owen (1978) attempted to examine the equipercentile assumption by tracing the percentile standing of Title I and non-Title I students in Portland, Oregon Public Schools over a four year period.
Four hundred nineteen students were followed from grades 4-7 and 327 students were followed from grades 7-10. Students were divided into four groups on the basis of initial percentile standing -- at or below 5th, 15th, 25th, and 35th.

Results showed that while many of the groups showed fairly consistent percentiles, others showed enough change over the last three years to cause problems with an evaluation based on the assumption of percentile maintenance. About the same amount of variability was noted in the Title I and the non-Title I students. The authors stated that it is difficult to tell whether the variability was due to real differences in student achievement, changes in test levels administered over the four years, or the amount of incomplete data involved in the study.

Kaskowitz and Norwood (1977) conducted a study of norm-referenced evaluation procedures as applied to Project Information Packages (PIPs). Each PIP is an information package of an exemplary education project. The RMC Corporation was contracted to select the projects and design the information packages. Kaskowitz and Norwood's study is an assessment of the norm-reference evaluation procedures RMC used in providing project effectiveness information in PIP packages. The norm-referenced procedures RMC used in these packages were the precursors for Model A of TIERs.

The essence of the norm-referenced procedures used in the PIP evaluations is the same as that of Model A. That is the equipercentile assumption. Kaskowitz and Norwood (1977) describe the procedure as follows:
The norm-reference procedure, as described by Horst et al. (1975) is constructed from a model of normal growth... The normal growth model sets a standard of performance for children who receive no treatment (i.e. children who would be considered members of comparison or control classrooms). In the absence of a control group, this approach relies on the norms of standardized tests to estimate how a group of children would have performed if no treatment had been present... The normal growth model adopted by RMC... assumes that the expected normal growth for children who are not in a special education program is such that, on the average, children maintain the same percentile from pretest to post-test with respect to the norm population. (p.1)

The major difference in the procedures as they were used in the PIP packages compared to Model A of TIERS is that the NCE scale had not yet been made part of the procedures. In the absence of the NCE scale, computations of group mean status in relation to the norm was performed through the use of test publishers' expanded standard scale scores. Individual scale scores were averaged and then the scale score average was converted to the percentile status of the score in the appropriate norm table.

Kaskowitz and Norwood applied the norm-referenced analysis procedures to a subset of 1970 MAT norming data obtained from the Psychological Corporation. The data base contained matched-set fall and spring student test records for grades 2-8. Kaskowitz and Norwood projected spring scale score grade means from fall scale score grade means using the equipercentile assumption. The projected scale score grade means were different (statistically significant at the 0.05 probability level) for all grades except grades 2 and 6 on MAT Total Math and Total Reading test scores. The percentile size difference between projected grade mean and actual ranged between -0.5 percentile
to +5.6 percentile on Total Reading and between -2.2 to +9.2 percentile on Total Math. These results which are longitudinal suggest cross-sectional norms may slightly underestimate average student growth. They also support arguments made in the literature that norms derived from cross-sectional samples of students may not provide the same information as norms based on longitudinal student samples.

Using the MAT data, Kaskowitz and Norwood also developed empirical student growth curves and compared them to equipercentile growth curves by grade and subject areas. Points on their empirical growth curve correspond to pretest scale scores by the average posttest scale scores of students for each pretest scale score. For example, if five students had a scale score of 16 on the pretest and different scores on the posttest, a point on the empirical growth curve is plotted for the coordinates, 16 on the pretest axis, and the posttest average of those five students. The number of students representing each point on the plot varies considerably across the range of the scale. Using this procedure, they found in almost every case (grade and subject area) the empirical curves tend to lie above the equipercentile curves for low pretest scores and they tend to lie below the equipercentile curve for extremely high pretest scores. They found the curves tracked each other closely in the midrange of pretest scores.

It's important to note that Kaskowitz and Norwood plotted average posttest scores of students who had the same pretest scores. This is very different from how projects are comprised in the field. That is,
projects are not made up of students that all score at the same exact point on a pretest, but rather they are comprised of students having a range of different pretest scores.

As a final note regarding the Kaskowitz and Norwood study, they also examined Follow Through Project data and Compensatory Reading program data from PIP information packages to provide information regarding growth of students in compensatory education programs and minority group students. Their data tend to support the hypothesis that use of norms based on the standardization group will lead to an expected posttest score that will be too high for students ordinarily in compensatory education programs.

Summary Conclusion

Under Model A, the no-treatment expectation is derived from normative data. It is assumed that the mean posttest score of project students will be at the same percentile rank as their pretest mean in the absence of the project. A wide range of issues and concerns exist which may jeopardize the validity of the assumption for estimating no-treatment expectations. Assuming constancy of percentile rank to estimate no-treatment expectations has not been an accepted procedure in the past. Relatively little is known regarding the validity of this procedure. Empirical studies are needed to examine the extent to which student groups maintain a constant percentile position to help assess the tenability of the equipercntile assumption as a no-treatment expectation.
CHAPTER III

PROCEDURES

The purpose of this study is to provide information to help assess how tenable it is to assume that student groups maintain a constant percentile position. This is done using a sizable representative subsample of the standardization group for the Metropolitan Achievement Test (MAT) for four grade levels. Class means are used as the unit of analysis. The degree to which classroom means maintain a constant percentile rank from fall to spring is examined for grades 2, 3, 4, and 5, on two MAT test scores, reading and math. Two kinds of NCE analyses are conducted for NCE scores derived from different norm distributions. Several statistics and displays are used to describe the extent to which class means maintain a constant percentile rank from fall to spring.

Study procedures are presented in this chapter. The information is organized into several subsections. The first subsection reviews the Metropolitan Achievement Tests (MAT). This is followed by subsections pertaining to the MAT standardization sample, the matched-set student sample for this study, and parts relating to data processing procedures. Comparative statistics for the matched-set student sample and the MAT fall standardization sample are provided. Descriptive statistics for class means derived from the matched-set student sample are presented. Statistical analyses used to assess the degree to which class means maintain a constant position are then described.

31
The Metropolitan Achievement Tests

The MAT is a broadly used standardized norm-referenced achievement test battery. It is published by the Psychological Corporation. The MAT has been in existence since 1930. The 1978 edition is the fifth edition of the MAT.

The 1978 MAT is a two-component or two-tiered system for assessing achievement. The two components are: Survey Test Battery and Instructional Component Tests. The Survey Battery is designed to provide an overall assessment of student performance in major areas of the curriculum, e.g. reading, math, language. The Instructional Component Tests are designed to provide detailed assessment within an individual curriculum area. The two components can be used separately or in conjunction with each other to provide a comprehensive system of assessment.

A consequence of a two-tiered assessment system was that the Survey Battery was made shorter than previous achievement test batteries. Additionally, single score instead of multiple subtest scores are provided for each major curriculum area. Performance in each content area on the Survey tests is assessed using a sample of items from the Instructional tests. The item sample yields a reliable estimate of overall performance in a content area. The item sample, however, is not large enough for subsets of items to form acceptably reliable subtest scores within a content area on the Survey tests.

The 1978 MAT Survey test consists of eight overlapping batteries (levels), covering the grade range from beginning Kindergarten to end of Grade 12. There are two alternate forms for each battery level. The first two battery levels yield one score each in the areas of
reading, mathematics, and language, as well as a total battery score. Besides these scores, later battery levels (grades 2-10) also provide a score in the areas of science and social studies. MAT complete survey batteries require about three hours to complete. The total number of items varies by battery level and ranges from 203 to 275 items ("MAT Teacher's Manual," 1978). Kuder-Richardson Formula 20 reliability coefficients for MAT subject area tests average approximately 0.88; reliability coefficients for complete battery test scores average approximately 0.98. The reliability coefficients for the reading test averages approximately 0.95 for grade levels used in this study; reliability coefficients for the math test averages approximately 0.89 for grade levels used in this study (MAT Special Report No. 11, 1978).

MAT Standardization Sample

The major index variables used in selecting and describing the MAT standardization group are given as follows (MAT Special Report, No. 8, 1978, p.5):

(a) a socioeconomic index composed of median family income and the percent of high school graduates among persons 25 years of age and over

(b) school system enrollment

(c) public versus nonpublic schools

(d) geographic region

Summary data of index variables for the standardization group in relation to the national population are given in Tables 2 to 5. Table 2 indicates the percentage of students in the standardization sample...
### TABLE 2

Percent of Pupils by School System Enrollment and System Type in National Population and Metropolitan Standardization Sample

<table>
<thead>
<tr>
<th>System Size and Type:</th>
<th>2501+ per grade</th>
<th>501-2500 per grade</th>
<th>201-500 per grade</th>
<th>200 and under per grade</th>
<th>All Non-Public Schools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metropolitan Sample(^a)</td>
<td>23%</td>
<td>30%</td>
<td>18%</td>
<td>19%</td>
<td>10%</td>
</tr>
<tr>
<td>National Population(^b)</td>
<td>25%</td>
<td>30%</td>
<td>18%</td>
<td>17%</td>
<td>10%</td>
</tr>
</tbody>
</table>


### TABLE 3

Percent of Pupils by Geographic Region in National Population and Metropolitan Standardization Sample

<table>
<thead>
<tr>
<th>Geographic Region:</th>
<th>Northeast</th>
<th>Midwest</th>
<th>Southeast</th>
<th>West</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metropolitan Sample(^a)</td>
<td>30%</td>
<td>31%</td>
<td>20%</td>
<td>19%</td>
</tr>
<tr>
<td>National Population(^b)</td>
<td>25%</td>
<td>26%</td>
<td>22%</td>
<td>25%</td>
</tr>
</tbody>
</table>


\(^a\) Metropolitan sample data are averages for fall and spring Survey Battery standardization groups.

