To the Graduate Council:

I am submitting herewith a dissertation written by David L. Cook entitled "An Objective Component for the Evaluation of Teaching." I have examined the final copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Education, with a major in Educational Administration and Supervision.

Gerald C. Ubben, Major Professor

We have read this dissertation
and recommend acceptance:

Accepted for the Council:

The Graduate School

AN OBJECTIVE COMPONENT FOR THE EVALUATION OF TEACHING

A Dissertation

Presented for the

Doctor of Education

Degree

The University of Tennessee, Knoxville

David L. Cook

March 1985

Copyright © David L. Cook, 1985

All Rights Reserved

DEDICATION

TO THE STUDENTS IN THE TENNESSEE PUBLIC SCHOOL SYSTEM

ACKNOWLEDGMENTS

I would like to express my graditude to the members of my doctoral committee: Dr. Gerald Ubben, Dr. Michael Rush, Dr. Frederick Venditti, and Dr. Michael Hannum. I am especially grateful to Dr. Ubben, chairman, for continued support and encouragement throughout this research project.

Also, I am indebted to Dr. William Sanders who provided valuable advice and direction in formulating the statistical design of this study.

Finally, I wish to acknowledge the individual help of Mrs. Melba Wilkins in the preparation of this manuscript.

ABSTRACT

The problem of this study was the need for a more objective method of evaluating classroom teachers. Specifically, the basic problem centered around the question as to whether standardized achievement tests in education could be used for teacher evaluational purposes. If yes, then what procedure or steps should be developed and followed to incorporate student achievement test data into the evaluational process? The purpose of this research was to determine how best standardized achievement tests can be utilized in assessing a classroom teacher's effectiveness.

The study used achievement test data for fifth grade students within the Blount County School System (Tennessee) for a two year time period, 1982-83 and 1983-84. Fifth grade achievement test data in the academic areas of reading, math and language arts, third grade achievement data from the same areas, and fourth grade school ability data for students enrolled in the fifth grade during the two year period were utilized in this study. Approximately 1,200 students and 50 classroom teachers in each academic area were subjects for this study.

It is possible and quite feasible to objectively evaluate teachers from student achievement data was the primary conclusion of this study. Adjusted mean scale scores were used to rank classroom teachers based upon these scores. An analysis of covariance was used to initially adjust fifth grade scores to a constant ability level. The analysis used third grade data and fourth grade school ability data

as independent variables. The scores were then subjected to the Henderson Best Linear Unbiased Prediction (BLUP) procedure to correct for the regression to the mean problem. Rankings of teacher effectiveness were developed for both single year periods and for a composite, two year time period.

The subjectivity of past and current teacher evaluation methodology has been attacked by many individuals both within and outside the field of education. This research project directly attacked this problem and has resulted in the creation of a design to utilize standardized achievement data to objectively evaluate classroom teaching. This design could help to bring the manpower evaluation practice in education into closer alignment with the methodology practiced by the business segment of our society.

TABLE OF CONTENTS

CHAPTER		PAGE
I.	INTRODUCTION	1
	Statement of the Problem	3
	Purpose of the Study	6
	Importance of the Study	7
	Assumptions	8
	Delimitations	9
	Definitions	9
II.	A REVIEW OF THE LITERATURE	12
	Introduction	12
	Criticism of Current Educational Personnel Evaluation	12
	Focus of Research on Teaching Evaluation	14
	Criticism of the Use of Standardized Achievement Data in Personnel Evaluation in Education	17
	Research Efforts Involving the Use of Standardized Test Data in Evaluation of a Teacher's Effectiveness and the Statistics Involved	30
	Merit Pay Plans Built Upon the Concept of Student Gain as Measured by Standardized Tests	51
	Summary	53
III.	INTRODUCTION	57
	The Population	57

CHAPTER PAGE

III. (Continued)

Data Collection Procedures 57

Analysis of the Data 60

Statistics of the Study 62

IV. INTRODUCTION 72

Presentation and Analysis 72

Summary 158

V. SUMMARY, CONCLUSIONS, DISCUSSION, AND RECOMMENDATIONS . . . 161

Summary 161

Conclusions and Discussion 164

Recommendations 181

BIBLIOGRAPHY 186

APPENDIX 193

VITA 202

LIST OF TABLES

TABLE		PAGE
I.	Multiple Regression Adjustments	37
II.	Results of An Analysis of Covariance in Which the Final Examination Scores of Students in Nine Discussion Sections are Adjusted for Selection Differences in Scholastic Aptitude	42
III.	Frequency Table of Scale Scores: Reading, Third Grade, 1981	90
IV.	Frequency Table of Scale Scores: Reading, Third Grade, 1983	92
V.	Frequency Table of Scale Scores: Math, Third Grade, 1981	94
VI.	Frequency Table of Scale Scores: Math, Fifth Grade, 1983	96
VII.	Frequency Table of Scale Scores: Language Arts, Fifth Grade, 1983	98
VIII.	Frequency Table of Scale Scores: School Ability, Fourth Grade, 1982	100
IX.	Frequency Table of Scale Scores: Language Arts, Third Grade, 1981	102
X.	Frequency Table of Scale Scores: Reading, Third Grade, 1982	104

TABLE PAGE

XI. Frequency Table of Scale Scores: Reading, Fifth Grade, 1984 106

XII. Frequency Table of Scale Scores: Math, Third Grade, 1982 109

XIII. Frequency Table of Scale Scores: Math, Fifth Grade, 1984 111

XIV. Frequency Table of Scale Scores: Language Arts, Third Grade, 1982 114

XV. Frequency Table of Scale Scores: Language Arts, Fifth Grade, 1984 115

XVI. Frequency Table of Scale Scores: School Ability, Fourth Grade, 1983 117

XVII. Reading, Fifth Grade, Unadjusted and Adjusted Mean Scores: 1983 121

XVIII. Math, Fifth Grade, Unadjusted and Adjusted Mean Scores: 1983 122

XIX. Language Arts, Fifth Grade, Unadjusted and Adjusted Mean Scores: 1983 123

XX. Reading, Fifth Grade, Unadjusted and Adjusted Mean Scores: 1984 124

XXI. Math, Fifth Grade, Unadjusted and Adjusted Mean Scores: 1984 126

XXII. Language Arts, Fifth Grade, Unadjusted and Adjusted Mean Scores: 1984 127

XXIII. Teacher Effect After Adjustment 179

TABLE PAGE

XXIV. BLUP (Best Linear Unbiased Prediction) of Teacher Effects* 131

XXV. Single Year Teacher Rank: Reading, 1983, Fifth Grade . . . 137

XXVI. Single Year Teacher Rank: Math, 1983, Fifth Grade 139

XXVII. Single Year Teacher Rank: Language Arts, 1983, Fifth Grade 140

XXVIII. Single Year Teacher Rank: Reading, 1984, Fifth Grade . . . 141

XXIX. Single Year Teacher Rank: Language Arts, 1984, Fifth Grade 142

XXX. Single Year Teacher Rank: Math, 1984, Fifth Grade 143

XXXI. Reading Composite Based Upon Two Year Achievement Data: 1983-1984 144

XXXII. Math Composite Based Upon Two Year Achievement Data: 1983-1984 146

XXIII. Language Arts Composite Based Upon Two Year Achievement Data: 1983-1984 147

XXXIV. Ranking Comparison: Principal Evaluation and Achievement Ranking 154

XXXV. Ranking Comparison: Supervisor Evaluation and Achievement Ranking 159

LIST OF FIGURES

FIGURE		PAGE
1.	Adjustment Through Multiple Regression Plot	36
2.	Teachers and Students in the Study	58
3.	Mean Adjustment--BLUP Estimates	66
4.	Bar Chart of Distribution of Scale Scores: Reading, Third Grade, 1981	75
5.	Bar Chart of Distribution of Scale Scores: Math, Third Grade, 1981	76
6.	Bar Chart of Distribution of Scale Scores: Language Arts, Third Grade, 1981	77
7.	Bar Chart of Distribution of Scale Scores: School Ability, Fourth Grade, 1982	78
8.	Bar Chart of Distribution of Scale Scores: Reading, Fifth Grade, 1983	79
9.	Bar Chart of Distribution of Scale Scores: Math, Fifth Grade, 1983	80
10.	Bar Chart of Distribution of Scale Scores: Language Arts, Fifth Grade, 1983	81
11.	Bar Chart of Distribution of Scale Scores: Reading, Third Grade, 1982	82
12.	Bar Chart of Distribution of Scale Scores: Reading, Fifth Grade, 1984	83

FIGURE PAGE

13. Bar Chart of Distribution of Scale Scores: Math, Third Grade, 1982 84

14. Bar Chart of Distribution of Scale Scores: Math, Fifth Grade, 1984 85

15. Bar Chart of Distribution of Scale Scores: Language Arts, Fifth Grade, 1984 86

16. Bar Chart of Distribution of Scale Scores: School Ability, Fourth Grade, 1983 87

17. Bar Chart of Distribution of Scale Scores: Language Arts, Third Grade, 1982 88

18. Relationship Between Teacher Effects Estimated in Consecutive Years, Reading, 1983 133

19. Relationship Between Teacher Effects Estimated in Consecutive Years, Math, 1983 134

20. Relationship Between Teacher Effects Estimated in Consecutive Years, Language, 1983 135

21. Reading Composite Ranking Graphic Presentation Based Upon Two Year Achievement Data, 1983-1984. 150

22. Math Composite Ranking Graphic Presentation Based Upon Two Year Achievement Data, 1983-1984 151

23. Language Composite Ranking Graphic Presentation Based Upon Two Year Achievement Data, 1983-1984 152

24. Reading Ranking Comparison (Principal Expected Ranking-Achievement Ranking) 155

FIGURE PAGE

25. Math Ranking Comparison (Principal Expected Ranking-Achievement Ranking) 156

26. Language Arts Ranking Comparison (Principal Expected Ranking-Achievement Ranking) 157

CHAPTER I

INTRODUCTION

Personnel evaluation whether by supervisors, administrators or students, can be a useful asset to the individual teacher and to the educational organization. Classroom teachers should be able to use evaluation data to analyze their teaching methods/procedures and alter those to enhance their students' learning. The educational organization can utilize these data in many ways. These data should most certainly be used in evaluating the organization's selection and training programs. Ideas for future staff development programs could also be ascertained from evaluation of personnel. If the organization implements a differentiated compensation plan (merit pay) effective evaluation would be essential. Personnel evaluation can be of benefit to both the individual and the organization. The problem with evaluation in education seems to be that no instrument or procedure has been compiled that can accurately or reliably and efficiently evaluate every teacher's effectiveness. In the area of teacher evaluation, we seem not to know what must be evaluated. It is quite obvious to most educators today that new criteria for evaluating teacher effectiveness are needed.

We have come to a point in education where administrators, students, parents, and non-parent taxpayers are calling for results and performance accountability. A review of current literature reveals multiple references regarding educational accountability and stricter measurement of teacher effectiveness. The recent A Nation at Risk

report issued by a Presidential Task Force can be seen as evidence of public concern over the quality of public education. In Tennessee, "The Comprehensive Education Reform Act of 1984" is evidence of this concern. Both address the need for increasing teaching effectiveness and the corresponding need for more effective personnel evaluation in education. The National Commission on Excellence in Education for example recommended in A Nation at Risk that:

> . . . Salaries for the teaching profession should be increased and should be professionally competitive, market-sensitive, and performance-based. Salary, promotion, tenure and retention decisions should be tied to an effective evaluation system that includes peer review so that superior teachers can be rewarded, average ones encouraged, and poor ones either improved or terminated.[1]

Teacher accountability is a broad concept, which has for too long been considered very narrowly. Teacher evaluation has often consisted of a one-shot, subjective, administration evaluation. This evaluation usually consists of a brief check-list of items only tangentially related to actual teaching effectiveness. Reliability and validity studies usually have not been available to substantiate such measures. This state of affairs has for the past few years come under heavy attack from many sources: parents, the general public, the press, researchers, and many state legislative bodies. The problem of measuring teaching effectiveness is being intensely debated by both educators and non-educators.

---

[1]The National Commission on Excellence in Education, A Nation at Risk: The Imperative for Educational Reform (Washington, D.C.: U. S. Department of Education, 1983), p. 30.

I. STATEMENT OF THE PROBLEM

Measuring a teacher's effectiveness is a broad and complex task. Frequently many school boards and/or state legislatures have enacted legislation or policies involving teacher accountability systems without realizing the complexity involved. The Stull Act of California is an example of such legislation.[2] Staff evaluation is a major component in the maintenance and advancement of any organization. The field of education is not immune to this rule. However, in practice evaluation is often entirely neglected or receives only minute attention. Evaluation does not operate in a vacuum. Evaluation should serve as the basis for personnel development programs and as the basis for other personnel decisions such as promotion and compensation. The Career Ladder Program for teachers, recently passed in the State of Tennessee as a component of the Tennessee Comprehensive Education Reform Act of 1984, does attempt to incorporate these concepts. Many leading experts in the business and industrial fields believe that appropriate compensation plans must be based upon job performance evaluation. The need for an adequate staff evaluation plan is clear.