\(^b\) National data are from School District Fifth Count Census Bureau Data, 1973.
and in the national population in various sizes and types of school systems. Table 3 shows the breakdown in terms of four geographic areas. Table 4 reports median family income data and the percentage of adult graduates from high school. Table 5 indicates the percentage of students with various ethnic backgrounds included in the norm samples along with comparable data for the national population. These tables are provided to show the care test developers take to defend their samples as representative of the nation as a whole.

Fall and spring norm samples were very comparable in terms of the various selection and stratification variables. Approximately 70% of the schools in the fall program also participated in the spring program, accounting for 83% of the spring sample (MAT Special Report, No. 8, 1978). This high percentage of overlap is not common for all standardized measures. Thus, MAT is an excellent data base for this study.

Both forms of the test were empirically normed twice during the school year, in October 1977 and April 1978. The two forms of the test were distributed randomly within grade, by classroom, within building throughout all schools participating in the program. Assignment of forms to each classroom was done by the publisher. Students who took Form J in the fall were given Form K in the spring and vice versa (MAT Special Report, No. 8, 1978).

MAT descriptions of the standardization group are given in terms of percentage figures. The total number of students tested in the standardization programs for all components of the series is given as being over 550,000 students (MAT Teacher's Manual, 1978).
### TABLE 4

**Socioeconomic Status of National Population and Metropolitan Standardization Sample**

<table>
<thead>
<tr>
<th>SES Variables:</th>
<th>Median Family Income</th>
<th>% of Adults Graduating from High School</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metropolitan Sample&lt;sup&gt;a&lt;/sup&gt;</td>
<td>$9760</td>
<td>54%</td>
</tr>
<tr>
<td>National Population&lt;sup&gt;b&lt;/sup&gt;</td>
<td>$9590</td>
<td>52%</td>
</tr>
</tbody>
</table>


<sup>a</sup> Metropolitan sample data are averages for fall and spring Survey Battery standardization groups.

<sup>b</sup> National data are from School District Fifth Count Census Bureau Data, 1973.

### TABLE 5

**Ethnic Breakdown of Pupils in National Population and Metropolitan Standardization Sample**

<table>
<thead>
<tr>
<th>Ethnic Breakdown:</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metropolitan Sample&lt;sup&gt;a&lt;/sup&gt;</td>
<td>75%</td>
<td>19%</td>
<td>4%</td>
<td>2%</td>
</tr>
<tr>
<td>National Population&lt;sup&gt;b&lt;/sup&gt;</td>
<td>76%</td>
<td>16%</td>
<td>6%</td>
<td>2%</td>
</tr>
</tbody>
</table>


<sup>a</sup> Metropolitan sample data are averages for fall and spring Survey Battery standardization groups.

Matched-Set Student Sample

Fifty school systems were involved in both the fall and spring norming programs. The large overlap in the fall and spring norm samples had been built into the standardization program to improve comparability of norms, and to provide data for a study of growth within the academic year (MAT Special Report, No. 8, 1978).

For internal studies on the MAT the test publishers had developed data tapes of matched-set test records of those students who were tested in both the fall and the spring. A subset of the data on these tapes was used as the data base for this study.

Original Data Source

The matched-set test records of students tested in both the fall and spring were obtained from a representative of The Psychological Corporation. Six computer tapes contained approximately 58,000 K-12 grade student matched-set test scores. The matching had been done by student name within a school district. There were approximately 5000-6000 cases per grade for grades 1-8, and about 2000-2500 at grades 9-12. Each individual student record on the tapes was 2585 columns long. Besides spring and fall test scores and identification information, a record contained individual item data, spring and fall, as well as data from the Otis-Lennon School Ability Test which had also been administered to all students in the fall standardization program.

Individual Student Data

Data with respect to the four grades of interest were copied from
the six MAT tapes and stored on a separate tape. This tape was used as the data base for this study. Not all the data contained on an individual student record on the MAT tapes were copied since not all of the data were relevant to the study. Only identification information and spring and fall test scores were necessary.

Student records from the MAT data tapes for grades 2, 3, 4, and 5 were stored on a separate tape. The number of student records contained on the tape per grade is as follows:

Grade 2: 6174  
Grade 3: 6077  
Grade 4: 5079  
Grade 5: 5088  
Combined Total: 22,418

A tape layout of the information for each individual record is given in Table 6. Basic classification information was copied including school district identification number, building number, grade of the student, level and form of the test, teacher name, student name, and fall and spring test scores. Classification information was used in data processing to check student records for possible errors and to process test scores for analysis.

Approximately half the students were administered Form J in the fall, Form K in the spring, and vice versa for the other half. The level of the test given each grade in fall and spring is listed in Table 7. In grades 2 and 3, the level of the test administered in the spring was different from the level administered in the fall. In grades 4 and 5, the level administered at each grade was the same fall and spring. Thus, the study contains conditions of testing in terms of use of different forms and levels which are similar to real school
### TABLE 6
Data Tape Record Layout

<table>
<thead>
<tr>
<th>Columns</th>
<th>Information</th>
<th>Columns</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-4</td>
<td>School System Number</td>
<td>58-61</td>
<td>School System Number</td>
</tr>
<tr>
<td>5</td>
<td>SES Index</td>
<td>62</td>
<td>SES Index</td>
</tr>
<tr>
<td>6-8</td>
<td>Building Number</td>
<td>63-65</td>
<td>Building Number</td>
</tr>
<tr>
<td>9-10</td>
<td>Grade Level</td>
<td>66-67</td>
<td>Grade Level</td>
</tr>
<tr>
<td>11</td>
<td>Test Level (fall)</td>
<td>68</td>
<td>Test Level (spring)</td>
</tr>
<tr>
<td>12</td>
<td>Test Form (fall)</td>
<td>69</td>
<td>Test Form (spring)</td>
</tr>
<tr>
<td>13</td>
<td>Sex</td>
<td>70</td>
<td>Sex</td>
</tr>
<tr>
<td>14-33</td>
<td>Teacher Name</td>
<td>71-90</td>
<td>Teacher Name</td>
</tr>
<tr>
<td>34-51</td>
<td>Student Name</td>
<td>91-108</td>
<td>Student Name</td>
</tr>
<tr>
<td>52-54</td>
<td>Reading Raw Score (fall)</td>
<td>109-111</td>
<td>Reading Raw Score (spring)</td>
</tr>
<tr>
<td>55-57</td>
<td>Math Raw Score (fall)</td>
<td>112-114</td>
<td>Math Raw Score (spring)</td>
</tr>
</tbody>
</table>

### TABLE 7
Test Levels Administered By Grade

<table>
<thead>
<tr>
<th>Grade</th>
<th>Test Level Administered</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spring</td>
</tr>
<tr>
<td>2</td>
<td>Primary 1</td>
</tr>
<tr>
<td>3</td>
<td>Primary 2</td>
</tr>
<tr>
<td>4</td>
<td>Elementary</td>
</tr>
<tr>
<td>5</td>
<td>Intermediate</td>
</tr>
</tbody>
</table>
testing patterns.

**Data Control Procedures**

There are many possibilities for missing data points and coding errors in a data base the size of the one used in this study. To identify and eliminate errors that could affect the analysis, checks were run for missing test scores and possible coding errors. Two hundred ninety-four (approximately 1.3%) student records on the tape were eliminated because one or more of a student's fall/spring test scores were missing. To search for possible coding errors, cross-tabulations were run on system identification number, grade level, and test level. The results were examined for anomalies. For example, if a student was coded as second grade in the fall then it was expected the student should have the same code for spring. Any mismatches were treated as coding errors and the records were eliminated from the file. A total of seventy-one (71) records were eliminated through spotting anomalies in cross-tabulation results.

**Adding NCE and MAT Scale Scores To Student Records**

The student scores on the data tape were raw score test values. For analysis purposes, corresponding NCE equivalents and MAT expanded scale score equivalents for raw score test values were added to the records. This was done by computer using conversion table listings. A table was made for each grade level, test (reading and Math), form, and test administration (fall and spring) combination in the data file. This required thirty-two different tables in all -- 4 different grade
levels by reading or math tests by two alternate test forms by fall or spring test administrations. Each table listed raw score test values with corresponding NCE and scale score equivalents from MAT norm tables.