The key word here is adequate, or better yet, optimal evaluation. The current problem is not that the educational organization is not evaluating its personnel but that it is not evaluating/assessing the relevant factors. In education the ultimate product is student development in the cognitive domain. And yet, this

---

[2]N. L. Gage, ed., Mandated Evaluation of Educators (Washington, D.C.: Capitol Publications, 1973), pp. 111-132.

important factor is often neglected in personnel evaluation. This state of affairs is slowly changing as more states move toward enacting legislation mandating more strict accountability systems in education. Many of these systems are connected to merit-pay compensation plans. Crucial to any teacher accountability system is the measurement of student achievement as a result of a classroom teacher's performance.

Measurement of student achievement for personnel evaluation purposes has become a major area of discussion concerning educational accountability systems. From a review of the literature, it would appear that very few educational experts can agree on this measure. Norm-referenced measures and criterion referenced measures have frequently been used to measure student achievement. Extensive criticism has been levied upon norm-referenced measures (standardized achievement tests). Popham identified two major deficiencies which make standardized achievement tests for personnel evaluation unsuitable.[3] First, such tests are designed so generally that there is little congruence between the emphasis of the test and that of a particular teacher. Second, norm-referenced tests are designed to produce considerable score variance which may lead to the elimination of test items that would be sensitive to detecting the instructional impact of a highly effective teacher. Popham believes that standardized achievement tests would penalize the high quality teacher while likely not identifying the teacher needing improvement. In contrast, criterion-referenced tests are skill specific and provide

---

[3]Ibid., p. 115.

more individually relevant data. Criterion-referenced tests indicate a pupil's status in relation to predetermined criterions. The area is usually linked to instructional objectives of a particular teacher or classroom. Norm-referenced tests reflect national objectives for a given grade level. Because of this differentiation, many educational experts believe that criterion-referenced tests are more appropriate when student performance is used as an index of teacher effectiveness. The argument presented is that a student's success in obtaining the educational objectives of his/her classroom teacher or local system should be more important to the evaluator than his/her score on a measure of more remote, national objectives. With such criticism evident, further research should provide valuable insight into the problem of using achievement test data in the evaluation process of educational personnel. The debate continues on many fronts as to the value of utilizing standardized achievement tests for evaluating a classroom teacher's effectiveness.

The basic problem then centers around the question as to whether standardized achievement tests in education can be used for teacher evaluational purposes. If yes, then what procedure or steps should be developed and followed to incorporate these test data into the evaluational process? Note that the researcher is not advocating that such tests become the sole evaluation tool in education, but that their potential as a valuable component of a broader, more complex evaluation system warrants further investigation. The traditional availability of standardized testing measures in most educational settings further confirms their potential usefulness for this area of research.

II. PURPOSE OF THE STUDY

A major problem lies in the area of valid measures of student achievement for personnel evaluation in the field of education. The purpose of this research was to determine how best standardized achievement tests can be utilized in assessing a classroom teacher's effectiveness. Secondly, the research would help establish procedures and/or a design for appropriately using student achievement data in staff evaluation within the educational setting. A final purpose was to examine the utility of such tests in the total professional evaluation framework within the educational organization.

The study attempted to answer the following questions:

1. Are there significant differences in teacher effectiveness?

2. Can standardized achievement tests be used appropriately and accurately to rank classroom teachers in terms of classroom (teaching) effectiveness?

3. What are the appropriate procedures or designs for using student achievement data in teacher evaluation?

4. How do student achievement test rankings correlate with principal and supervisory ratings?

5. Is yearly achievement testing a requirement if one wishes to evaluate teaching effectiveness through the use of achievement test score data?

III. IMPORTANCE OF THE STUDY

Appropriately and accurately evaluating a teacher's effectiveness is a priority if we want to provide the best education possible for students. It is also a pre-requisite for establishing a valid personalized, work-oriented compensation-reward system. If the compensation-reward system is closely related to job performance, employee performance can be improved.

> The operant conditioning theory of Skinner, the expectancy theory of Vroom, and the path-goal theory of Lawler were all shown to lead to the conclusion that there must be a critical linkage between performance and reward if a reward system is to be effective in motivating human performance.[4]

The educational organization has neglected this basic business concept for too long.

Personnel evaluation in education needs to become more objective. Subjective measures such as administrative checklists, peer ratings, and student checklists have dominated teacher evaluation. The use of standardized achievement tests as an objective measure of a teacher's effectiveness has great potential for school administrators working to more accurately determine individual teacher effectiveness. Such information will be valuable for making personnel decisions (merit pay, promotion, and dismissal) and also for planning necessary growth opportunities for employees (in-service education). Obtaining more

---

[4] Lawrence A. Klatt, Robert G. Murdick, and Fred E. Schuster, Human Resources Management: A Behavioral Systems Approach (Homewood, Illinois: Richard D. Irwin, Inc., 1978), p. 215.

objective data in teacher evaluation is necessary to enhance the validity of such evaluation of personnel in education.

The purpose of this research was to show that standardized achievement test data, usually readily available to school administrators, could be used to document objectively a teacher's effectiveness. Secondarily, how best to use this data in the evaluation process was explored.

IV. ASSUMPTIONS

Assumptions pertinent to the study include the following:

1. A more objective measure of a teacher's effectiveness is needed in evaluation of teaching and can be derived from standardized achievement test data.

2. Standardized achievement tests, if selected to match curriculum goals and objectives, are a good measure of student academic growth.

3. Student academic growth as measured by standardized achievement tests is a result of teacher instructional effectiveness.

4. There is a need to identify a set of procedures or a design for utilizing standardized achievement data in evaluating teacher effectiveness.

5. Standardized achievement test data can become an objective component of a broader form of teacher evaluation and thus a valuable addition to principal and supervisory subjective evaluation.

6. Present literature does not include sufficient research on the use of achievement test data for personnel evaluation purposes.

V. DELIMITATIONS

The study was limited to faculty members and students at the fifth grade level in the Blount County, Tennessee, School System for the 1982-83 and 1983-84 school years.

VI. DEFINITIONS

Adjusted Score. A score that has been adjusted through some statistical analysis, such as the analysis of covariance, to equate for initial differences on some variable. Such treatment is undertaken to assure the score is a result of treatment.

Analysis of Covariance. A control technique used both in studies in which already-formed, not necessarily equal, groups are involved and in experimental studies in which either existing groups or randomly formed groups are involved. The analysis of covariance is a basic method for controlling for pre-test differences in a pre-test/post-test design.

Basic Skills First. An educational program being implemented in the State of Tennessee. The program identifies basic skills in the area of math and reading at each grade level. Mastery tests are available to test the student on each identified skill (criterion mastery tests).

BLUP (Best Linear Unbiased Prediction). A statistical procedure developed by C. R. Henderson that controls the problem of the regression toward the mean. It does this by "pulling" the prediction closer to the mean as a function of distance, ratio of the variances and quantity of information available.

$$BLUP = M + (X - M) \frac{\sigma^2 \text{ teachers}}{\sigma^2 \text{ teachers} + \frac{\sigma^2 \text{ student}}{N \text{ students}}}$$

Criterion Tests. A criterion-referenced test measures whether a learner has mastered a particular type of intended behavior (Robert Glasser).[5]

Henderson's Mixed Model Solution (MMS). A statistical procedure developed by C. R. Henderson for computing BLUE (Best Linear Unbiased Estimates) of the estimable linear functions of the fixed elements of a model, and for computing best linear unbiased predictions of the random elements of a model. The procedure is used extensively in the area of animal breeding applications. Used in the area of measuring teaching effectiveness through achievement test gain, Henderson's MMS can control for the "regression to the mean" problem and to adjust for known differences among students in order to alleviate the problem of nonrandom assignment of students to teachers. The statistical procedure allows for the inclusion of both continuous and class fixed variables into the analysis concurrently with one or more random class variables.

Professional Career Ladder System. A system recently created and passed by the Tennessee Legislature (1984). Briefly, the system creates a series of career steps with specific teacher requirements and responsibilities outlined. Differential compensation is built into the system as well as a more professional objective evaluation system.

---

[5]Gage, Ibid., p. 116.

Rankings. The order according to some statistical characteristic (as score on a test).

Raw Score. The actual score obtained on a standardized achievement test. The score has not been adjusted in any manner. Such scores suffer from unreliability and susceptibility to distortion by the regression effect.

Standardized Achievement Test. Achievement tests measure the current status of individuals with respect to proficiency in given areas of knowledge or skill. Standardized achievement tests are carefully developed to include measurement of objectives common to many school systems. They measure knowledge of facts, concepts, and principles.

Supervisor. An individual within a school system who has supervisory, but not teaching, duties. This individual is not assigned to a particular building but works throughout the school system.

Teacher Effectiveness. A measure of how successful a classroom teacher is in teaching his/her students (a subject, curriculum, set of basic skills). Effectiveness, for the purpose of this study, will be limited to the demonstrated capacity of a teacher to effect change in behaviors measured by standard achievement tests.

CHAPTER II

A REVIEW OF THE LITERATURE

I. INTRODUCTION

This review of the literature was organized into the following areas: criticism of current educational personnel evaluation, focus of research on teaching evaluation, criticism of using standardized test data in evaluation in education, review of research efforts involving the use of standardized test data in teacher evaluation and the statistics involved, and merit pay plans built upon concept of student gain as measured by standardized tests.

II. CRITICISM OF CURRENT EDUCATIONAL PERSONNEL EVALUATION

Staff evaluation is an important component in the maintenance and advancement of any organization. The field of education, unfortunately, has often ignored this rule. Evaluation should serve as the basis for personnel development programs and as the basis for other personnel decisions such as promotions and compensation. Unfortunately this latter point is not practiced very extensively at this time. Leading experts in the field of personnel management are in agreement that effective compensation plans must be based upon job performance evaluation.[6] The need for an adequate staff evaluation plan in education is real.

---

[6]Lawrence A. Klatt, Robert G. Murdick, and Robert E. Schuster, Human Resources Management: A Behavioral Systems Approach (Homewood, Illinois: 1978, Richard D. Irwin, Inc., 1978), p. 238.

A wide variety of evaluational methods or techniques can be found in use in the field of education. Student ratings, self-ratings, administrative and peer ratings, and classroom observations are discussed in much of the current literature.[7] Each of these methods have come under criticism, due primarily to its subjectivity.

> The difficulty with ratings of teacher effectiveness characteristically supplied by administrators, but also obtainable from students and peers, is that different raters have different notions regarding what it is that constitutes good teaching.[8]

The same teacher rated high by one individual for flexibility in planning an instructional program may be rated low by another individual because of poor discipline. This issue of subjectivity also comes into play in regard to self-ratings, administrative and peer ratings and to some extent with classroom observation.

> Subjectivity in the evaluation of teacher effectiveness, long deplored continues to frustrate those who are evaluated, those who evaluate, and those who depend upon evaluation to strengthen programs of teacher selection, teacher training, and in-service growth.[9]

Hawkins and Stoops (1966) conducted a study involving a comparison of three methods of evaluation: the process of salary determination

---

[7]John D. McNeil and James W. Popham, "The Assessment of Teacher Competence," Second Handbook of Research on Teaching (New York: Rand McNally, 1973), Chapter 7, pp. 231-233.

[8]W. James Popham, "Alternative Teacher Assessment Strategies," University of California, Los Angeles and Instructional Objectives Exchange, 1973, p. 2.

[9]Edward E. Hawkins and Emery Stoops, "Objective and Subjective Identification of Outstanding Elementary Teachers," The Journal of Educational Research, LIX, (April 1966), p. 344.

(training and experience), formal evaluation (principal judgements), and informal non-administrative evaluation (selections by peer teachers, non-teaching staff members, and school-associated parents).[10] The formal evaluation process failed to exhibit any superiority over informal subjective evaluation. The researchers concluded that the formal evaluation process fails to provide a ranking of teachers sufficient to meet the needs of a merit pay system. The limited number of schools in this study (10 of 1,017 elementary schools in Los Angeles County) would not warrant a generalization of the researchers' conclusions.[11] However, it is included in this review because of relevance to the problem of subjectivity in teacher evaluation. The problem of subjectivity in evaluation of educational personnel is discussed frequently in the educational literature bearing on this topic.

III. FOCUS OF RESEARCH ON TEACHING EVALUATION

The major focus of most educational research for the past 20 years has been on teacher behavior and traits. The emphasis has been upon what behaviors or characteristics make a teacher effective. There is extensive research in the literature concerning this process-based approach. Under this branch of research one could find topics such as directiveness, reinforcement and behavior modification, teacher use of praise, and so forth. The focus of such research is upon behaviors

---

[10] Ibid., pp. 344-346.

[11] Ibid., p. 344.

and actions of the teacher. The reader interested in a further review of this approach to educational teacher research should consult The Study of Teaching by Dunkin[12] or the Second Handbook on Research on Teaching[13] by McNeil and Popham. Both sources provide a good discussion of the research in teaching over the past few years. An extensive bibliography is provided in each for further reference.

The product based research of teaching is not an extensive as the process based. Product based research or evaluation focuses upon what the teacher produces or his/her effectiveness. The student or student's performance becomes the focal point. A clear distinction can be seen in the following definitions provided by Pembroke and Goedert:

> Process-oriented--where the focus is aspects of teaching behavior considered to indicate teaching competence, such as teaching methods, classroom management, and student-teacher interaction. Product-oriented--where the focus is on the end results of teaching, evidenced by measurable or observable aspects of student growth or change in attitude or behavior that can be attributed to the effect of the teacher.[14]

The major focus of this study concerns product teacher evaluation.

A shift can be detected in the focus of teacher evaluation from process-based to product based. This shift in emphasis has been well

---

[12]Michael J. Dunkin and Bruce J. Biddle. The Study of Teaching (New York: Holt, Rinehart and Winston, Inc., 1974).

[13]John D. McNeil and W. James Popham. Second Handbook of Research on Teaching (New York: Rand McNally, 1973).

[14]Eiben Pembroke and Edmund R. Goedert, "What is the Key to Developing an Effective Teacher Evaluation System?" NASSP Bulletin, V66, N458, (December 1982), p. 30.

documented.[15] The rising accountability issue has created much of the force behind this change. Currently the nation's educational enterprise is being subjected to increasing public scrutiny. This concern for accountability in education is seen in an increased emphasis upon teacher evaluation.

> Between April 1963 and 1972, 23 states enacted legislation that authorized or mandated some type of educational accountability. By April 1973, 16 additional states had created some form of accountability legislation to be acted upon.[16]

An excellent example of this increased accountability issue upon education would be the passage of the Stull Act in California in 1971 with its mandated product teacher evaluation.[17] Marvin Alkin provides the following definition of accountability:

> Accountability is a negotiated relationship in which the participants agree in advance to accept specified rewards and costs on the basis of evaluation findings as to the attainment of specified ends.[18]

---

[15]W. James Popham., "Alternative Teacher Assessment Strategies," University of California, Los Angeles and Instructional Objectives Exchange, 1973; see also John D. McNeil and W. James Popham, Second Handbook on Teaching, New York: Rand McNally, 1973); Marvin C. Alkin and Stephen P. Klein, "Accountability Defined. Evaluating Teachers for Outcome Accountability," UCLA Evaluation Comment, VC, N3, (May 1972), pp. 1-11.

[16]G. F. Ovard, "The Practitioner's Guide to Research Teacher Effectiveness and Accountability," NASSP Bulletin, V59, (January 1975), p. 91.

[17]N. L. Gage, ed., Mandated Evaluation of Educators (Washington, D.C.: Capitol Publications, 1973), pp. 111-132.

[18]Marvin C. Alkin and Stephen P. Klein, "Accountability Defined. Evaluating Teachers for Outcome Accountability,." UCEA Evaluation Comment, V3, N3 (May 1972), p. 2.

Alkin further identifies three accountability types: goal accountability, program accountability, and outcome accountability.[19] He believes that all three must be in operation to have an accountability system. Different participants are involved within each type. Within goal accountability--the school board and central office administration must accept major responsibility; within program accountability--the individual school's administration accepts major responsibility; and within outcome accountability--the classroom teacher becomes the responsible figure.[20] The increasing emphasis upon accountability in education effects every level of the educational enterprise. The effect at the classroom level, Alkin's outcome accountability level, has resulted in a change in emphasis in teacher evaluation and research. Increasingly, the focus in teacher evaluation is upon the product of education or upon what the child has learned or achieved.

IV. CRITICISM OF THE USE OF STANDARDIZED ACHIEVEMENT DATA IN PERSONNEL EVALUATION IN EDUCATION

Although there is a need for more objective evaluation of teaching in education, the method of objective evaluation is a point of controversy. Objective evaluation of teachers based upon their students' gains on standardized achievement tests is one method frequently discussed. Since this research study involves using such

---

[19] Ibid., p. 3.

[20] Ibid.

data in teacher evaluation, it is felt that both sides of the issue should be reviewed and presented.

Robert and Ruth Soar, distinguished researchers and authors in the field of education, have written extensively against using achievement test data and other objective outcome measures in the evaluation of educational personnel. As a matter of fact, much of their material has been used by the National Education Association, the largest teacher organization in the United States, in its fight against such evaluation. The Soars list the following reasons or problems in using pupil outcomes for teacher evaluation:

> 1. What the pupil brings to the classroom in terms of ability, previous knowledge, home and peer influence, motivation, and other influences is clearly very powerful in determining academic standing at the end of the year.
> 2. Student achievement reflects only a small portion of the total set of objectives for which schools increasingly are being held accountable.
> 3. Taking pupil standing at year-end as an indicator of teaching effectiveness frequently does not recognize standing at the beginning of the year, in addition, the problems of adjusting for prior standing are extremely serious and rarely recognized.
> 4. The accountability system would reward the teacher who teaches to the test or who gives primary attention to those pupils who are below minimum standards but in reach of them.
> 5. Such an evaluation system would probably reward teaching behavior which promotes low cognitive level learning and penalize teaching which promotes complex learning.[21]

With the rising pressures for measuring the outcomes in education, the Soars believe that the field of education is erroneously being compared to modern industry.[22] In many industrial settings the

---

[21]Robert S. Soar and Ruth M. Soar, "Problems in Using Pupil Outcomes for Teacher Evaluation," (Washington, D.C.: National Education Association Publication, April 1975), pp. 6-17.

[22]Ibid., p. 6.

worker is compensated quantitatively according to the number of units produced. One cannot apply this procedure to teaching, however, because a number of problems emerge which are often overlooked. The worker in industry usually receives parts or other raw materials which are identical and will follow a set of standard procedures in assembly. The finished products should be identical and of similar quality and is easily evaluated. This is where the comparison in education falls apart, however. When students come into a classroom, they do so differing in abilities, backgrounds, achievement levels, socio-economic status and in many other ways. The classroom teacher must approach each child as an individual. As a matter of fact, the whole concept of individualization in education is built upon the recognition of these unique student characteristics. The identification of key "outside" factors that influence a student's performance in the classroom is important. The Soars maintain that what goes on in the classroom is not the only and probably not the major influence on the student's growth in achievement.

Entry Level Ability

The fact that differences in pupils before they enter a classroom have major influence on achievement has been documented by research. Among the outside influences, the Soars present: entry ability, socioeconomic status, family attitude and support for learning in the home.[23] Soar and Soar (1973) completed a study of

---

[23] Ibid.

the relationship between fall IQ and spring achievement.[24] They found correlations between class averages for the two that ranged from +.85 to +.90.[25] They arrive at the conclusion that as much as 80 percent of the difference in class averages for student achievement annually can be accounted for by student characteristics which existed before the child entered the classroom and are completely out of teacher's control.[26] The reader interested in gaining a deeper insight into research and theories emphasizing this genetic component in determining the cognitive attainment of students should consult the following: Jensen (1969, 1979),[27] Banfield (1974),[28] Eysenck (1971),[29] Herrnstein (1973), [30] and Wilson (1975).[31] The major contention of these researchers is that environment (teachers are part of this environment) has little to do with cognitive abilities such as IQ and achievement.

---

[24]Ibid., p. 7.

[25]Ibid.

[26]Ibid.

[27]Arthur R. Jensen, "How Much Can We Boost IQ and Scholastic Achievement?" Harvard Educational Review (Winter 1969), pp. 1-123.

[28]Edward C. Banfield, The Unheavenly City, 2d., ed., (Boston: Little-Brown, 1974).

[29]Hans J. Eysenck, The IQ Argument: Race, Intelligence and Education (New York: Library Press, 1971).

[30]Richard J. Herrnstein, IQ in the Meritocracy (Boston: Little-Brown, 1973).

[31]Edward O. Wilson, Sociolbiology (Cambridge, Mass.: University Press, 1975).

Socioeconomic Status

The second major outside influence in pupil achievement discussed in the literature is socioeconomic status. The issue is discussed by the Soars in their work. This influence was initially presented by Coleman in 1966.[32] The Coleman Report stated that schools (including the classroom teacher) have little effect upon the student's academic growth. This basic idea of the influence of socioeconomic status on student achievement has been more recently studied by Mosteller and Moynihan[33] in 1972, by Jencks[34], et al., (1972, 1979) and to a lesser degree by the Rand Corporation (Averch, 1972).[35] The basic contention of these researchers is that a student's family background and socioeconomic background are the most important factors relating to student achievement. The classroom teacher plays a secondary role of importance.

Family Influence

A third major outside influence effecting a student's achievement in school involves family attitudes toward and support of learning. The research supporting this influence is not as extensive as that cited for socioeconomic influences. Garber and Ware in 1972

---

[32]James S. Coleman, et al., Equity of Educational Opportunity (Washington, D.C.: U. S. Government Printing Office, 1966).

[33]Frederick Mosteller and David P. Moynihan, On Equity of Educational Opportunity (New York: Random House, 1972).

[34]Christopher Jencks, et al., Inequality: A Reassessment of the Effect of Family and Schooling in America (New York: Basic Books, 1972).

[35]Harvey Averch, et al., How Effective is Schooling? (Santa Monica, Calif.: Rand Corporation, 1972).

found a significant relationship between a support of learning from the home and student achievement in a study involving a group of Black and Spanish American students.[36]

Peer Attitude

Peer group attitude has also been identified as another factor which affects a student's achievement in school. Research in this area is limited also, but support for this position has been supported by Anderson in 1970.[37]

The Soars along with other major researchers in the field of education are critical of using student achievement data for evaluation of teachers because of the major influence these outside factors (socio-economic status, family influence, entry level, ability level and etc.) have on the student's ultimate achievement. They believe that these crucial influences are not usually given consideration when student achievement measures are used for teacher evaluation. Such measures are viewed as being misleading in terms of attempting to measure a teacher's or school's effectiveness.

Measuring Change

The measurement of change that is involved in using student achievement as an effectiveness measure is cited by the Soars and

---

[36]M. Garber and W. B. Ware, "The Home Environment as a Predictor of School Achievement," Theory Into Practice 11 (June 1972), pp. 190-195.

[37]G. J. Anderson, "Effects of Classroom Social Climate on Individual Learning," American Educational Research Journal 7 (March 1970), pp. 135-153.

others as another major problem inhibiting the use of objective outcome measures in education. The statistical problem of the regression to the mean is cited as a major problem in measuring the change in student achievement. In terms of student achievement scores, the Soars define the regression to the mean problem as the tendency for extreme scores to "regress" or move toward the mean on administration of the second test.[38] The regression effect can be expected whenever prediction is less than perfect.[39] In educational predictions, a large proportion of total range of scores can be due to this regression effect.

Lower Cognitive Emphasis

The last two major problems, cited by the Soars, in their defense against the use of student outcomes for teacher evaluation are that such a system would stress the emphasis of lower cognitive learning and that studies of student achievement fail to distinguish between complex achievement growth and simple growth.[40] Simple learning growth is usually characterized by memory of facts/rote memory. Complex learning involves problem solving, abstracting, and generalization. A number of studies have concluded that the teaching behaviors required to teach these two types of knowledge are different

---

[38] Soar, op. cit., p. 12.

[39] F. M. Lord, "Elementary Models for Measuring Change," Problems in Measuring Change, (Edited by C. W. Harris). (Madison: University of Wisconsin Press, 1963), Chapter 2, p. 21.

[40] Soar, op. cit., p. 15.

(Solomon, Bezdek, and Rosenberg,[41] 1963; Soar[42]; Soar and Soar,[43] 1972, 1973). When teachers are evaluated on pupil achievement, probably teachers emphasizing simple learning would receive better evaluations than teachers emphasizing more complex learning.

The Soars' last major criticism of using student outcome measures for personnel evaluation is that such a system encourages the teacher to "teach to the test."[44] This is especially likely in settings where financial compensation is tied to student achievement, as found in some merit pay systems. Teachers who tend to "teach to the test" to facilitate high achievement scores of their students severely restrict the curriculum being offered to their students. Such a mode of instruction emphasizes rote-memory learning at the expense of more complexed, reasoning-type learning skills.

Popham, noted educational researcher and author, has also been very critical of using achievement test data for evaluational purposes. Unlike the Soars, Popham believes criterion-referenced tests and

---

[41]D. Solomon, W. E. Bezdek, and L. Rosenberg, Teaching Styles and Learning Chicago: Center for the Study of Liberal Education of Adults, 1963).

[42]R. S. Soar, "Optimum Teacher-Pupil Interaction for Pupil Growth," Educational Leadership Research Supplement 2 (December 1968), pp. 275-280.

[43]R. S. Soar and R. M. Soar, Classroom Behavior, Pupil Characteristics and Pupil Growth for the School Year and for the Summer (University of Florida, Gainsville: National Institute of Mental Health, Education, and Welfare, 1973).

[44]Robert S. Soar and Ruth M. Soar, "Problems in Using Pupil Outcomes for Teacher Evaluation," (Washington, D.C.: National Education Association Publication, April 1975), p. 17.

teacher performance tests do offer possibilities in the area of objective personnel evaluation in education. The reader interested in more details within these two areas should consult the Bibliography of this study for references. The scope of this study does not permit any extensive discussion of either. Popham bases his argument against norm-referenced achievement test use in teacher evaluation on the following reasons:

> 1. The descriptions of what is measured by norm-referenced tests are far too loose;
> 2. Evaluative expectations are unclear;
> 3. The necessity for norm-referenced tests to produce sufficient variance in examinee responses makes it less likely that such tests will measure pupil mastery of key knowledge and skills; and
> 4. Most norm-referenced tests do not differentiate among students chiefly on the basis of what those students have learned in school--rather their differentiations are based heavily on the nature of the semantic experience that students bring to school.[45]

Popham's position on the use of norm-referenced tests in evaluation corresponds to some degree with that of the Soars' previously discussed. He restates the adverse effects of certain parental or subcultural environments on the performance of students on such tests.[46] Popham's major argument against such test usage, however, primarily involves the creation of such tests.