Assignment of Class Identification Numbers

The unit of analysis in this study is the classroom mean. A two-digit class identification number was assigned to individual student records to serve as a grouping variable for obtaining class means. Teacher name was used to assign class identification numbers.

To assign class identification numbers, a listing of all unique teacher fall and spring name combinations by school system and grade was printed along with the number of students having that combination. The list contained 1540 different fall-spring teacher name combinations. There were a number of instances in which the fall and spring teacher names were different and a number of instances in which only one or a few students had a particular combination. Sometimes a combination was created because of a misspelling of the teacher's name or the reversal of the teacher's last name and first. Duplicate sets of names in consecutive grades were found because of combination classes e.g. a class contained part second graders and part third graders. Any number of additional reasons could account for teacher names not being the same on an individual's fall and spring record. Some speculative examples are: (1) the teacher may have been absent and a substitute did one of the fall or spring test administrations; (2) a student may have been absent and then taken the test in another classroom, or a
counselor or principal may have done the make-up testing; (3) leveling of classroom sizes by the administration may have occurred after the fall testing resulting in student transfers; (4) the teacher name itself may have changed from fall to spring.

To eliminate errors and set a lower limit on class size, the list of 1540 teacher name combinations was hand screened. A lower limit of ten was set on class size. Screening procedures were followed to prevent class means from entering the analysis that would have been based on fewer than ten students. Screening procedures involved making one of three decisions with respect to each unique teacher fall-spring name combination on the list of 1540. One decision choice was to accept the teacher fall-spring name combination as forming one class unit. This was done if the teacher names were the same fall and spring for a number of students, or if a sufficient number of students had the same two-name set. A second alternative was to combine a set of students having one combination with one or more sets of students with other combinations of names to form a single class unit. This formation was done when there seemed to be clear reason for doing so e.g. misspelling of a teacher name, first/last name reversals, or clear indications of student transfers within a school. The third alternative was to eliminate the combination from the file. This occurred when there were only a few cases (fewer than ten) of a particular combination and there was no way to determine to which class within the school the students belonged. It also occurred where it seemed apparent that a dual grade combination class existed in a school but there were fewer than ten students for one of the grades in the combination. For ex-
ample, if a teacher had 14 third graders and 8 second graders in his/her class, the 14 third graders were kept as a class unit in the grade three data and the 8 second graders were eliminated from the grade two data.

No teacher name combinations were left to stand by themselves or be used as a basis for calculating class means that had fewer than 10 students. Students in fall-spring teacher name combinations that met the criteria listed in the foregoing as designating a class were assigned two digit class identification numbers. The class identification numbers were used to form the units of analysis. The screening process of teacher names resulted in reducing the number of teacher name combinations on which to obtain classroom means from 1540 to 1006. The total number of individual records eliminated was 419 (approximately 1.9% of the sample).

Comparability of Matched-Set Student Sample To MAT Standardization Sample

After data cleaning and score conversions were completed, descriptive statistics of the student sample on the tape were obtained. The size of matched-set student samples by grade were as follows:

<table>
<thead>
<tr>
<th>Grade</th>
<th>Students</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5,872</td>
<td>281</td>
</tr>
<tr>
<td>3</td>
<td>5,834</td>
<td>272</td>
</tr>
<tr>
<td>4</td>
<td>4,957</td>
<td>230</td>
</tr>
<tr>
<td>5</td>
<td>4,971</td>
<td>223</td>
</tr>
<tr>
<td>Combined</td>
<td>21,634</td>
<td>1006</td>
</tr>
</tbody>
</table>

Scale score means and standard deviations for the sample by grade for the fall test administration are given in Tables 8 and 9. Comparable statistics for the total MAT standardization sample are also presented.
Because the test standardization procedure uses the individual student as the unit of analysis, the comparison is presented at the student level and not the unit of this study, i.e. the class. The summary statistics of the matched-set student subset serving as the data base for this study are very similar to the total MAT standardization sample. Mean scale score differences between the subset and the total standardization sample differ, on the average, less than a tenth of the national sample standard deviations.

**TABLE 8**

<table>
<thead>
<tr>
<th>Grade</th>
<th>Study Mean</th>
<th>Total Mean</th>
<th>Study S.D.</th>
<th>Total S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>558.2</td>
<td>569.9</td>
<td>84.2</td>
<td>88.1</td>
</tr>
<tr>
<td>3</td>
<td>632.0</td>
<td>641.6</td>
<td>71.7</td>
<td>74.1</td>
</tr>
<tr>
<td>4</td>
<td>662.7</td>
<td>666.4</td>
<td>69.8</td>
<td>73.8</td>
</tr>
<tr>
<td>5</td>
<td>697.7</td>
<td>701.6</td>
<td>71.0</td>
<td>74.8</td>
</tr>
</tbody>
</table>

**TABLE 9**

<table>
<thead>
<tr>
<th>Grade</th>
<th>Study Mean</th>
<th>Total Mean</th>
<th>Study S.D.</th>
<th>Total S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>453.0</td>
<td>462.6</td>
<td>74.5</td>
<td>74.5</td>
</tr>
<tr>
<td>3</td>
<td>519.3</td>
<td>528.7</td>
<td>73.9</td>
<td>77.1</td>
</tr>
<tr>
<td>4</td>
<td>573.9</td>
<td>576.8</td>
<td>85.4</td>
<td>90.2</td>
</tr>
<tr>
<td>5</td>
<td>621.6</td>
<td>623.3</td>
<td>79.2</td>
<td>85.5</td>
</tr>
</tbody>
</table>
Class Means Sample

The class mean is the unit of analysis used in this study. The matched-set MAT fall/spring student sample was the data base from which the class mean sample for this study was derived. The number of classes, average class size, and standard deviation of class size per grade, for the derived sample of class means used in this study, are listed in Table 10. The number of classes per grade ranges from 223 to 281. Average class size across grades is 21.8. Average standard deviation of class size across grades is 5.3.

**TABLE 10**

<table>
<thead>
<tr>
<th>Grade</th>
<th>Number of Classes</th>
<th>Average Class Size</th>
<th>Standard Deviation of Class Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>281</td>
<td>21.3</td>
<td>5.0</td>
</tr>
<tr>
<td>3</td>
<td>272</td>
<td>21.8</td>
<td>5.3</td>
</tr>
<tr>
<td>4</td>
<td>230</td>
<td>21.8</td>
<td>4.8</td>
</tr>
<tr>
<td>5</td>
<td>223</td>
<td>22.5</td>
<td>6.1</td>
</tr>
<tr>
<td>Combined</td>
<td>1006</td>
<td>21.8</td>
<td>5.3</td>
</tr>
</tbody>
</table>

NCE' Scores

Two kinds of NCE analysis are conducted as part of this study. The primary series, as it replicates procedures in the field, is conducted using NCE scores derived from MAT norms for individuals. A supplemental series of analyses is conducted using NCE scores derived from the observed class scale score percentile distribution of classes.
in the sample. This score is referred to as NCE' in this study.

NCE' scores were derived as follows. Frequency distributions of class scale score means were printed by grade and test subject area, and test administration (grades 2-5 by reading and math by fall and spring). Percentile distributions were derived for each class mean scale score distribution. A conversion table was made for each grade/test/administration combination. Each table listed scale score mean values with their corresponding NCE' equivalent. The NCE' scores were then added by computer to the class means matrix for analysis.

Analysis Procedures

Under Model A, the no-treatment expectation is derived from the equipercentile assumption. That is, in the absence of special treatment, student groups are expected to maintain a constant percentile position in relation to national norms. This assumption is empirically investigated in this study for class means from the national norm. Constancy of class percentile rank is examined for four different grade levels on MAT reading and math tests. Two kinds of NCE analyses are conducted for NCE scores derived from different score distributions.

NCE Analyses

The first series of analyses are directly applicable to current practice in that NCE scores for these analyses were obtained in accordance with TIERS guidelines. The same procedures were followed in this study as those that are used in performing evaluations in the
Individual student raw scores were converted to their NCE score equivalent using MAT norm tables for individuals. The individual NCE scores were then averaged to obtain the classroom mean, which is the group unit of analysis used in this study. Fall/spring class mean NCE differences were then calculated. This analysis series then provides information on how constant the class average NCE score stays on the NCE scale for individuals, as test norms for individuals are used as the underlying referent points in TIERS calculation procedures.

**NCE' Analyses**

The second series of analyses is supplemental. It provides information regarding how constant class means stayed in direct relation to each other, instead of how stable they stayed in relation to MAT norm scores for individuals. NCE scores for these analyses were obtained by deriving actual fall and spring class percentile distributions of class scale score means of the sample. NCE' values were assigned accordingly. This analysis series then provides information on how constant class means stayed in relation to each other as opposed to how stable they remained in relation to norms for individuals. The underlying referent points are the observed percentile distributions of class scale score means of the sample.

**Common Statistical Treatments**

Several statistics and displays are provided regarding the extent to which class units maintain a constant rank from fall to spring. The first series of analyses is performed using NCE class scores based
on MAT norms for individuals. The second series is a repeat of the first using NCE' scores based on the empirical percentile distributions of class scale score means of the sample. Both series of analyses, for each grade level and test subject area, include mean NCE fall to spring differences i.e. the average deviation score difference for spring minus fall NCE class units, minimum and maximum NCE class unit changes, standard deviations of fall to spring deviation scores, correlation and regression statistics and displays. An analysis of variance of deviation scores is performed to determine whether stability of class units appear to vary by grade level or test subject area (reading, math). Deviation score frequency distributions are presented for each test area by grade level. These analyses are used to draw conclusions with respect to the following five questions.

<table>
<thead>
<tr>
<th>Question</th>
<th>Relevant Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Do classes maintain a constant percentile rank?</td>
<td>1. NCE deviation score statistics and regression analysis.</td>
</tr>
<tr>
<td>2. Does the degree to which classes maintain a constant percentile rank vary by grade levels and/or test subject areas?</td>
<td>2. Two-way ANOVA of NCE deviation scores.</td>
</tr>
<tr>
<td>3. How stable are class means in relation to each other?</td>
<td>3. NCE' deviation score statistics and regression analysis.</td>
</tr>
<tr>
<td>4. How much change can be expected to occur in class mean?</td>
<td>4. NCE deviation score frequency distributions and confidence intervals.</td>
</tr>
<tr>
<td>5. How sound is the equipercentile assumption as a no-treatment expectation?</td>
<td>5. Synthesis of results across analyses and replications.</td>
</tr>
</tbody>
</table>
In summary, if the equipercentile assumption holds, classes' fall and spring NCE scores should be the same. Class NCE spring minus fall deviation scores then should be zero, or near zero given that some variation can be expected due to measurement error. Likewise, if the equipercentile assumption holds, the correlation between fall and spring class NCE scores should be one, or almost one, and regression analyses should indicate that fall NCE class mean is a perfect or near perfect predictor for spring NCE class mean. The same logic applies for assessing the equipercentile assumption with respect to NCE' data analyses.
CHAPTER IV

RESULTS

This chapter is divided into two major sections. The first section provides information regarding how constant NCE class means remained from fall to spring at four different grade levels on MAT reading and math tests. NCE class means were obtained according to procedural guidelines specified in the TIERS implementation manual and given in the previous chapter. The second section repeats the analyses of the first section for NCE' class scores derived from the fall and spring empirical percentile class mean scale score distributions of means in the sample.

NCE Analysis

Several statistics and displays are used to describe the extent to which class means maintain a constant percentile rank from fall to spring. The information is organized into four subsections. In the first subsection, central tendency and dispersion statistics are presented for NCE class mean deviation scores. Deviation score statistics are for spring NCE class mean minus fall NCE class mean. In the second subsection, regression statistics and plots of fall NCE class mean on spring NCE class mean are presented. In the third subsection, an analysis of variance of class mean NCE deviation scores is presented. And in the final subsection, histograms of deviation score frequency distributions are presented. In each subsection, statistics...
are presented for each grade level and test subject area. Summary statistics across grade levels for each test subject area are also presented.

Central Tendency and Dispersion of NCE Class Means

Tables 11 and 12 present the NCE mean and standard deviation of class means by grade for fall and spring. Table 11 presents these statistics for reading means, Table 12 for math. Combined average results across grades are also presented.

Tables 13 and 14 present NCE deviation score statistics in class means fall to spring for reading and math respectively. The statistics reported are for NCE deviation scores obtained by subtracting a class's fall NCE mean from its spring NCE mean for all classes in the sample. Each table presents the overall NCE mean change fall to spring, the standard deviation of NCE mean changes, and the largest negative and positive NCE change that occurred in class means by grade.

The equipercentile assumption is that the normal growth expectation is for groups to maintain a constant percentile rank relative to the norm. Based on this assumption, the prediction would be for fall class NCE mean to equal spring class NCE mean. Data in Tables 13 and 14 indicate the NCE distribution of class means shifted approximately 1.5 NCE units up from fall to spring on the MAT reading test and approximately 2.0 NCE units up on the MAT math test. The standard deviation of deviation scores in the reading area was approximately 4 NCE units and in math it was approximately 6 NCE units. The largest fall to spring decrease in a class mean on the MAT reading test was 12.35 NCE units;
### TABLE 11
Fall and Spring NCE Reading Means and Standard Deviations of Class Means By Grade

<table>
<thead>
<tr>
<th>Grade</th>
<th>n</th>
<th>Fall Reading Means</th>
<th></th>
<th>Spring Reading Means</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( \bar{X} )</td>
<td>S.D.</td>
<td>( \bar{X} )</td>
<td>S.D.</td>
</tr>
<tr>
<td>2</td>
<td>281</td>
<td>47.89</td>
<td>9.72</td>
<td>49.43</td>
<td>10.13</td>
</tr>
<tr>
<td>3</td>
<td>272</td>
<td>47.47</td>
<td>10.14</td>
<td>49.21</td>
<td>10.28</td>
</tr>
<tr>
<td>4</td>
<td>230</td>
<td>48.1</td>
<td>9.45</td>
<td>49.47</td>
<td>9.37</td>
</tr>
<tr>
<td>5</td>
<td>223</td>
<td>48.74</td>
<td>8.84</td>
<td>49.84</td>
<td>8.27</td>
</tr>
<tr>
<td>Combined (grades 2-5)</td>
<td>1006</td>
<td>48.01</td>
<td>9.59</td>
<td>49.47</td>
<td>9.6</td>
</tr>
</tbody>
</table>

### TABLE 12
Fall and Spring NCE Math Means and Standard Deviations of Class Means By Grade

<table>
<thead>
<tr>
<th>Grade</th>
<th>n</th>
<th>( \bar{X} )</th>
<th>S.D.</th>
<th>( \bar{X} )</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>281</td>
<td>48.48</td>
<td>10.45</td>
<td>50.72</td>
<td>11.3</td>
</tr>
<tr>
<td>3</td>
<td>272</td>
<td>47.63</td>
<td>9.99</td>
<td>50.1</td>
<td>11.53</td>
</tr>
<tr>
<td>4</td>
<td>230</td>
<td>49.42</td>
<td>9.7</td>
<td>51.16</td>
<td>10.68</td>
</tr>
<tr>
<td>5</td>
<td>223</td>
<td>49.29</td>
<td>8.88</td>
<td>50.49</td>
<td>8.75</td>
</tr>
<tr>
<td>Combined (grades 2-5)</td>
<td>1006</td>
<td>48.65</td>
<td>9.83</td>
<td>50.62</td>
<td>10.7</td>
</tr>
</tbody>
</table>
### TABLE 13

Reading NCE Deviation Score Mean Differences, Standard Deviations and Largest Negative and Positive Deviation Scores In Classroom Means By Grade

<table>
<thead>
<tr>
<th>Grade</th>
<th>n</th>
<th>Mean of Deviation Scores</th>
<th>Standard Deviation of Deviation Scores</th>
<th>Largest Deviation In Class Means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Negative</td>
</tr>
<tr>
<td>2</td>
<td>281</td>
<td>1.54</td>
<td>4.53</td>
<td>-12.35</td>
</tr>
<tr>
<td>3</td>
<td>272</td>
<td>1.74</td>
<td>3.72</td>
<td>-7.57</td>
</tr>
<tr>
<td>4</td>
<td>230</td>
<td>1.37</td>
<td>3.80</td>
<td>-7.46</td>
</tr>
<tr>
<td>5</td>
<td>223</td>
<td>1.10</td>
<td>4.58</td>
<td>-9.06</td>
</tr>
<tr>
<td>Combined (grades 2-5)</td>
<td>1006</td>
<td>1.46</td>
<td>4.17</td>
<td>-12.35</td>
</tr>
</tbody>
</table>

### TABLE 14

Math NCE Deviation Score Mean Differences, Standard Deviations and Largest Negative and Positive Deviation Scores In Classroom Means By Grade

<table>
<thead>
<tr>
<th>Grade</th>
<th>n</th>
<th>Mean of Deviation Scores</th>
<th>Standard Deviation of Deviation Scores</th>
<th>Largest Deviation In Class Means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Negative</td>
</tr>
<tr>
<td>2</td>
<td>281</td>
<td>2.24</td>
<td>7.21</td>
<td>-11.67</td>
</tr>
<tr>
<td>3</td>
<td>272</td>
<td>2.47</td>
<td>6.41</td>
<td>-15.36</td>
</tr>
<tr>
<td>4</td>
<td>230</td>
<td>1.87</td>
<td>4.22</td>
<td>-7.62</td>
</tr>
<tr>
<td>5</td>
<td>223</td>
<td>1.2</td>
<td>4.82</td>
<td>-10.88</td>
</tr>
<tr>
<td>Combined (grades 2-5)</td>
<td>1006</td>
<td>1.97</td>
<td>5.92</td>
<td>-15.36</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
the largest increase was 40.17 NCE units. The largest fall to spring decrease in a class NCE mean on the MAT math test was 15.36 NCE units; the largest increase was 46.49 NCE units.