---

[45]W. James Popham, "Teacher Evaluation--The Wrong Tests for the Right Job," Adapted from a symposium presentation at a jointly sponsored session during the Annual Meetings of the National Council of Measurement in Education and the American Educational Research Association, Los Angeles, Calif., April 1981, Research in Education (July 1981), pp. 1-9.

[46]Ibid., p. 3.

> The major problem with norm-referenced achievement tests, for purposes of teacher evaluation, stems from the psychometric traditions governing the creation of such tests. Central among these traditional imperatives is the necessity for the norm-referenced tests to produce sufficient variance in examinee responses so that fine-grained comparisons can be made among those examinees. These comparisons are made, of course, in relationship to the performance of the original norm group. Test items that contribute most to the production of response variance are those which are answered correctly by about half of the examinees. If too few or too many examinees answer an item correctly, then it fails to do its share in spreading out examinee performance. Such items are modified or removed form the test.[47]

Due to this method of achievement test construction teachers may find themselves in "Catch--22" dilemma.[48] Simply stated this dilemma involves the fact that the more a teacher emphasizes a topic and the more effective he/she is in presenting the topic, the better students will perform on items measuring those topics. As a result of the students performing well on test items, the probability increases that those items will be dropped from the next version of the test. Unfortunately, the more successful a teacher is in promoting student mastery of key knowledge, the less likely it will be that the tests will include items that measure these very skills and knowledge.[49] Popham contends that the very method of construction of norm-referenced

---

[47] Ibid., p. 5.

[48] W. James Popham, "Catch--22 for Teachers: The Penalty for Doing Well," Principal, V61, N5, (May 1982), p. 35.

[49] W. James Popham, "Teacher Evaluation--The Wrong tests for the Right Job," Adapted from a symposium presentation at a jointly sponsored session during the Annual Meetings of the National Council of Measurement in Education and the American Educational Research Association, Los Angeles, Calif., April 1981, Research in Education (July 1981), pp. 1-9.

tests, precludes their actual relevance to the classroom instructional program. This lack of congruence between test items and classroom goals and objectives should eliminate the use of norm-referenced tests for evaluation of teaching.

A second unique aspect in arguing against the use of norm-referenced tests for evaluating teachers related to actual test construction is that test developers often use subtle verbal variations in designing discrimination among test items. The end result is that the more verbal a child is the better he/she will perform on the test.

> Most norm-referenced achievement tests do not differentiate among students chiefly on the basis of what those students have learned in school; rather, their differentiations are based heavily on the nature of the sematic experience that students bring to school.[50]

Popham contends that norm-reference tests are designed to favor the more verbal child. He further contends that such tests base a large segment of their discriminating efficiency on factors unrelated to the effects of school-based instruction. Norm-reference tests may in reality be formed in such a matter as to evaluate a teacher's influence as minimum.

Berliner, as a result of his work at the Far West Laboratory for Educational Research and Development, is also very critical of use of standardized tests for evaluational purposes. Berliner contends that although standardized tests usually have adequate curriculum content validity and may be fairly successful in predicting future

---

[50] Ibid., p. 16.

academic success, they do not reflect what was taught in any one teacher's classroom.[51] This view of Berliner is clearly related to Popham's criticism of using norm-reference tests because of the way they are created and to conclusions of the Soars in regard to outside influences. Berliner contends that standardized achievement tests cannot be used to evaluate a teacher's effectiveness because they are not sensitive to an individual's classroom. Similar findings came out of studies conducted by M. D. Gall, again at the Far West Laboratory for Educational Research and Development.[52] Another criticism expressed by Berliner is that standardized tests do not measure multivariate outcomes. Affective responses are usually totally ignored within standardized testing.[53] Because standardized tests do not reflect an individual classroom's goals and objectives and because most standardized tests ignore the affective domain, Berliner urges caution in using them for evaluational purposes.

The last major area to be considered in this section involving criticism of the use of standardized achievement tests for teacher evaluational purposes involves the issue of stability of teacher effectiveness. Critics contend that since teacher effectiveness is unstable any attempt to evaluate such effectiveness

---

[51]David C. Berliner, "Impediments to the Study of Teacher Effectiveness," Washington, D.C.: National Institute of Education, Department of Health, Education and Welfare (1975), p. 9.

[52]M. D. Gall, "The Problems of Student Achievement in Research on Teacher Effects," San Francisco, California: Far West Laboratory for Educational Research and Development (Report A73-2), 1973.

[53]Berliner, op. cit., p. 11.

through achievement test data is futile. Brophy conducted a study in an attempt to measure the stability of teacher effectiveness. He selected two grade levels for his study from a city school system. The study included all teachers in these two grades who had taught at the same grade level for three years. Brophy found moderate consistency across subtests within years, consistency within years across sex of student and stability within subtests across time.[54] The limited sample used and the many uncontrolled factors present (absenteeism, class size, student aptitude, and so forth) warrant cautious interpretation of these findings. This note of caution is even voiced by the researcher in his closing comments:

> . . . The fact that teacher effectiveness shows great individual differences and only moderate stability from year to year in an unselected sample and the fact that yearly "class effects" were observed despite statistical controls, even in highly consistent teachers (though to a lesser degree), argue against the indiscriminate use of student gain or general achievement tests for assessing teacher accountability. Until the sources of instability become known and eliminated or controlled, this procedure is inappropriate, and certainly unfair to many teachers.[55]

Prior to this work of Brophy, a review of studies of stability in teacher effectiveness by Rosenshine (1970) was conducted. Rosenshine could locate only four long-term investigations with stability coefficients reflecting teacher consistency across time in

---

[54]Jere E. Brophy, "Stability of Teacher Effectiveness," American Educational Research Journal, Vol. 10, No. 3 (Summer 1973), pp. 247-250.

[55]Ibid., p. 247.

relation to producing student gains.[56] The stability coefficients were generally low and the four studies very difficult to compare. These studies by Rosenshire (1970) and Brophy (1972) show that a very small percentage of the variation in achievement from group to group reflects the stable influence of the classroom teacher. The Soars estimate that data from about 20 classes would be necessary to make reliable decisions about individual teachers.[57] Critics of the use of standardized achievement data for teacher evaluation, state that the collection of such a large collection of data is impractical.

IV. RESEARCH EFFORTS INVOLVING THE USE OF STANDARDIZED TEST DATA IN EVALUATION OF A TEACHER'S EFFECTIVENESS AND THE STATISTICS INVOLVED

In the past, the evaluation of a classroom teacher has in essence been held equivalent to determining the instructional competence of the teacher's ability to affect desired changes in students. Appropriate and valid decisions concerning teachers rest on many kinds of information. It is essential that among these many data that there be some consideration of those concerning teacher competencies. Often this latter type of information is overlooked in favor of subjective evaluations of the teacher which frequently are

---

[56]Barak Rosenshine, "The Stability of Teacher Effects Upon Student Achievement," Review of Educational Research, 40 (1970), p. 648.

[57]Robert S. Soar and Ruth M. Soar, "Problems in Using Pupil Outcomes for Teacher Evaluations," Washington, D.C.: National Education Association Publications (April 1975), p. 8.

concerned with the instructor's personal characteristics and instructional techniques as discussed earlier in this review. A vast number of investigators have directed their energies toward development of schemes for analyzing teaching and correlating process and product instructional variables. Experiments are now needed to determine whether the teacher's use of these various instructional variables that have been analyzed can produce desired effects in students.

The research and amount of literature concerning the product component of the evaluation of a teacher's effectiveness was found to be more limited than the other major areas of this review. This may be a result of the vast criticism involving the topic and the strong opposition from the educational community. However, there were found several very relevant studies involving the use of standardized achievement tests that provide direction for this study.

Before discussing the particulars of major studies and reports, some consideration should be directed toward how gain should be measured in using standardized tests for evaluational purposes. The controversy between the concepts of adjusted gain versus raw gain is found in much of the literature. A distinction between the terms is required to understand the controversy. Raw gain refers to the difference between a student's pre- and post-test scores and adjusted gain refers to a more complex score obtained from more detailed calculations designed to create a measure of gain which is independent of the student's initial standing and more

representative of change.[58] The statistical methods chosen to calculate these adjusted scores might include the analysis of covariance, multiple regression, and so forth. Although raw gain scores are used in some situations for measuring effectiveness, the adjusted gain score concept appears to be the preferred method. There are a number of reasons for this preference. Borich identifies two major weaknesses in using the raw gain concept.[59] Raw gain scores suffer from unreliability and distortion by the regression to the mean effect. The regression effect as explained previously in this review, is the tendency of extreme scores to move toward the mean on later assessments. The unreliability aspect of raw gain scores derives from the fact that the score inherits the unreliability from both the pre-test and post-test.[60] This aspect is frequently overlooked, and the raw gain score is viewed as the actual gain and indicative of the teacher's effectiveness. Several researchers and writers have identified this problem involved in using the gain score concept for evaluational purposes. Cronbach, in his article "How We Should Measure 'Change'--Or Should We?" condemns the use of raw scores.[61] Raw change or raw gain scores

---

[58]Gary D. Borich, The Appraisal of Teaching--Concepts and Processes (Reading, Massachusetts: Addison-Wesley Publishing Co., 1977), p. 51.

[59]Ibid.

[60]Ibid., p. 52.

[61]Lee J. Cronbach and Lita Furby, "How We Should Measure 'Change'--Or Should We?" Psychological Bulletin, Vol 74, No. 1 (1970), pp. 68-80.

obtained by subtracting pre-test scores from post-test scores lead to erroneous conclusions, primarily because such scores are systematically related to any random error of measurement.[62] Veldman and Brophy show a distinct preference for the use of the adjusted gain concept when they state that a residual gain score is better than a simple pre-test difference score as a measure of teacher influence.[63] This preference among researchers to use an adjusted gain concept in studies investigating the use of standardized tests for evaluational purposes will be seen more clearly within the discussion of the particular studies. It should also be noted that the use of the adjusted score concept as advocated by many research specialists eliminate some of the major criticisms discussed earlier within this chapter. The problem of student prior ability differences and the statistical problem of the regression to the mean are two examples.

This preference among investigators to use an adjusted gain in measuring teacher effectiveness was summarized effectively by Morsh and Wilder.

> Of the several methods used to measure pupil change, residual pupil gains (i.e., the difference) between actual gain and predicted gain is becoming more widely used as a criterion of instructor effectiveness. . . . Its main advantage is that a more adequate attempt is made to hold constant student factors other than the effort of the instructor. The chief

---

[62] Ibid., p. 68.

[63] Douale J. Veldman and Jere E. Brophy, "Measuring Teacher Effects on Pupil Achievement," Paper presented at a symposium of annual meeting of American Educational Research Association (New Orleans, Louisiana, February 25--March 1, 1973), p. 2.

> disadvantages are its dependence upon the availability of valid instruments for measuring student growth, the excessive time required to obtain necessary data, and the rather elaborate statistical assumptions and analysis involved. With all its difficulties, however, this appears to be the best criteria of instructor effectiveness.[64]

If the adjusted gain concept is accepted as the best means of measuring achievement gain, the next logical question would be how this is to be accomplished. The methods of adjusting the gain score are many and varied. The purpose for which the data is to be used often dictates the complexity and how detailed an adjustment procedure will be employed.

A review of the literature offers insight into the process. Cronbach identifies multiple regression procedure for situations where gain scores are desired.[65] He believes that this procedure is most appropriate where treatment groups are non-randomly formed. Multiple regression procedures can make use of more information than other statistical procedures often used. Klein (and Alkin)[66] and Barro[67] also advocate the regression analysis approach in calculating gain. Klein offers a very clear description of how this technique, in its simplest form, works. Klein's five step procedure is presented for reader clarification as follows:

---

[64]J. E. Morsh and E. W. Wilder, Identifying the Effective Instructor: A Review of the Quantitative Studies 1900-1952 (Chanute AFB, Ill.: Air Force Personnel and Training Center, 1954), p. 53.

[65]Cronbach, op. cit., p. 80.

[66]Alkin, op. cit., p. 8.

[67]Stephen M. Barro, "An Approach to Developing Accountability Measures for the Public Schools," Phi Delta Kappan 52, No. 4 (1970), p. 202.

1. Administer a pre-test to all the students within a given grade or age level. This test should assess each student's performance on all or a good sample of the relevant objectives for students at that level although it is not necessary to have separate scores for each objective. Thus, one might either use a nationally-normed standardized test if it matches the district's objectives, or construct a measure specifically for the objectives in question. Such a test should not take more than one or, at the most, two hours of testing time.

2. At the end of the year, administer a post-test covering the same objectives that were assessed with the pre-test. It would probably be a good idea to use a different set of items for the two tests, however, so as to minimize potential biases.

3. Plot the two sets of scores (pre-test and post-test) on each pupil within a given grade (or age level). The pupil's teacher should also be identified in this process. The scores of five pupils in each of three classes have been plotted in Figure 1 to illustrate this procedure.