**Regression of Fall on Spring NCE Class Means**

Regression statistics and plots of fall on spring NCE class means are presented in this section. Tables 15 and 16 present regression analysis statistics based on using fall NCE class means to predict spring NCE class means. Table 15 presents regression statistics for reading class means and Table 16 for math. Each table includes: (1) the correlation coefficient between fall and spring class means; (2) $r^2$ squared, the percent of variance accounted for using fall class mean to predict spring class mean; (3) the standard error of the estimate; (4) the slope of the regression line; and (5) the 95% confidence limits for predicting spring NCE class means.

Based on the equipercentile assumption, the correlation coefficient should be one; $r^2$ should be one; the slope of the regression line should be one. Statistics in Tables 15 and 16 indicate correlation coefficients for the study sample are approximately 0.91 on the MAT reading test and approximately 0.84 on the MAT math test. Corresponding $r^2$ values are 0.82 and 0.70 respectively. The slopes are approximately 0.91 on both MAT reading and MAT math tests. The standard error of the estimate is approximately 4 NCE units on the reading test and approximately 6 NCE units on the math test; corresponding respective 95% confidence intervals are approximately plus or minus 8 NCE units for a reading class mean and plus or minus 11.5 NCE units.
### TABLE 15
Regression Statistics Using Fall NCE Reading Class Means To Predict Spring NCE Reading Class Means

<table>
<thead>
<tr>
<th>Grade</th>
<th>n</th>
<th>r</th>
<th>r²</th>
<th>Standard Error of Estimate</th>
<th>Slope</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>281</td>
<td>.90</td>
<td>.80</td>
<td>4.50</td>
<td>.94</td>
<td>± 8.72</td>
</tr>
<tr>
<td>3</td>
<td>272</td>
<td>.93</td>
<td>.87</td>
<td>3.69</td>
<td>.95</td>
<td>± 7.23</td>
</tr>
<tr>
<td>4</td>
<td>230</td>
<td>.92</td>
<td>.85</td>
<td>3.71</td>
<td>.91</td>
<td>± 7.27</td>
</tr>
<tr>
<td>5</td>
<td>223</td>
<td>.86</td>
<td>.74</td>
<td>4.25</td>
<td>.80</td>
<td>± 8.33</td>
</tr>
<tr>
<td>Combined (grades 2-5)</td>
<td>1006</td>
<td>.91</td>
<td>.82</td>
<td>4.08</td>
<td>.91</td>
<td>± 8.00</td>
</tr>
</tbody>
</table>

### TABLE 16
Regression Statistics Using Fall NCE Math Class Means To Predict Spring NCE Math Class Means

<table>
<thead>
<tr>
<th>Grade</th>
<th>n</th>
<th>r</th>
<th>r²</th>
<th>Standard Error of Estimate</th>
<th>Slope</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>281</td>
<td>.78</td>
<td>.61</td>
<td>7.04</td>
<td>.85</td>
<td>± 13.80</td>
</tr>
<tr>
<td>3</td>
<td>272</td>
<td>.83</td>
<td>.69</td>
<td>6.41</td>
<td>.96</td>
<td>± 12.56</td>
</tr>
<tr>
<td>4</td>
<td>230</td>
<td>.92</td>
<td>.84</td>
<td>4.23</td>
<td>1.01</td>
<td>± 8.29</td>
</tr>
<tr>
<td>5</td>
<td>223</td>
<td>.85</td>
<td>.72</td>
<td>4.61</td>
<td>.84</td>
<td>± 9.04</td>
</tr>
<tr>
<td>Combined (grades 2-5)</td>
<td>1006</td>
<td>.84</td>
<td>.70</td>
<td>5.85</td>
<td>.91</td>
<td>± 11.47</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
for a math class mean.

Regression plots for fall on spring NCE class means are presented in Figures 2 to 11. Figures 2 to 5 present regression plots for each grade, grade 2 through grade 5, on the MAT reading test. Figure 6 presents the reading regression plot for grades 2-5 combined. Figures 7 to 11 present the grade by grade math regression plots. Figure 11 presents the math regression plot for grades 2-5 combined. Based on the equipercntile assumption each plot should look like a straight line. To the extent that data points appear away from a line at a 45 degree angle to the base line, the assumption does not hold.

Analysis of Variance

An analysis of variance of class mean NCE deviation scores was conducted to assess systematic differences by grade and test subject area. Table 17 presents the Two-Way Analysis of Variance Summary Table for the one between- and one within- subjects variable design used. The between subject independent variable is grade levels (grades 2-5); the within subject independent variable is test area score (reading and math); and the dependent variable is class mean NCE deviation scores. The F ratio is not statistically significant at the .05 level for between grade level differences; it is statistically significant at the .05 level for the main effect of reading NCE class deviation score differences versus math NCE class deviation score differences; interaction effects are not statistically significant. Tables 18 and 19 list respectively for reading and math class mean deviation scores and the standard deviations of class mean deviation scores by grade level. A
FIGURE 2

Regression Plot of Grade 2 Reading NCE Class Means Fall (Across) to Spring (Down)
FIGURE 3

Regression Plot of Grade 3 Reading NCE Class Means Fall (Across) to Spring (Down)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
FIGURE 4

Regression Plot of Grade 4 Reading NCE
Class Means Fall (Across)
to Spring (Down)
FIGURE 5
Regression Plot of Grade 5 Reading NCE
Class Means Fall (Across)
to Spring (Down)
FIGURE 6
Regression Plot of Grades 2-5 Reading NCE
Class Means Fall (Across)
to Spring (Down)
FIGURE 7

Regression Plot of Grade 2 Math NCE Class Means Fall (Across) to Spring (Down)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
FIGURE 9
Regression Plot of Grade 4 Math NCE Class Means Fall (Across) to Spring (Down)
FIGURE 10
Regression Plot of Grade 5 Math NCE Class Means Fall (Across) to Spring (Down)
FIGURE 11
Regression Plot of Grades 2-5 Math NCE Class Means Fall (Across) to Spring (Down)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
### TABLE 17

ANOVA Summary Table of NCE Class Mean Deviation Scores For Grade Level By Test Area

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Class Mean Deviations</td>
<td>1005</td>
<td>34,414.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade Levels</td>
<td>3</td>
<td>262.78</td>
<td>87.59</td>
<td>2.57</td>
</tr>
<tr>
<td>Class Mean Deviations/Grade Levels</td>
<td>1002</td>
<td>34,151.40</td>
<td>34.08</td>
<td></td>
</tr>
<tr>
<td>Within Class Mean Deviations</td>
<td>1006</td>
<td>18,397.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test Areas</td>
<td>1</td>
<td>135.26</td>
<td>135.26</td>
<td>7.49*</td>
</tr>
<tr>
<td>Grade Levels/Test Areas</td>
<td>3</td>
<td>37.71</td>
<td>12.57</td>
<td>0.7</td>
</tr>
<tr>
<td>Class Mean Deviations and Test Areas/Grade Levels</td>
<td>1002</td>
<td>18,096.60</td>
<td>18.06</td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>2001</td>
<td>52,811.27</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Significant at .05 level.

### TABLE 18

Average Class Mean NCE Deviation Score By Grade Level and Test Subject Area

<table>
<thead>
<tr>
<th>Grade</th>
<th>(1) Mean Reading Deviation Scores</th>
<th>(2) Mean Math Deviation Scores</th>
<th>Combined (Reading &amp; Math) Average of (1) and (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>281</td>
<td>1.54</td>
<td>2.24</td>
</tr>
<tr>
<td>3</td>
<td>272</td>
<td>1.74</td>
<td>2.47</td>
</tr>
<tr>
<td>4</td>
<td>232</td>
<td>1.37</td>
<td>1.87</td>
</tr>
<tr>
<td>5</td>
<td>223</td>
<td>1.10</td>
<td>1.2</td>
</tr>
<tr>
<td>Combined (2-5)</td>
<td>1006</td>
<td>1.46</td>
<td>1.97</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
TABLE 19
Standard Deviations of Class Mean Deviation Scores
By Grade Level and Test Subject Areas

<table>
<thead>
<tr>
<th>Grade</th>
<th>n</th>
<th>(1) Standard Deviation of Class Reading Deviation Scores</th>
<th>(2) Standard Deviation of Class Math Deviation Scores</th>
<th>(3) Combined (Reading &amp; Math)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>281</td>
<td>4.53</td>
<td>7.21</td>
<td>6.03</td>
</tr>
<tr>
<td>3</td>
<td>272</td>
<td>3.72</td>
<td>6.41</td>
<td>5.25</td>
</tr>
<tr>
<td>4</td>
<td>230</td>
<td>3.80</td>
<td>4.22</td>
<td>4.02</td>
</tr>
<tr>
<td>5</td>
<td>223</td>
<td>4.58</td>
<td>4.82</td>
<td>4.7</td>
</tr>
<tr>
<td>Combined (2-5)</td>
<td>1006</td>
<td>4.17</td>
<td>5.92</td>
<td>5.13</td>
</tr>
</tbody>
</table>

plot of grade level mean deviation scores of NCE class means for both reading and math is given in Figure 12.