4. Fit a line among these points on the plot that would represent the average or typical relationship between pre-test and post-test scores. The statistical procedure called regression analysis can be used for this purpose.

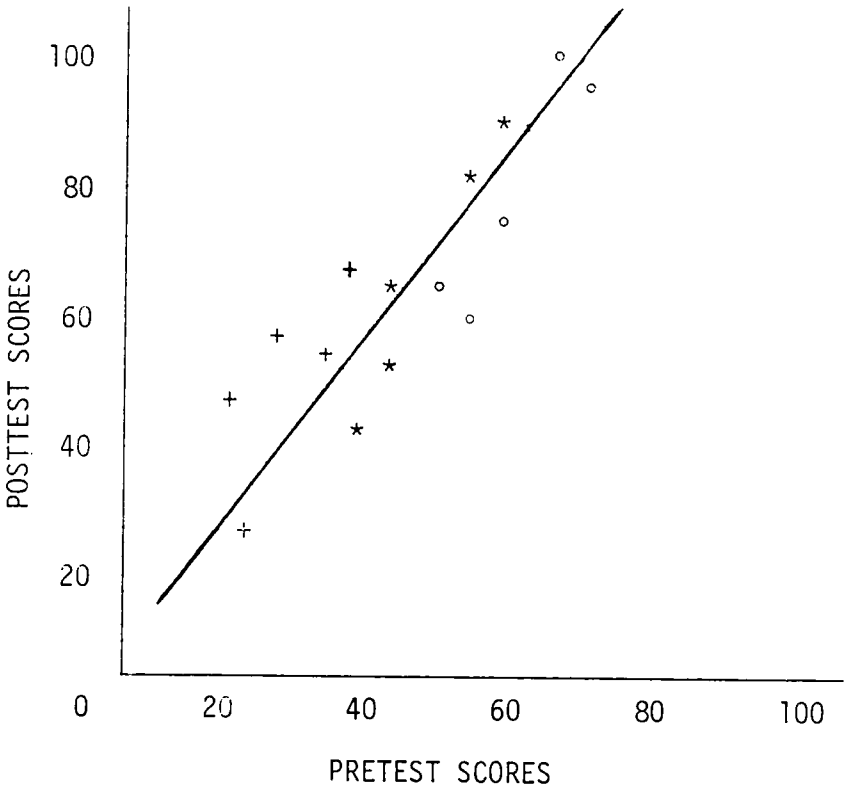
5. Inspect the results in terms of whether a teacher's class tends to fall above or below the line of expected performance as well as whether the class performance tends to be above or below this line. Table I is an example of how these results might be summarized. An examination of this table reveals that although Teacher A's class had relatively poor pre-test scores, they gained more in relation to their starting position than did the students in either Teacher B's or Teacher C's class.[68]

Borich identifies two methods of adjusting gain scores. A residual gain is calculated by correlating the pre-test and post-test scores of all pupils, predicting an expected post-test score for each student on the basis of the pre-test score and subtracting this from the post-test score.[69] This gain score is more representative of the change that has occurred as a result of instruction, because it is independent of the student's initial standing. A second method

---

[68]Alkin, op. cit., pp. 8-9.

[69]Borich, op. cit., p. 51.



Line indicates the average gain in performance between the pretest and the posttest.

+ = Teacher A's pupils

* = Teacher B's Pupils

° = Teacher C's Pupils

FIGURE 1

ADJUSTMENT THROUGH MULTIPLE REGRESSION PLOT

Source: Marvin C. Alkin and Stephen P. Klein, "Accountability Defined. Evaluating Teachers for Outcome Accountability," UCEA Evaluation Comment, V3, N3 (May 1972), p. 9.

TABLE I

MULTIPLE REGRESSION ADJUSTMENTS

Category	Teacher A	Teacher B	Teacher C
Average Pre-test Score	35	50	65
Average Post-test Score	57	65	73
Expected Post-test Score	50	65	80
Difference Between Expected and Actual Post-test Scores	7	0	7
Percent of Pupils Who Are:			
Above Expectancy	80%	40%	20%
At Expectancy	0	10%	0
Below Expectancy	20%	40%	80%

Source: Marvin C. Alkin and Stephen P. Klein, "Accountability Defined. Evaluating Teachers for Outcome Accountability," UCEA Evaluation Comment, V3, N3 (May 1972), p. 9.

identified by Borich is the analysis of covariance, which can statistically hold constant the effect of pre-test scores. Both procedures provide measures of gain independent of the student's initial standing (adjusted gain).

Various research studies have used regression analysis to calculate an adjusted gain. In a study conducted by Veldman and Brophy on measuring teacher effects on student achievement, regression models were used to obtain a measure of gain.[70] They found that it was important to include more than the simple pre-test variable in the regression analysis. The selection and inclusion of the appropriate variables in the regression analysis increases accuracy. Evertson and Brophy in a study of high-inference behavioral ratings as correlates of teaching effectiveness used the regression analysis technique.

> Student's residual gains on the subtests of the Metropolitan battery were computed by using scores from the beginning of the year as pre-score covariates for adjusting gains between testing and the following testing a year later at the beginning of the following grade.[71]

Dyer in his discussion of the School Effectiveness Index Program operating in the New York City Schools again describes the use of a series of regression analyses.[72] The regression model used in

---

[70]Veldman, op. cit., p. 2.

[71]Carolyn M. Evertson and Jere E. Brophy, "High Inference Behavioral Ratings as Correlates of Teaching Effectiveness," Texas University, Texas: Research and Development Center for Teacher Education (March 1973), p. 1.

[72]Henry S. Dyer, "Toward Objective Criteria of Professional Accountability in the Schools of New York City," Phi Delta Kappan 51, No. 4 (1970), p. 210.

calculating an individual school's effectiveness index is presented in detail by Dyer. The reader interested in a complete description of this model is referred to the Bibliography. For the purpose of this review, it is sufficient to note that the discrepancy between predicted and actual level of performance is used to determine a school's effectiveness index in the various content areas through a series of regression analyses. A review of current educational research in the area of teacher effectiveness as determined by achievement data, reveals extensive use of the regression technique to calculate an adjusted gain.

The use of the regression technique to measure gain was found in at least one state education department document serving as a summary report of achievement in reading and math. This particular summary report was issued by the Texas Education Agency, Austin, 1978.[73] As a measure of school quality or effectiveness, this educational agency selected grades four and five for detailed study. Fifth grade achievement scores were regressed on the fourth grade scores. This regression technique provided a set of prediction equations from which fifth grade outcomes could be predicted, based upon fourth grade measures. The differences between average expected scores and actual scores became the agency's measure of a school's effectiveness. This method of the Texas Education Agency to measure the effectiveness of a school is included in this review, because it

---

[73]Texas Education Agency, Interpretive Analyses of 1978-79 Research and Evaluation Reports Volume II. Austin, Texas (1978), pp. 627-650.

uses a regression technique in deriving a measure of gain and because it is somewhat unique in that the agency is measuring a school's effectiveness while most of the literature focuses upon the individual teacher's effectiveness.

The concept of adjusted gain concept versus a raw gain is discussed in much of the literature involving the use of achievement gain. The predominately preferred measure of gain to use in measuring achievement gain for evaluational purposes is the adjusted gain. Various models of regression analysis appear to be a major method of obtaining or calculating the adjusted gain. The reasoning underlying the regression approach is the belief that the teacher's effectiveness should be judged only on the portion of the student's performance that is related to classroom instruction. In addition to taking into account a student's prior knowledge and skill level, this technique is flexible enough through a procedure called multiple regression to take into account several factors simultaneously.

Another frequently used statistical technique in adjusting scores is the analysis of covariance. Gay defines the analysis of covariance (ANCOVA) as a form of ANOVA (analysis of variance) that can be used to equate groups on one or more variables.[74] It involves matching groups on the variable or variables to be controlled. Jason Millman states that it is statistically possible to adjust test scores

---

[74]L. R. Gay, Educational Research Competencies for Analysis and Application (Columbus, Ohio: Charles F. Merrill Publishing Co., 1981), p. 322.

to estimate what they would be if each student in a particular teacher's class had the same values on the characteristics measured as the average value of all the students in the classes being compared (analysis of covariance).[75] Millman further states that it is required for all students to have both their scores on a common achievement measure and their scores on one or more common measure collected before or at the beginning of the school year.[76] The latter measures (covariates) should be measures of student abilities or characteristics that relate to their final achievement. Scores on measures such as previous achievement tests, aptitude tests, and previous course grades have often been used as covariates. The decision on the selection of the appropriate covariate or covariates is crucial if adequate adjustment is to occur.

The analysis of covariance predicts what a student would score on the dependent or criterion variable based upon the student's score on the covariates. In using this procedure, teachers having students who perform better than expected are judged to be more effective than those teachers with students performing below prediction.

Millman provides an excellent example of the use of the analysis of covariance. For reader clarification, this illustration will be presented in its entirety:

> Table II shows the results of a computer analysis of the final examination scores earned by students in an introductory course at an eastern university. That term the course had nine

---

[75]Jason Millman, "Student Achievement as a Measure of Teacher Competence," Handbook of Teacher Evaluation, Chapter 9 (Beverly Hills: Sage Publications, 1981), p. 161.

[76]Ibid.

TABLE II

RESULTS OF AN ANALYSIS OF COVARIANCE IN WHICH THE FINAL EXAMINATION SCORES OF STUDENTS IN NINE DISCUSSION SECTIONS ARE ADJUSTED FOR SECTION DIFFERENCES IN SCHOLASTIC APTITUDE

Discussion Section	Teacher	Class Size	Mean (SAT)	Mean (Final)	Deviation Final From Weighted Average	Deviation Adjusted for SAT
1	A	9	1144	37.9	-0.8	-0.6
2	A	11	1147	39.0	0.3	0.5
3	B and A	12	1191	38.6	-0.1	-0.7
4	B and C	12	1138	37.7	-1.0	-0.7
5	C	10	1147	37.6	-1.1	-0.9
6	C	11	1200	42.3	3.5	2.8
7	D	14	1119	36.8	-1.9	-1.2
8	E	13	1228	40.1	1.4	0.1
9	E	11	1108	38.8	0.1	1.0
Weighted Average			1159	38.7	0.0	0.0

Source: Jason Millman, "Student Achievement as a Measure of Teacher Competence," Handbook of Teacher Evaluation, Ch. 9 (Beverly Hills: Sage Publications, 1981), p. 162.

> discussion sections, each taught by one, or a pair, of five instructors. The College Board's Scholastic Aptitude Test (SAT) was the covariate; the dependent variable was the score on the common final examination. The average SAT scores and the average scores on the final examination showed some variation among the sections.
>
> Consider the data for sections 8 and 9, both taught by Teacher E. The average score on the final for Section 8 was 40.1 or 1.4 points higher than the overall course average of 38.7. Because students in Section 8 had a relatively high average (1228) on the covariate, the students were predicted to do relatively well on the final examination. Thus, the deviation of 1.4 was adjusted downward to .1. That is, the section average of 40.1 is just one-tenth of a point higher than predicted. In Section 9, the students had relatively low SAT scores yet managed to score about average on the final (38.8 versus 38.7). This average performance on the final, accomplished in spite of the handicap of relatively low SAT scores, produced an adjusted final examination average one point higher than predicted. To summarize, although Section 8 had a higher average score on the final than Section 9, the effectiveness of Teacher E was judged somewhat greater in Section 9 (1.0) than in Section 8 (1.1).
>
> With the possible exception of Section 6, the performance of students in all the sections was very close to that predicted from their aptitude scores. It is not surprising that differences among instructors were small, since they met only once a week for an hour. Most of the material covered on the final examination was presented in large group lectures and in the textbook and other readings.[77]

Borich in a discussion on the adjusted gain concept also recognizes the merit of this technique (ANCOVA) as an adjustment measure.[78] He concludes that the procedure creates a measure of gain which is independent of the student's initial standing and is more representative of the change that has occurred during a prescribed instructional period.

To provide the reader with greater insight into the research and current status of using standardized test data for evaluational

---

[77]Ibid., pp. 162-163.

[78]Borich, op. cit., p. 51.

purposes, a more in depth discussion of several studies will be included in this review. It should be noted once again that research in this particular area was found to be very limited.

Several very relevant and interesting results were obtained through a study conducted by Veldman and Brophy[79] involving measuring teacher effects on student achievement. The researchers selected a large urban school district's second and third grade students for subjects.[80] Four years of test scores (Metropolitan) were used to compute mean residual gain scores for 170 teachers who taught during this time span.[81] Veldman and Brophy utilized the residual gain concept and regression analysis to measure a teacher's effectiveness. Their study investigated multiple covariates (pre-test, squared pre-test, pupil sex, year of testing and teacher as predictor of post-test score). Using five different regression models, one for each covariate, they found that the pre-test was the only effective predictor and therefore eliminated the other four. Though only teachers with five or more years experience were selected for the project, they were unable to randomly assign subjects. After computing average gain scores for each teacher for three consecutive years, the researchers found stability correlations from .20 to .80, a rather impressive result.[82] From their study the researchers determined the following results:

---

[79]Veldman, op. cit., p. 1.

[80]Ibid.

[81]Veldman, op. cit., p. 1.

[82]Veldman, op. cit., p. 2.

1. Variation among teachers is substantial.
2. Teacher impact was consistent. They concluded that three-year averages provide a reasonable level of reliability as estimates of teacher impact on student learning.
3. Reasonably stable estimates of teacher influence can be obtained from standardized achievement data.[83]

The Veldman and Brophy study utilized the adjusted gain concept through regression analysis to measure a teacher's effectiveness from achievement test data.

McGowan conducted a study on teacher effect on fifth grade achievement. The research attempted to answer the following questions:

1. What is the effect of individual teachers on elementary school pupil achievement test scores?
2. Does the effect of individual teachers on elementary school pupil achievement test scores remain stable over time, subject area, and ability level of pupils?[84]

The particular procedures of the study were as follows:

1. Iowa Test of Basic Skills results of all sixth, seventh, and eighth grade students were reviewed in the winter of 1982 from 35 elementary schools.
2. If an adequate number of sixth, seventh, and eighth grade students records verified their presence in the same fifth grade classroom for the years 1978-79, 1979-80 and 1980-81, then that particular teacher was selected for the study.
3. Fifteen teachers were selected for this study. Fourth and fifth grade scores were recorded for each student and the fourth grade students divided into three groups for assignment to a level of the ability variable based on their scores.
4. A teacher by level of analysis of variance was performed for the three year period in reading and math.
5. Stability over time and subject was computed by rank correlations.[85]

---

[83]Veldman, op. cit., p. 4.

[84]William E. McGowan, Classroom Effect on Fifth Grade Achievement (Dekalb, Illinois: Northern Illinois University, 1982), p. 4.

[85]Ibid., pp. 49-50.

Teacher effects over time were found to be relatively stable with a Kendall coefficient calculated for complete sample of .20.[86] Across Spearman rank correlations coefficients for three years were .932, .943, and .946, all significant.[87] Other results of the study were that the teacher does have a significant effect upon student achievement and that this effect is distributed across all levels of student ability.

In addition to McGowan's study, several other studies have helped to establish the stability of teacher effectiveness over time and subject. The reader will remember that the issue of stability of this teacher effectiveness was a major criticism surrounding the issue of using achievement test data for evaluation of teachers. Fortune in a study in 1966 used randomly chosen teachers from a Head Start Program to teach the same topic to two different classes. He found significant correlations such as .45.[88] In a very similar study, Justiz used student teachers to present two 30 minute lessons on projects not taught in the high schools. Random selections of students were placed within classes randomly assigned to teachers. Spearman rank order correlations over the two lessons computed from class mean post-test scores were from .60 to .90.[89] Interpretations of both studies need

---

[86] Ibid., p. 64.

[87] Ibid., pp. 73-75.

[88] Rosenshine, op. cit., p. 657.

[89] T. B. Justiz, "A Reliable Measure of Teacher Effectiveness," Educational Leadership Research Supplement 3, No. 1, (1969), p. 51.

to include the fact that random selection of both students and teachers was employed. This is not characteristic of the normal classroom setting. McGowan's study did not employ randomization.[90] However, these studies and others have helped to verify that teacher effectiveness is stable over time, thus helping to eliminate a major criticism involving the use of achievement data in evaluation.

A most recent study using achievement data to evaluate a teacher's effectiveness was completed by Dr. Robert McLean and Dr. William Sanders within the Knox County School System (Knoxville, Tennessee) during the summer of 1983. The study collected achievement data for grades two through five for three school years (1981, 1982, and 1983).[91] One of the primary reasons why Knox County was chosen for this study was because the system had a written curriculum and had given extensive attention to selecting a standardized test to match the curriculum. The California Achievement Test is administered at every grade level annually within the school system. McLean and Sanders hypothesized that one could evaluate teacher effectiveness using student achievement data.

Sanders and McLean attempted to answer or solve many of the problems that opponents of using achievement test data for evaluation have been very vocal in expressing. The problem of mean achievement scores representing factors other than teacher influence was directly

---

[90]McGowan, op. cit., p. 49.

[91]Robert A. McLean and William L. Sanders, Objective Component of Teacher Evaluation--A Feasibility Study (December 1983), p. 6.

attacked. The work of Robert and Ruth Soar, James Popham, and others has been cited previously in this chapter to help the reader gain a perspective of these criticisms. The influences of family background, genetic make-up, socioeconomic status, and so forth on a child's performance on achievement tests have received much attention by many in denouncing the use of achievement test data for evaluational purposes. Sanders and McLean believe that the use of gain scores can eliminate many of these problems. "Gain scores (simple differences between scores obtained in different time periods) allow students to serve as their own controls."[92] They believe that gain scores can be used to measure a student's growth in situations where the tests reflect the curriculum and where there are metrics which are designed to measure sequential learning. Not all subject content is sequential, however, and in these situations the gain score concept cannot be used. The researchers identify the analysis covariance and multiple regression as possible alternatives in these cases of nonsequential subject content.[93] They believe these methods can be used quite effectively with problems containing only fixed effects.

The field of education, however, is characterized by random effects.

> If it is recognized that teachers differ in their effectiveness in influencing student learning, it is reasonable to assume that a distribution of teacher effects exists for the entire population of teachers who are teaching similar objectives.[94]

---

[92]Ibid., p. 3.

[93]Ibid.

[94]Ibid., pp. 4-5.

This randomness effect is made even clearer when the problem of regression to the mean is considered. This also is a problem often cited by those critical of using achievement data for evaluational purposes. Regression to the mean occurs when students scoring extremely high or low on prior tests tend to move toward the mean on subsequent tests. Since most school classrooms are characterized by non-random grouping, the regression to the mean problem will bias teacher effectiveness measures derived from student achievement data. Teacher effectiveness measures will be biased under nonrandom grouping due to the inconsistent direction of the regression to the mean effect. Both high and low scores in non-equal numbers would be within the same classroom and the regression to mean effect would be in different directions, extreme pupil scores moving toward the mean. Highest scoring students would show less increase than the lower scoring students, thus biasing any measure of teacher effectiveness designed from unadjusted gain scores.

Sanders and McLean chose to use statistical procedures developed by C. R. Henderson to control for the regression to the mean problem. Henderson's mixed model solution provides a best linear unbiased prediction of the random effects.[95] This best linear unbiased solution procedure offers a solution to the problem of regression to the mean by essentially pulling the prediction closer to the mean, as a function of distance, ratio of the variance and quantity of information available.[96] Sanders and McLean explain the regression to the mean

---

[95] Ibid., p. 5.

[96] Ibid., p. 52.

problem as a result of the initial tests being biased predictors of the individuals true ability and find the best solution to the problem is to use unbiased predictors.[97] Such models may contain both fixed and random classification variables, with some of the fixed variables in the form of continuous covariates.[98] A fixed class variable does not have a probability distribution; it is an entity in itself. Examples of a fixed class variable would be male/female or democrat/republican. However, a random class variable such as teacher effect has its own entity, but it is also a member of a distribution. A continuous variable is one that assumes an unlimited number of values. Such variables as age/income are continuous because they can be infinitely divided. The analysis of covariance using prior measures of student performance as covariates has been used extensively in prior studies, under the assumption that the covariates are continuous fixed effects. Since teacher effects are random, however, the analysis is a mixed model problem. Henderson's model provides a solution to this mixed model problem.[99]

Sanders and McLean's study provided answers to the following questions:

> 1. Are there measurable effects of teachers on student gain scores?
>
> Yes. There were significant differences among teachers for grades 3,4 and 5 for math, reading, and language.
>
> 2. Do teacher effects tend to be consistent from year to year?
>
> Yes. Teacher effects estimated independently for two years were positively correlated.

---

[97] Ibid.

[98] Ibid., p. 5.

[99] Ibid., p. 5.

3. How much were estimated teacher effects influenced by differences in ability of students?

Virtually none. The relationship between estimated mean gain scores and prior class mean was virtually unrelated in this study.

4. What was the relationship between teacher effects estimated separately for math, reading, and language?

Highly correlated.

5. Are teacher effects biased due to the regression to the mean problem?

No. Henderson's MMS procedure provides unbiased predictions of random effects.

6. Was there evidence that better or poorer teacher estimates were clustered within geographical areas of Knox County?

No. No obvious pattern of teacher effects existed relative to geographic location of the schools.[100]

V. MERIT PAY PLANS BUILT UPON THE CONCEPT OF STUDENT GAIN AS MEASURED BY STANDARDIZED TESTS

Several merit pay plans in the field of education utilize student test scores in their evaluation of teachers for merit pay and were included in this review. The Educational Research Service reported in November of 1983 that eight districts used test scores.[101] Two of these eight systems will be described because they offer unique aspects related to the topic of using achievement test data for education of teaching.

The Seiling Public Schools System, Seiling, Oklahoma, utilizes a pre-test/post-test design.[102] The district uses both standardized

---

[100] Ibid., p. 1.

[101] Educational Research Service, Merit Pay Plans for Teachers: Status and Descriptions (Arlington, Virginia: Educational Research Service, Inc., 1983), p. 14.

[102] Gerald Daugherty and Gladys Dronberger, "A Merit Pay Program That Works: The Seiling, Oklahoma, Experience," Merit Pay Plan for Teachers: Status and Descriptions (Arlington, Virginia: Educational research Services, Inc., 1981), p. 112.

and objective-referenced tests. From the pre-test information, a predicted or targeted increase is projected.[103] At the end of a school year, the amount of gain made by each student is calculated and a class average determined and compared to the predicted gain.

> The amount of bonus the teacher receives depends on the average gain in normal curve equivalency points for all the students in the school and the average gain in each class he/she teaches. Thus, the system is designed to reward both faculty and individual performance.[104]

Record Keeping and calculations are handled by a TRS-80 Model I microcomputer.[105] The computer is programmed, once pre-test and post-test scores are typed in, to print the score, the gains, class average, school average, and the teacher's bonus. The Seiling Public School System's merit pay plan is described because it is based upon the gain score concept utilizing standardized achievement tests, a predicted score in calculation of bonuses, and combined a total school faculty and individual bonus plan.

The Dallas Independent School District's merit pay plan is also described within the Educational Research Service Report on Merit Pay Plans for Teachers. The Dallas Merit Pay Plan also uses standardized test scores in evaluating teachers for merit pay. The plan is designed to reward an entire faculty for outstanding performance. A very brief description of how the system operates is as follows:

---

[103] Ibid.

[104] Ibid.

[105] Ibid.

1. Grade and subject area achievement is projected utilizing a cohort longitudinal regression approach.
2. Each school is projected against its past student achievement trends and those across the system.
3. Confidence intervals are calculated and each school identified as to whether they fall below, at, or above the expectations.
4. Faculties with documented exceptional progress are awarded the bonus compensation.[106]

Statistical procedures to determine school effects from established expected levels of achievement on the basis of achievement history and on the basis of achievement trends of like students in the Dallas District have been tested and found extremely reliable. The use of longitudinal data to calculate predicted scores to use in calculation of a gain score is the unique aspect of the Dallas merit pay plan.

The Dallas School System and the Seiling Public School System report increases in student achievement and improved teacher morale after implementation of their merit pay plans. The reader interested in more specifics of how these two merit pay plans (or those of the other six systems) based upon test scores operate should consult the Educational Research Service publication.

VI. SUMMARY

This review of the literature has attempted to trace the move within education toward using achievement test data for evaluating classroom teachers. A brief history and review of current methods used

---

[106]F. S. Calhoun and N. J. Protheroe, "Dallas Merit Pay Plan," Merit Pay Plans for Teachers: Status and Descriptions (Arlington, Virginia: Educational Research Services, Inc., 1981), pp. 128-131.

to evaluate teaching, such as administrative rating, student ratings, and administrative classroom observations have been discussed. McNeil and Popham, Popham and Hawkins and Stoops have been cited to highlight the subjective nature of these frequently used methods of teacher evaluation. The accountability movement affecting the educational field today has been examined and found to be adding to the increasing pressure for more objective accounting or evaluation of teachers. Education can no longer stand apart from this movement. This review has shown a strong need for a more objective means of evaluating the classroom teacher. The use of achievement test data as a more objective measure of teacher effectiveness has been suggested and investigated by several researchers. However, this proposal has been opposed by many and has initiated a very intense debate as to its merit. Major criticisms surrounding the use of achievement test data for teacher evaluation are listed below:

1. Two many non-school factors affect a student's performance on achievement tests (prior ability, previous knowledge, home and peer influence, socioeconomic status, family attitude and support for learning at home, etc.).

2. The structure of achievement tests is such that they do not differentiate among students chiefly on the basis of what those students have learned in schools.

3. Achievement tests are not always congruent with individual classroom objectives.

4. The use of achievement test data for evaluation of teachers can result in a tendency for teachers to "teach the test" and often to emphasize lower cognitive learning.

5. Achievement tests do not measure multivariate outcomes.

6. A teacher's effectiveness as measured by achievement test data is unstable over time and across subject areas.

7. The statistical problems in measuring change are of such a magnitude that a system of teacher evaluation based upon student achievement data is not possible.

Each of these major criticisms was discussed with reference to current research and literature. Dunkin and Biddle, McNeil and Popham, Robert and Ruth Soar, Coleman, Popham, Berliner, Brophy, Gall, and Rosenshine provided direction in this discussion.