NCE Deviation Score Frequency Distributions

Presented next are the NCE spring minus fall class mean deviation score frequency distributions. The equipercentile assumption is that the normal growth expectation is for groups to maintain a constant percentile position, i.e. the deviations should all be zero. The degree to which NCE class means changed fall to spring by grade levels and test subject areas is displayed in histograms of the deviation score frequency distributions with a line in the histogram drawn to show the assumed no growth point. Figures 13 to 16 present deviation score frequency distributions for each grade on the MAT reading test. Figure 17 presents the deviation score frequency distribution for
grades 2-5 combined. Figures 18 to 21 present deviation score frequency distributions grade by grade on the MAT math test. The grade 2-5 combined results are presented in Figure 22.

NCE' Analysis

The analyses in the first section of this chapter were conducted on class mean NCE scores calculated according to procedures specified in User's Guide: ESEA Title I Evaluation and Reporting System, 1978. Using TIERS procedures for calculating NCE class means, and calculating the amount of change in class means fall to spring, provided data
FIGURE 13
Distribution of NCE Deviation Scores of Grade 2 Reading Class Means

FIGURE 14
Distribution of NCE Deviation Scores of Grade 3 Reading Class Means
FIGURE 15

Distribution of NCE Deviation Scores of Grade 4 Reading Class Means

FIGURE 16

Distribution of NCE Deviation Scores of Grade 5 Reading Class Means
FIGURE 17

Distribution of NCE Deviation Scores of Grades 2-5 Reading Class Means.

FIGURE 18

Distribution of NCE Deviation Scores of Grade 2 Math Class Means.
FIGURE 19

Distribution of NCE Deviation Scores of Grade 3 Math Class Means

FIGURE 20

Distribution of NCE Deviation Scores of Grade 4 Math Class Means

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
FIGURE 21
Distribution of NCE Deviation Scores of Grade 5 Math Class Means

Range of Deviation Scores: -11 to +26
2σ = 12% % Graphed: 98.7%

FIGURE 22
Distribution of NCE Deviation Scores of Grades 2-5 Math Class Means

Range of Deviation Scores: -15 to +46
2σ = 12% % Graphed: 99.4%

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
for describing the stability of class means fall to spring on the NCE scale for individuals.

Statistics and displays presented in this section are based on class mean NCE' scores. NCE' values were derived from actual fall and spring percentile distributions of MAT scale score class means of the sample. The distribution of class means is the referent point for assigning NCE' values. Analyses of NCE' values, then, indicate how much class means moved fall to spring in relation to each other, as opposed to how much class means changed on the NCE scale for individuals.

The NCE' statistics and displays presented parallel the analyses presented in the first section. The analyses are presented in the following order: Central Tendency and Dispersion of NCE' Class Means, Regression of Fall on Spring NCE' Class Means, and NCE' Deviation Score Frequency Distributions.

Central Tendency and Dispersion of NCE' Class Means

The mean and standard deviation of the NCE scale are 50 and 21.06 respectively. NCE' values were assigned to class scale score means based on the observed percentile distributions. The resulting NCE' class mean fall and spring distributions then should have means of 50 and standard deviations of 21.06. Tables 20 and 21 present the NCE' mean and standard deviation of class means by grade for fall and spring. The means and standard deviation are within rounding error of each distribution having the expected mean of 50 and standard deviation of 21.06.
TABLE 20
Fall and Spring NCE' Reading Means and Standard Deviations of Class Means By Grade

<table>
<thead>
<tr>
<th>Grade</th>
<th>n</th>
<th>Fall Reading Means</th>
<th>Spring Reading Means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>X</td>
<td>S.D.</td>
</tr>
<tr>
<td>2</td>
<td>281</td>
<td>50.00</td>
<td>20.70</td>
</tr>
<tr>
<td>3</td>
<td>272</td>
<td>50.01</td>
<td>20.76</td>
</tr>
<tr>
<td>4</td>
<td>230</td>
<td>49.97</td>
<td>20.79</td>
</tr>
<tr>
<td>5</td>
<td>223</td>
<td>50.05</td>
<td>20.76</td>
</tr>
<tr>
<td>Combined (2-5)</td>
<td>1006</td>
<td>50.01</td>
<td>20.72</td>
</tr>
</tbody>
</table>

TABLE 21
Fall and Spring NCE' Math Means and Standard Deviations of Class Means By Grade

<table>
<thead>
<tr>
<th>Grade</th>
<th>n</th>
<th>Fall Math Means</th>
<th>Spring Math Means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>X</td>
<td>S.D.</td>
</tr>
<tr>
<td>2</td>
<td>281</td>
<td>49.97</td>
<td>20.73</td>
</tr>
<tr>
<td>3</td>
<td>272</td>
<td>50.06</td>
<td>20.72</td>
</tr>
<tr>
<td>4</td>
<td>230</td>
<td>49.97</td>
<td>20.79</td>
</tr>
<tr>
<td>5</td>
<td>223</td>
<td>49.98</td>
<td>20.72</td>
</tr>
<tr>
<td>Combined (2-5)</td>
<td>1006</td>
<td>50.00</td>
<td>20.71</td>
</tr>
</tbody>
</table>
Tables 22 and 23 present NCE' deviation score statistics in class means fall to spring for reading and math respectively. The statistics reported are for NCE' deviation scores obtained by subtracting a class's fall NCE' mean from its spring NCE' mean. Each table presents the overall NCE' mean change fall to spring, standard deviation of NCE' mean changes, and the largest negative and positive NCE' changes that occurred among class means by grade. The means of the deviation scores as already indicated round to zero. The standard deviations of deviation scores are approximately 9 NCE units in reading and 12 NCE units in math. The largest fall to spring decrease in a class mean on the MAT reading test was 25.2 NCE' units; the largest increase was 58.9 NCE' units. The largest decrease in a mean on the MAT math test was 29.4 NCE' units; the largest increase was 65.4 NCE' units.

**TABLE 22**

Reading NCE' Deviation Score Mean Differences, Standard Deviations, and Largest Negative and Positive Deviation Scores In Classroom Means By Grade

<table>
<thead>
<tr>
<th>Grade</th>
<th>n</th>
<th>Mean of Deviation Scores</th>
<th>Standard Deviation of Deviation Scores</th>
<th>Largest Deviation In Class Means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Negative</td>
</tr>
<tr>
<td>2</td>
<td>281</td>
<td>0.02</td>
<td>9.42</td>
<td>24.9</td>
</tr>
<tr>
<td>3</td>
<td>272</td>
<td>0.07</td>
<td>8.51</td>
<td>25.2</td>
</tr>
<tr>
<td>4</td>
<td>230</td>
<td>0.00</td>
<td>8.46</td>
<td>20.8</td>
</tr>
<tr>
<td>5</td>
<td>223</td>
<td>-0.06</td>
<td>10.54</td>
<td>24.2</td>
</tr>
<tr>
<td>Combined (2-5)</td>
<td>1006</td>
<td>0.01</td>
<td>9.22</td>
<td>25.2</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
<table>
<thead>
<tr>
<th>Grade</th>
<th>n</th>
<th>Mean of Deviation Scores</th>
<th>Standard Deviation of Deviation Scores</th>
<th>Largest Deviation In Class Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>281</td>
<td>-0.04</td>
<td>13.34</td>
<td>-28.8</td>
</tr>
<tr>
<td>3</td>
<td>272</td>
<td>-0.11</td>
<td>13.19</td>
<td>-29.4</td>
</tr>
<tr>
<td>4</td>
<td>230</td>
<td>0.04</td>
<td>8.48</td>
<td>-22.3</td>
</tr>
<tr>
<td>5</td>
<td>223</td>
<td>0.12</td>
<td>11.68</td>
<td>-23.4</td>
</tr>
<tr>
<td>Combined (2-5)</td>
<td>1006</td>
<td>-0.01</td>
<td>11.96</td>
<td>-29.4</td>
</tr>
</tbody>
</table>

Regression of Fall on Spring NCE' Class Means

Regression statistics and plots of fall on spring NCE' class means are presented in this section. Tables 24 and 25 present regression analysis statistics based on using fall NCE' class means to predict spring NCE' class means. Table 24 presents regression statistics for reading class means and Table 25 for math. Each table includes: (1) the correlation coefficient between fall and spring class means; (2) \( r^2 \), the percent of variance accounted for using fall class mean to predict spring class mean; (3) the standard error of the estimate; (4) the slope of the regression line; and (5) the 95% confidence limits for predicting spring NCE class means. NCE' correlation coefficients are approximately 0.9 for reading scores and 0.84 for math scores; corresponding \( r^2 \) values are 0.81 and .70 respectively. The slopes are approximately 0.9 in reading and 0.83 in math. The standard
### TABLE 24
Regression Statistics Using Fall NCE' Reading Class Means To Predict Spring NCE' Reading Class Means

<table>
<thead>
<tr>
<th>Grade</th>
<th>n</th>
<th>r</th>
<th>$r^2$</th>
<th>Standard Error of Estimate</th>
<th>Slope</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>281</td>
<td>.90</td>
<td>.80</td>
<td>9.20</td>
<td>.90</td>
<td>± 18.03</td>
</tr>
<tr>
<td>3</td>
<td>272</td>
<td>.92</td>
<td>.84</td>
<td>8.37</td>
<td>.92</td>
<td>± 16.40</td>
</tr>
<tr>
<td>4</td>
<td>230</td>
<td>.92</td>
<td>.84</td>
<td>8.29</td>
<td>.92</td>
<td>± 16.25</td>
</tr>
<tr>
<td>5</td>
<td>223</td>
<td>.87</td>
<td>.76</td>
<td>10.24</td>
<td>.87</td>
<td>± 20.06</td>
</tr>
<tr>
<td>Combined (2-5)</td>
<td>1006</td>
<td>.90</td>
<td>.81</td>
<td>9.01</td>
<td>.90</td>
<td>± 17.65</td>
</tr>
</tbody>
</table>

### TABLE 25
Regression Statistics Using Fall NCE' Math Class Means To Predict Spring NCE' Math Class Means

<table>
<thead>
<tr>
<th>Grade</th>
<th>n</th>
<th>r</th>
<th>$r^2$</th>
<th>Standard Error of Estimate</th>
<th>Slope</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>281</td>
<td>.79</td>
<td>.63</td>
<td>12.63</td>
<td>.79</td>
<td>± 24.76</td>
</tr>
<tr>
<td>3</td>
<td>272</td>
<td>.80</td>
<td>.64</td>
<td>12.55</td>
<td>.80</td>
<td>± 24.60</td>
</tr>
<tr>
<td>4</td>
<td>230</td>
<td>.92</td>
<td>.84</td>
<td>8.29</td>
<td>.91</td>
<td>± 16.25</td>
</tr>
<tr>
<td>5</td>
<td>223</td>
<td>.84</td>
<td>.71</td>
<td>11.27</td>
<td>.85</td>
<td>± 22.08</td>
</tr>
<tr>
<td>Combined (2-5)</td>
<td>1006</td>
<td>.84</td>
<td>.70</td>
<td>11.46</td>
<td>.83</td>
<td>± 22.45</td>
</tr>
</tbody>
</table>
error of the estimate is approximately 9 NCE units for reading scores and 11.5 NCE units for math. Corresponding respective 95% confidence intervals are approximately plus or minus 18 NCE's for reading and plus or minus 22.5 NCE's for math.

Regression plots for fall on spring NCE class means are presented in Figures 23 to 32. Figures 23 to 26 present grade by grade regression plots for reading. Figure 27 presents the reading regression plot for grades 2-5 combined. Figures 28 to 31 present the grade by grade math regression plots. Figure 32 presents the math regression plot for grades 2-5 combined. If the assumptions hold, the plots should show a line at a 45 degree angle to the base.

NCE' Deviation Score Frequency Distributions

The degree to which NCE class means changed their position from fall to spring by grade levels and test subject areas is displayed in histograms of deviation score frequency distributions. Figures 33 to 36 present grade by grade deviation score frequency distributions on the reading test. Figure 37 presents the combined grade 2-5 reading deviation score frequency distribution. Figures 38 to 41 present math grade by grade deviation score frequency distributions. The combined grades 2-5 math deviation score frequency distribution is presented in Figure 42. A line at zero change is drawn on each histogram.
FIGURE 23
Regression Plot of Grade 2 Reading NCE
Class Means Fall (Across)
to Spring (Down)
FIGURE 24
Regression Plot of Grade 3 Reading NCE¹
Class Means Fall (Across)
to Spring (Down)
FIGURE 25
Regression Plot of Grade 4 Reading NCE
Class Means Fall (Across) to Spring (Down)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
FIGURE 26
Regression Plot of Grade 5 Reading NCE'
Class Means Fall (Across)
to Spring (Down)
FIGURE 27
Regression Plot of Grades 2-5 NCE Class Means Fall (Across) to Spring (Down)
FIGURE 28

Regression Plot of Grade 2 Math NCE
Class Means Fall (Across)
to Spring (Down)
FIGURE 29
Regression Plot of Grade 3 Math NCE
Class Means Fall (Across)
to Spring (Down)
FIGURE 30
Regression Plot of Grade 4 Math NCE
Class Means Fall (Across)
to Spring (Down)
FIGURE 31
Regression Plot of Grade 5 Math NCE¹
Class Means Fall (Across)
to Spring (Down)
FIGURE 32
Regression Plot of Grades 2-5 Math NCE
Class Means Fall (Across) to Spring (Down)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
### FIGURE 33

Distribution of NCE' Deviation Scores of Grade 2 Reading Class Means

<table>
<thead>
<tr>
<th>Range of Deviation Scores: -25 to +57</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 2x = 75 )</td>
</tr>
<tr>
<td>Graphed: 97.6%</td>
</tr>
</tbody>
</table>
FIGURE 34
Distribution of NCE' Deviation Scores of Grade 3 Reading Class Means
FIGURE 35

Distribution of NCE' Deviation Scores of Grade 4 Reading Class Means

Range of Deviation Scores: -21 to +43

2σ = 15 t Graphed: 98.3%
FIGURE 36

Distribution of NCE Deviation Scores of Grade 5 Reading Class Means
FIGURE 37

Distribution of NCE' Deviation Scores of Grades 2-5 Reading Class Means
FIGURE 38

Distribution of NCE' Deviation Scores of Grade 2 Math Class Means
Range of Deviation Scores: -29 to +47

FIGURE 39

Distribution of NCE\(^{1}\) Deviation Scores of Grade 3 Math Class Means

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
FIGURE 40
Distribution of NCE' Deviation Scores of Grade 3 Math Class Means

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
FIGURE 41

Distribution of NCE Deviation Scores of Grade 5 Math Class Means

Range of Deviation Scores: -23 to +60
2X = 12% % Graphed: 98.8%
FIGURE 42

Distribution of NCE Deviation Scores of Grades 2-5 Math Class Means

Range of Deviation Scores: -29 to +65
$2\sigma = 1\%$ Graphed: 97.9%
CHAPTER V

CONCLUSIONS

The basic question of interest in this study is the degree to which student groups can be expected to maintain a constant percentile position. This was investigated using a matched-set subsample of the 1978 MAT standardization group. Class means were used as the unit of analysis. The degree to which class means maintain a constant percentile rank fall to spring was reported for four grade levels on two subject area tests, reading and math. Two kinds of NCE analyses were conducted for NCE scores derived from different score distributions. Several analysis techniques were used to describe the extent to which class means maintained a constant percentile rank. Conclusions based on the results of these analyses follow.

Do Classes Maintain A Constant Percentile Rank?

TIERS calculations procedures were used to obtain NCE class means. Based on the no-treatment expectation of Model A, one would predict that the deviation score frequency distribution would have a mean of zero and a variance of zero. The mean of class deviation scores obtained in this study across grades on the reading test was approximately 1.5 NCEs and the variance (standard deviation squared) was approximately 17.5 NCEs. On the math test, the mean of class deviation scores across grades was approximately 2.0 NCEs and the variance was approximately 35 NCEs.
Additionally, regression statistics were obtained using the fall NCE class mean to predict spring class mean. Based on the equipercen-
tile assumption, the fall NCE mean of a class should be the same as its spring NCE mean; the correlation coefficient between fall and spring
class means should be one. In this study, the correlation coefficient
for class means fall to spring across grades on the reading test was
approximately 0.91; it was approximately 0.84 on the math test. The
standard error of the estimate across grades in predicting spring NCE
class mean from fall NCE class mean was approximately 4 NCEs (error
variance of approximately 16) on the reading test, and 6 NCEs (error
variance of approximately 36) on the math test. The 95% confidence
range across grades for predicting spring NCE class mean from fall NCE
class mean was approximately plus or minus 8 NCEs (range of 16) on the
reading test and plus or minus 11.5 NCEs (range of 23) on the math
test.