An attempt was also made within the literature review to answer many of these criticisms with current research. For example, the concerns revolving around the statistical problems of measuring change were discussed in light of the work of Sanders and McLean (1983). These researchers found that the best linear unbiased predictor (BLUP) procedure of C. R. Henderson could resolve the statistical problem of the regression to the mean. This is only one example, and no attempt will be made in this summary to review each criticism and its response. There has also been no attempt to claim that all criticisms can be explained or proven invalid. However, many of the criticisms, if studied carefully in light of the research, are not as serious as they may appear at first. The work and research of Borich, Cronbach,

Veldman and Brophy, Klein and Alkin, and Evertson and Brophy have been used to counter many of these major criticisms. In addition, the detailed discussion of the research of McGowan (1982), Fortune (1975), Justiz (1969), and Sanders and McLean (1983) provided valuable insight into using achievement test data to measure a teacher's effectiveness.

The last section of this review included a review of recent studies involving the use of achievement test data for evaluating a teacher's effectiveness. These studies (Justiz, 1979; Veldman and Brophy, 1973; Fortune, 1975; McGowan, 1982; McLean and Sanders, 1983) highlight the potential for future studies involving the concept of using achievement data to evaluate a teacher's effectiveness. This does not imply that achievement test data provide the best instrument for teacher evaluation, but practical considerations in schools encourage the use of achievement test data as at least a component in a broader teacher evaluation plan. The availability of achievement data and their established validity and reliability make them attractive for further investigation into their use in teacher evaluation.

CHAPTER III

INTRODUCTION

This chapter is divided into the following sections: the population, data collection procedures, analysis of the data, and a review of the statistics.

I. THE POPULATION

The population for this study consisted of all fifth grade students and teachers within the Blount County School System (Tennessee) for the 1982-83 and 1983-84 school years. Permission to conduct the study was obtained from the Blount County School Board in May of 1984. A copy of the letter requesting permission to conduct the study can be found in the Appendix. All faculty members and students in the fifth grade for the selected years were used as subjects for this study (Figure 2).

II. DATA COLLECTION PROCEDURES

1. Computer tapes of Blount County test data were obtained from the State Testing and Evaluation Center. A letter authorizing the release of such data was sent to the Center by the Superintendent of Blount County Schools. A copy of this authorization letter can be found in the Appendix. The study used achievement test data from fifth grade classes in three subject areas: reading, math, and language arts. Achievement test data and school ability test data from the third and fourth grades were also used. Computer tapes contained

NUMBER OF TEACHERS AND NUMBER OF STUDENT MEAN SCORES BY SUBJECT

Year	Grade 5 Math		Grade 5 Reading		Grade 5 Language	
	Number of Teachers	Number of Students	Number of Teachers	Number of Students	Number of Teachers	Number of Students
1983	21	581	24	499	23	586
1984	19	550	26	542	21	553
	40	1131	50	1041	44	1139

Total Student Mean Scores 3,311

COMBINED DATA--NUMBER OF TEACHERS AND STUDENTS WITHIN STUDY*

Subject	Number of Teachers	Number of Students
Reading	32	1041
Math	25	1131
Language	28	1139

*Combined data includes teachers having taught the same subject for both years of the study.

FIGURE 2

TEACHERS AND STUDENTS IN THE STUDY

information on third and fifth grade achievement test data and fourth grade school ability test data for the following school years: 1981-82, 1982-83, 1983-84.

2. In order to match records to enable the calculation of adjusted scores for individual students, an alphanumeric merging system using student names as the identifying variables was employed.

3. Individual records were edited to ensure consistent spelling of student names in successive years. A loss of records occurred during this editing process for several reasons: (a) student not taking the test in successive years, (b) student transferring to another school, and (c) incongruent name spelling between years. This editing process was completed by using a line editor (XEDIT) in the CMS system on an IBM mainframe computer.

4. Alphabetized lists sorted by school, year, subject and grade were developed.

5. Roll books were collected from participating elementary schools to provide information on teacher identification. A letter was sent to Blount County Elementary principals requesting them to collect the books for two school years (1982-83 and 1983-84) and to bring them to the central office where the elementary supervisor collected them. A copy of this request letter can be found in the Appendix.

6. Teacher identification taken from the roll books was added to the student record data using XEDIT in the CMS system.

7. Elementary principals were asked to subjectively rate their fifth grade teachers in terms of where they believed the teacher would

fall in a ranking system developed county wide from achievement data. This ranking scale can be found in the Appendix.

8. Elementary supervisors from the Blount County School System were asked to comment on ranking of fifth grade teachers obtained from this study. The ranking was divided into three levels and data as to where they would place individual teachers in reading, math and language arts were recorded. This ranking form is also found in the Appendix.

III. ANALYSIS OF THE DATA

1. To detect any potential ceiling/floor effect problem with the data set, a frequency distribution study was performed initially. Ceiling/floor effects refer to the extent a test is too easy or too difficult. The ceiling effect occurs when the test is too easy so the students "top out." It is not possible to tell how much more the student knows because the items were too easy. The floor effect occurs when the test is too hard. The student's scores are like those the student would have received by chance. It is impossible to determine what the student does know, because the items were too hard. The rationale behind this analysis is if these data exhibited any ceiling or floor effect problem, they would invalidate any rating or ranking system designed from the data and be unfair and unjust to the teachers involved.

A distribution frequency study was completed through the SAS (Statistical Analysis System) computer program: univariate. A frequency distribution was obtained for third and fifth grade scores in reading, math, and language arts for 1981, 1982, 1983, and 1984.

2. Longitudinal data (fourth grade school ability and third grade achievement test data) from grades three and four were used to adjust student fifth grade achievement scores in reading, math, and language. An analysis of covariance was performed. The model or design included the teacher as a classification variable and fourth grade school ability/third grade achievement test scores (indicators of prior ability) as covariables. The fifth grade achievement score was the dependent or criterion variable. This analysis was performed to adjust for different student ability levels to obtain teacher effects independent of ability levels of students and to determine if there were significant differences among teachers in terms of effectiveness.

The analysis of covariance adjusts post-test scores for initial differences on some variable (in this case third grade achievement and fourth grade school ability) and allows a comparison of the adjusted scores. Groups or students are equalized in regard to the control variables. Any variable that is correlated with the criterion or dependent variable can be controlled by using covariance.

3. An adjusted score per student was calculated in the areas of reading, math and language arts. These adjusted individual scores were then used in calculating a score per teacher for each year. Statistical procedures developed by C. R. Henderson were used in further analysis in this study to eliminate the statistical problem of the regression to the mean. Henderson's mixed model solution method permitted control of the problem of the regression to the mean and to

adjust for nonrandomness.[107] It is recognized by both educators and noneducators that teachers differ in their effectiveness with different students. Therefore, teacher effects should be viewed as a random class of effects instead of being fixed. From Henderson's mixed model solution method, best linear unbiased prediction (BLUP) of random effects can be obtained. (A more detailed discussion of this statistical procedure is included in the final section of this chapter, entitled Statistics of the Study.)

4. A correlation coefficient was calculated between BLUP estimates for the two consecutive school years within the study. Four correlation coefficients were calculated: reading, math, language arts, and total subjects.

5. A composite score for each teacher was calculated from the BLUP estimates for the two school years of this study. A ranking of teachers from these combined scores was developed.

6. Percentages of agreement were calculated between both principal and supervisory ratings and the ranking developed from achievement data.

IV. STATISTICS OF THE STUDY

Analysis of Covariance

An analysis of covariance was used within this study to adjust the fifth grade achievement scores to a constant ability level. One of

---

[107] C. R. Henderson, "Best Linear Unbiased Estimation and Predictions Under a Selection Model," Biometrics 31 (1975), pp. 432-477.

the major criticisms involving the use of achievement data for evaluation of educational personnel, centers on the issue of prior student ability and differences. The ANCOVA is a statistical procedure designed to make such an adjustment. In this study the dependent or criterion variable was the fifth grade achievement score. Covariables selected were third grade achievement data and a fourth grade school ability score. As reported in the literature, pre-test performance, IQ, readiness, and specific aptitude are frequently controlled in educational research studies through the use of analysis of covariance.[108] Therefore, the selection of covariables for this particular study appear to be well founded.

The particular model used in the calculation of the ANCOVA was as follows (math example):

Math 5th = Teacher math 84 Math - 3 - 82 School Ability - 4 - 83

The purpose or rationale of this model was to get teacher effects independent of ability levels of students and determine if there are significant differences among teachers in terms of their effectiveness.

Many previous studies have concluded with effectiveness measures calculated through the ANCOVA. This study, however, went one step further to control for the regression to the mean problem that the ANCOVA does not solve in cases of nonrandomness, a characteristic of most classroom settings.

---

[108]L. R. Gay, Educational Research Competencies for Analysis and Application (Columbus, Ohio: Charles E. Merrill Publishing Co., 1981), p. 324.

Best Linear Unbiased Prediction (BLUP)--C. R. Henderson

A frequently discussed problem involved in using student achievement data for evaluation of teachers is the problem of regression to the mean. The regression to the mean refers to the tendency of students who have scored either extremely high or low on prior tests due partly to testing process errors to move toward the mean of all students on subsequent tests. The regression to the mean is also a feature of any linear prediction rule that is best in the "least squares" sense.[109] The problem is intensified due to the fact that this effect operates unequally on students. Since most classrooms are characterized by nonrandom assignment to teachers, teacher effects calculated from student achievement data will be biased due to the regression to the mean problem. Henderson's model--Best Linear Unbiased Prediction (BLUP) was utilized in this study to eliminate the problem of the regression to the mean.

It is generally accepted that teachers differ in their effectiveness or success in influencing student learning outcomes. It should be clear that a distribution of teacher effectiveness exists for the entire population of teachers teaching toward similar objectives. Therefore, it is reasonable to view teacher effects as a random class of effects and not fixed. With the mixed model solution (MMS) procedure Best Linear Unbiased Prediction (BLUP) of these random effects can be obtained. BLUP estimates of teacher effects offer a solution to the regression to the mean problem.

---

[109]William L. Hays, Statistics for the Social Sciences (New York: Holt, Rinehart and Winston, Inc., 1973), p. 626.

Another positive feature of the Henderson's MMS procedure in terms of its use in teacher evaluation from student achievement data is that it can also be used to adjust for known differences among students in order to eliminate the problem of nonrandom classrooms. Henderson's MMS will allow one to include continuous and class fixed variables (achievement test scores) simultaneously with one or more random class variables (teacher effect).[110] The analysis can provide BLUP estimates of the teacher effects and adjust these effects via covariables. This characteristic of Henderson's model has great potential in cases where such adjustment is essential to obtain an unbiased estimate of teacher effects.

A brief example of how the BLUP analysis of Henderson operates is included for clarification. An example will help to clarify this analysis:

1. If one had a normal distribution of sixth grade achievement gain, the distribution would resemble a normal bell shape curve (Figure 3).

2. If a principal was considering hiring a new teacher, but had no direct information of his or her effectiveness--the best guess or estimate would be the mean.

3. But if the principal had an achievement gain score on one of the prospective teacher's former students and this student had received a high gain score, where then is the best estimate of where

---

[110]Charles R. Henderson and C. R. Henderson, "Analysis of Covariance in Mixed Models with Unequal Subclass Numbers," Communication in Statistics--Theory and Methods (1979), pp. 751-787.

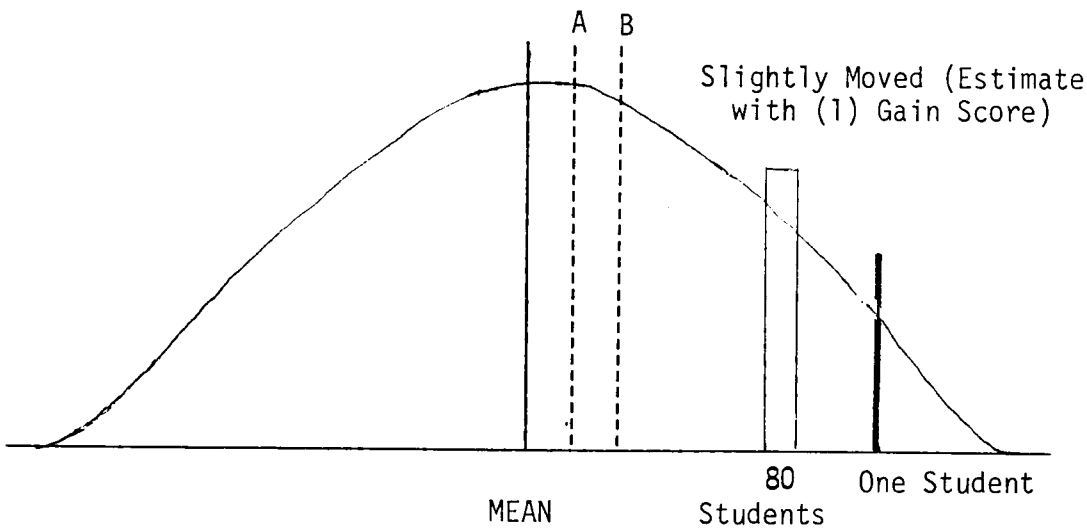


FIGURE 3

MEAN ADJUSTMENT--BLUP ESTIMATES

this teacher falls in relation to the distribution of all teachers? The original estimate would need to be adjusted. In this example, the estimate would move slightly to the right of the mean (line A).