The no-treatment expectation under Model A is based on the assump-
tion that student groups maintain a constant percentile position. Data
from this study indicate a considerable amount of variance can be
expected in the degree to which class means maintain a constant percen-
tile rank fall to spring. In addition, this variability occurs across
the range of the NCE scale and is not concentrated at either end, as
can be seen in the regression plots.

Do Expectations Vary By Grade Levels And/Or Subject Areas?

The degree to which class means maintain a constant percentile
rank fall to spring was examined for grades 2, 3, 4, and 5 on MAT
reading and math tests. A two-way analysis of variance (one between and one within subject variable design) of class deviation scores (spring class mean minus fall class mean) was conducted to determine whether the average degree to which classes maintained their position varied by grade and test subject area. The F ratio was statistically significant at the .05 level for mean differences between reading and math class deviation scores. F ratios were not significant for either mean differences between grade levels or interaction effects between grade levels and test subject areas.

The mean of the deviation score distributions across grades is 1.72 NCEs (Table 18). The means of the deviation score distributions for each grade are all positive and approximately the same size. The standard deviations of the deviation score distributions are approximately the same size for each grade. The size of the standard deviations varies somewhat, as do the sizes of the means for these distributions, but no distinct between-grade-level pattern emerges in either the means or the standard deviations of class deviation scores.

The mean of class deviation scores on the math test is larger than the mean of class deviation scores on the reading test at each grade level (Table 18). Similarly, the same pattern exists with respect to the grade by grade reading and math standard deviations. That is, the standard deviation of class deviation scores is larger on the math test than the reading test at each grade level.

The overall mean of class mean spring minus fall deviation scores across grades on the reading test is 1.45 NCEs. The overall mean across grades on the math test is 1.97 NCEs. The average mean dif-
ference between the two means is approximately 0.5 NCE. The overall standard deviation of class mean spring minus fall deviation scores across grades on the reading test is 4.17 NCEs. The overall standard deviation across grades on the math test is 5.92 NCEs. The difference between the two standard deviations is approximately 2.0 NCEs.

Therefore, data in this study suggest that the degree to which class means maintain their position fall to spring is likely to be more variable in math than it is in reading, and changes in class mean position fall to spring are likely to be slightly larger in math than in reading. These two patterns held across the four grade levels examined in this study. No distinct between-grade-level patterns were apparent.

How Stable Are Class Means In Relation To Each Other?

When TIERs procedures are followed to calculate group mean NCE scores, the underlying referent points for making the calculations are test norms for individuals. Since percentile norms for individuals are the underlying referent points, group status is being described in terms of scale points for individuals. In calculating the amount of change in a group mean from one point in time to another, then, one is describing how much the mean has moved on the scale for individuals from one point in time to another.

Since distributions for group scores are not the same as distributions for individuals, a group's mean score on the percentile scale for individuals could be considerably different from its position in a distribution of group scores for that unit. Similarly, changes in
group status from one point in time to another could vary considerably, depending on whether individual or group distributions are used as the underlying referent points to describe change.

Besides examining the degree to which class means remain constant, fall to spring, using TIERS procedures and test norms for individuals, procedures were also used in this study to investigate how stable a position class means hold, fall to spring, in direct relation to each other. This was done by obtaining empirical fall and spring percentile distributions of MAT scale score class means of the sample. NCE's values were then assigned in accordance with the class' mean scale score percentile rank among classes. The difference in a class' fall and spring values was used as an indicator of how much the class' position had shifted in relation to the other classes in the sample.

The standard deviation of deviation scores for the degree to which class position changes in relation to other classes in the sample was twice that of standard deviations of deviation scores obtained when TIERS procedures were followed and test norms for individuals were used as the underlying referent points. The standard deviation of deviation scores across grades in reading for the former was approximately 9.22 NCE's compared to 4.17 NCEs for the latter. The corresponding respective estimates in math were 11.96 NCE's compared to 5.92 NCEs. Based on these results, the degree to which class means maintain their position fall to spring in relation to each other can be expected to be considerably more variable when related to each other than the degree to which class means maintain their position when related to test norms for individuals. Likewise, it follows
that if one were to argue that group norms should be used as the basis for interpreting results with respect to groups, maintenance of the equipercentile assumption as a no-treatment expectation would be even less tenable than it is under the present procedures.

**How Much Change Can Be Expected To Occur In Class Means?**

The no-treatment expectation under Model A of TIERs is that, without special treatment, student groups are expected to maintain a constant percentile position in relation to national norms. The degree to which class means maintain a constant position fall to spring in the norm group was examined in this study.

The degree to which class means maintained a constant position fall to spring was examined for grades 2-5 on MAT reading and math tests. TIERs procedures were used to calculate NCE class means, and class spring minus fall NCE deviation scores were obtained. The mean deviation score across grades on the reading test was approximately 1.5 NCEs. The standard deviation of deviation scores across grades on the reading test was approximately 4 NCEs. On the math test, the mean deviation score across grades was approximately 2 NCEs, and standard deviation was approximately 6 NCEs. Based on these estimates and assuming normality for deviation score distributions, one could expect the following fall to spring NCE changes to occur in class reading means:

1. NCE changes between -2.5 to +5.5 approximately 68% of the time.
2. NCE changes between -6.3 to +9.3 approximately 95% of the time.

In math, one could expect the following fall to spring NCE changes to
occur in class means:

1. NCE changes between $-4$ to $+8$ approximately 68% of the time.
2. NCE changes between $-9.8$ to $+13.8$ approximately 95% of the time.

How Sound Is The Equipercentile Assumption
As A No-Treatment Expectation?

This study examines the appropriateness of the equipercentile assumption in relation to class means. Data were obtained to assess whether classes maintain a constant percentile position. That is, whether the growth expectation is for classes to maintain a constant percentile position in relation to national norms. Study results indicate classes generally do not maintain a constant percentile position. Class means were found to vary considerably from fall to spring in relation to national norms. The 95% confidence interval for changes in reading was estimated at $-6.3$ to $+9.3$ NCEs. The 95% confidence interval for changes in math was estimated at $-9.8$ to $+13.8$ NCEs. Given these results, in a subset of the norming sample the expectation that groups maintain a constant position in relation to national norms is not tenable.

The sample of class means used in this study was derived from a national standardization group. The classes represent a wide range of educational conditions, student characteristics, teacher characteristics, etc. The basic question of interest in this study was in determining whether such differences were likely, overall, to have an equalizing effect. That is, even though classes are known to vary on a wide range of variables, the overall effect of these differences
is that classes can basically be expected to maintain a constant percentile position. This was not found to be the case at all. The degree to which class position in general can be expected to change in relation to national norms is not insignificant but sizable. Additional empirical work should be done regarding student group growth expectations. Any number of variables could affect student group growth expectations. Based on the results of this study, however, there is no reason to assume student groups maintain a constant percentile position as a normal growth expectation.

Limitations And Implications
For Further Research

This study provides estimates of the extent to which class units maintain a constant percentile position in relation to national norms. These estimates are for a particular time interval on a particular test for four grade levels in two subject areas. Varying any major design feature of this study such as grade levels, test, test subject areas, group unit, or population characteristics from which the sample was drawn, could produce significantly different estimates. The study estimates therefore apply to group units that are like those of this study and not to others.

Although statistical estimates provided in this study should not be applied to group units dissimilar to those of this study, the general implication can be drawn that just as the equipercentile assumption does not hold for the units of this study, there is little reason to suppose it would hold for other group units. While data from this study
support the argument being made, further verification of the argument and research on the topic are needed.

One variable of practical importance which should be investigated is group size. The size of Title I projects is highly variable, ranging from only a few students per grade to thousands of students per grade. Presently, Model A guidelines make no reference to treatment group size as a factor which may affect constancy of group percentile rank. At the same time, we know that means which are based on large numbers of students are likely to be more stable than means which are based on only a few students. Consequently, research should be conducted to determine the extent to which varying treatment group size affects the validity of the equipercen tile assumption as a no-treatment expectation.

An even more important area which needs investigation is the extent to which Title I project students are comparable to students in the norm group with whom they are compared. Under Model A, the assumption is made that Title I project students are similar to students in the norm group who scored at the same percentile level as Title I project students. How comparable the two groups are is questionable, as, by definition, the samples used to develop national norms on standardized tests are not like the specially defined subpopulation to which Title I compensatory education programs are directed. Because of student population differences between the Title I student subpopulation and the cross-section of students comprising the national norm, Title I students may progress at either a faster or slower rate than expectations derived from the national norm. To the extent that they
do, Model A will provide a biased estimate of project impact. How significant that bias may be, and the direction of that bias both bear investigation.
BIBLIOGRAPHY


Faddis, B. J., & Estes, G. D. Fall-to-spring vs. fall-to-fall evaluations of a large Title I program with a comparison group design. Paper presented at the annual meeting of the American Educational Research Association, Toronto, April 1978.