4. However, if the teacher had been hired on a probationary basis for one year and scores on 80 students (as shown on curve) had been recorded, the estimate would be moved further from the mean (line B). The more information available, the more accurate the prediction.

The concept of allowing estimates to be pulled from the mean of the entire population as we get more information is the concept of the Best Linear Unbiased Prediction and solves the regression to the mean problem. This "pulling" of the prediction closer to the mean as a function of distance, ratio of the variances, and quantity of information forms the basis of the BLUP concept.[111]

The particular BLUP equation used in this study was as follows:

$$= + (- \quad - =) \qquad \frac{\sigma^2 \text{teachers}}{\sigma^2 \text{ teachers} + \frac{\sigma^2 \text{students}}{N}}$$

$y \quad y_T \quad y$

$\bar{\bar{y}}$ = mean gain score for all students over entire population

$\bar{y}_T$ = sample mean gain score for a specific teacher

$\sigma\text{-}^2$teacher = variation among all teacher means in population

---

[111]Robert A. McLean and William L. Sanders, Objective Component of Teacher Evaluation, A Feasibility Study (December 1983), p. 52.

$o\text{-}^2$student = variation among all students scores within teacher

N = number of student records for a teacher.

The procedure allows the pulling from the mean to be a function of the number of observations and of the unintentional variability of teachers relative to the unintentional variability of students.

An example of this BLUP procedure, taken from this particular study is presented for further clarification.

$$BLUP = M + (\underset{X}{-} - m) \quad \frac{\sigma^2 \text{teachers}}{\sigma^2 \text{teachers} \; \frac{\sigma^2 \text{students}}{N}}$$

$$= 169 + (175 - 169) \quad \frac{18.19}{18.19 + \frac{92.53}{21.94}}$$

$$= \frac{18.19}{18.19 + 4.22}$$

$$= \frac{18.19}{22.41}$$

$$= 169 + (6) \quad (.809)$$

BLUP Estimate = 174

169 - mean score for all students in population

175 - sample mean score for a specific teacher

18.19 - variance component among all teachers within population

92.53 - variation among all student scores within teacher

21.94 - number of student records for a teacher

A second example of the BLUP procedure taken from the work of Sanders and McLean in its entirety is presented:

Given that the population mean and variance of true IQ is 100 and 225, respectively, and if an individual takes one IQ test and scores 130 on a test that has test error variance of 25, what is the best prediction of true IQ for that individual?

In this example,

$$\text{Prediction of true IQ} = u + (X - u)\frac{\sigma^2_I}{\sigma^2_I + \frac{\sigma^2_T}{N_T}}$$

$$= 100 + (130 - 100)\frac{225}{225 + (25/1)} = 127$$

where:

u = population mean of true IQ and IQ test scores

X = mean of IQ tests taken by the individual

σ^2_I = population variance of true IQ

σ^2_T = variance of test error

N_T = number of tests taken

The best predictions of this individual's true IQ is not 130 rather 127. Why? This expression for the conditional mean of true IQ is given IQ test score may be obtained from the joint distribution of true IQ and IQ test score if both true IQ and the errors of the test are both assumed to be normally distributed. Note that this prediction of true IQ is "pulled" ever closer to the population mean as the ratio of test error variance to population variance increases or as NT becomes smaller. Thus, if a little information is available, a prediction close to the population mean tends to be best. If more information is available, a prediction closer to the sample mean is best. Henderson's Best Linear Unbiased (BLUP) procedures controls for the regression to the mean problem. This problem has been cited by many in denouncing the use of achievement data for purposes of evaluation of educational personnel.[112]

---

[112] Ibid., pp. 51-52.

Variance Component Calculation

The formula or model for the variance component calculation was as follows:

Mean Square* - Error Mean Square = Expected Mean Square Variance Error

$$*\text{Mean Square} = \frac{\text{Sums of Squares}}{\text{DF}}$$

An example is provided for further clarification

Source	D.F.	S.S.	Expected M.S.	Error Mean Square
TRead5	23	10,687.7	20.46	92.52

$$\frac{10687}{23} - 92.53 = 20.46\ \sigma^2_{\text{TRead 5}}$$

$$464.68 - 92.53 = 20.46\ \sigma^2_{\text{TRead 5}}$$

$$\frac{464.68 - 92.53}{20.46} = \hat{\sigma}^2_{\text{TRead 5}}$$

$$\frac{372.15}{20.46} = \hat{\sigma}^2_{\text{TRead 5}}$$

$$18.19 = \hat{\sigma}^2_{\text{TRead 5}}$$

Z Score Formula--Standardizing Scores

In this study, a statistical problem arose as a result of a change in test editions between the two years of the study. In 1984, the fifth grade Stanford Achievement Test was changed from the 1973 edition to the 1983 edition. The scale scores from the 1983 edition were not comparable with the other data within the study. Scale scores had to be standardized to provide comparable scores.

(Z) scores are developed through linear transformation, and therefore, retain the exact numerical relations of the original scores.[113] In (Z) score calculation, the difference is found between an individual's raw score and the mean of the normative group and divided by the standard deviation of the normative group. The formula is as follows:

$$Z = \frac{X - M}{SD}$$

X = any raw score in a frequency distribution

M = mean of the distribution

SD = standard deviation of the distribution

To eliminate negative numbers in these calculations, these (Z) scores can be adjusted to a mean of 50 and a standard deviation of 10. In converting an original standard to a new scale, it is necessary to multiply the standard score by the desired SD and subtract it from the desired mean. Using this scale, a score of 50 corresponds to the mean and the standard deviation is 10. This standardization procedure provides scores that are comparable.

---

[113]Ann Anastasi, Psychological Testing--Third Edition, The MacMillan Company, 1968, pp. 52-53.

CHAPTER IV

INTRODUCTION

This chapter contains the introduction, presentation, and analysis of data used in developing a teacher evaluation system based upon student achievement tests results. The data and subjects were obtained from fifth grade students and teachers within the Blount County School System (Tennessee) for two consecutive school years: 1982-83 and 1983-84.

The data analysis and presentation was broken into the following areas of discussion:

1. Distribution diagnostics
2. Data adjustments
3. Development of Best Linear Unbiased Prediction Estimates
4. Correlation study of consecutive year estimates
5. Composite teacher rankings
6. Two-level independent collaboration of teacher rankings

I. PRESENTATION AND ANALYSIS

Distribution Diagnostics

Before beginning the data manipulation in developing the evaluation system of teachers from student achievement data, a diagnostic measure was undertaken. To detect any potential ceiling or floor effect problem with the data set, a frequency distribution study was conducted. The ceiling effect occurs when the test is too easy

and, therefore, the students "top out." It is impossible to determine how much more the student knows because the items were too easy. The floor effect occurs when the test is too difficult. The student receives scores that are like those he would have received by chance. It is impossible to determine what the student does know because the items were too difficult. Both conditions are, it is hoped, considered and eliminated by the test publisher in the actual test construction stage. An appropriate amount of stretch or elasticity should be built into the test. However, one should not rely upon this assumption. Also, there are several non-related test construction factors that could result in both a ceiling and a floor effect problem. Two of these factors are: (1) a poor fit between the curriculum and test; (2) the violation of the integrity of the testing process.

The rationale supporting this aspect of this study is that if a data set exhibited any ceiling or floor effect problem, this problem would tend to discredit any rating or evaluation system designed from the data and should certainly be a consideration in its interpretation.

The distribution frequency study was completed through the Statistical Analysis System computer program: Univariate. This section presented a distribution study that was obtained for third and fifth grade achievement data in reading, math and language arts and for the school ability tests in the fourth grade for students in the fifth grade during the two school years of this study, 1982-83 and 1983-84.

Distribution bar charts were developed and are presented for third, fourth, and fifth grade test data. In addition, frequency distribution tables are presented. A test distribution free of ceiling

and floor effect should represent a normal bell shape curve. A visual inspection of the data set bar graphs provides information about the nature of the various distribution patterns.

Distribution patterns for third grade achievement in reading Figure 4)/math (Figure 5)/language arts (Figure 6) for 1981, fourth grade school ability distribution for 1982 (Figure 7), and fifth grade achievement in reading (Figure 8)/math (Figure 9)/language arts (Figure 10) for 1983 are included in this first data set. Consistently normal distributions were exhibited in all the bar graphs with one exception. In language arts, third grade (Figure 6), there appeared to be a heavier distribution at the upper end of the scale. This variation from the normal bell shape curve distribution would indicate the existence of a slight ceiling effect.

The distribution patterns for the achievement and school ability tests for the second school year of this study, 1983-84, were included in the second data set. Third grade achievement (reading/ math/language) for 1982, school ability test data for 1983, and fifth grade achievement (reading/math/language) for 1984 are included in this data set.

Normal distribution patterns were found in reading--third grade, 1982 (Figure 11); reading--fifth grade, 1984 (Figure 12); math--third grade, 1982 (Figure 13); math--fifth grade, 1984 (Figure 14); language arts--fifth grade, 1984 (Figure 15); and school ability--fourth grade,1983 (Figure 16). A slight ceiling effect is once again demonstrated in the distribution pattern of language--third grade, 1982 (Figure 17).

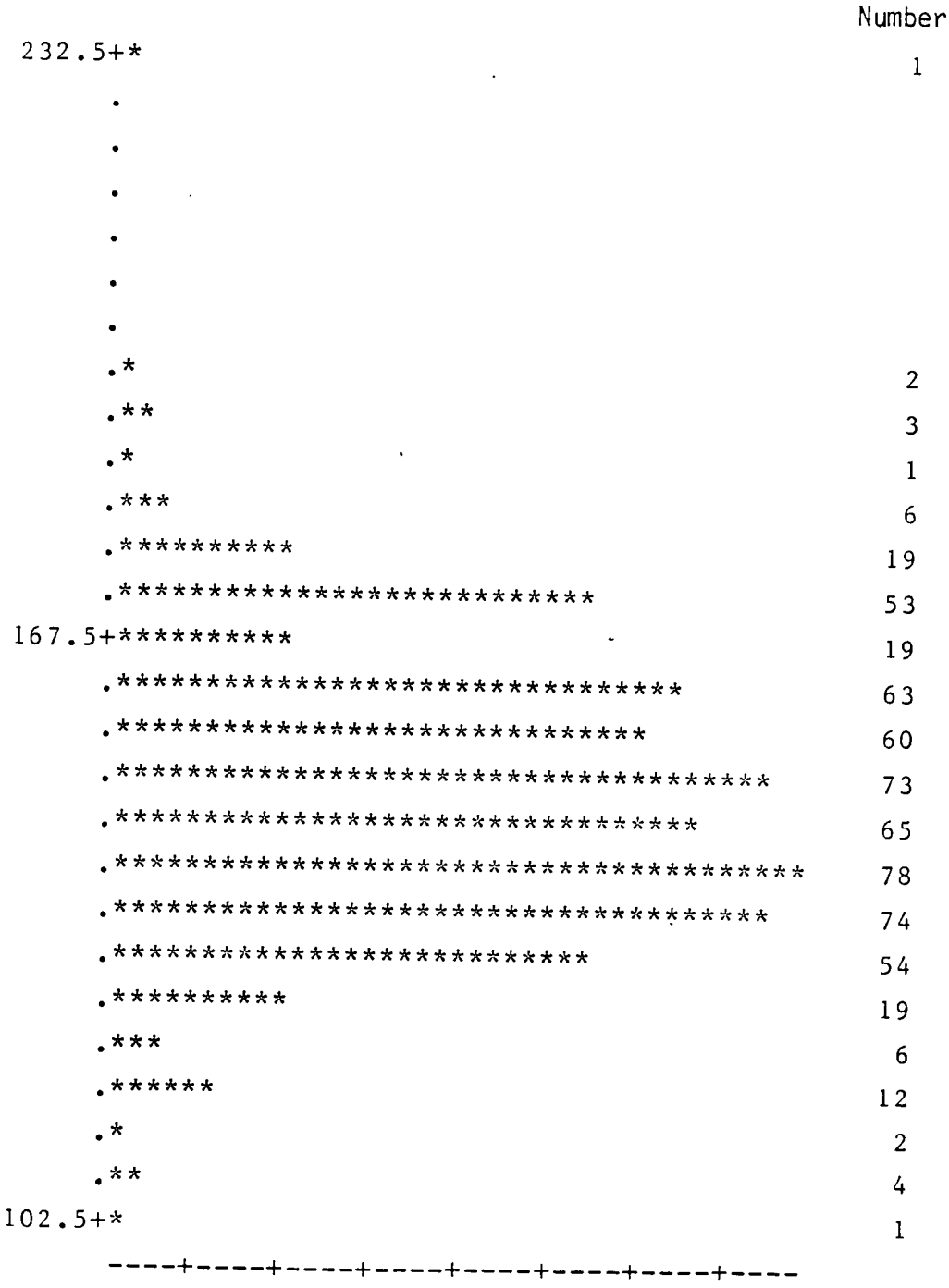


FIGURE 4

BAR CHART OF DISTRIBUTION OF SCALE SCORES:
READING, THIRD GRADE, 1981

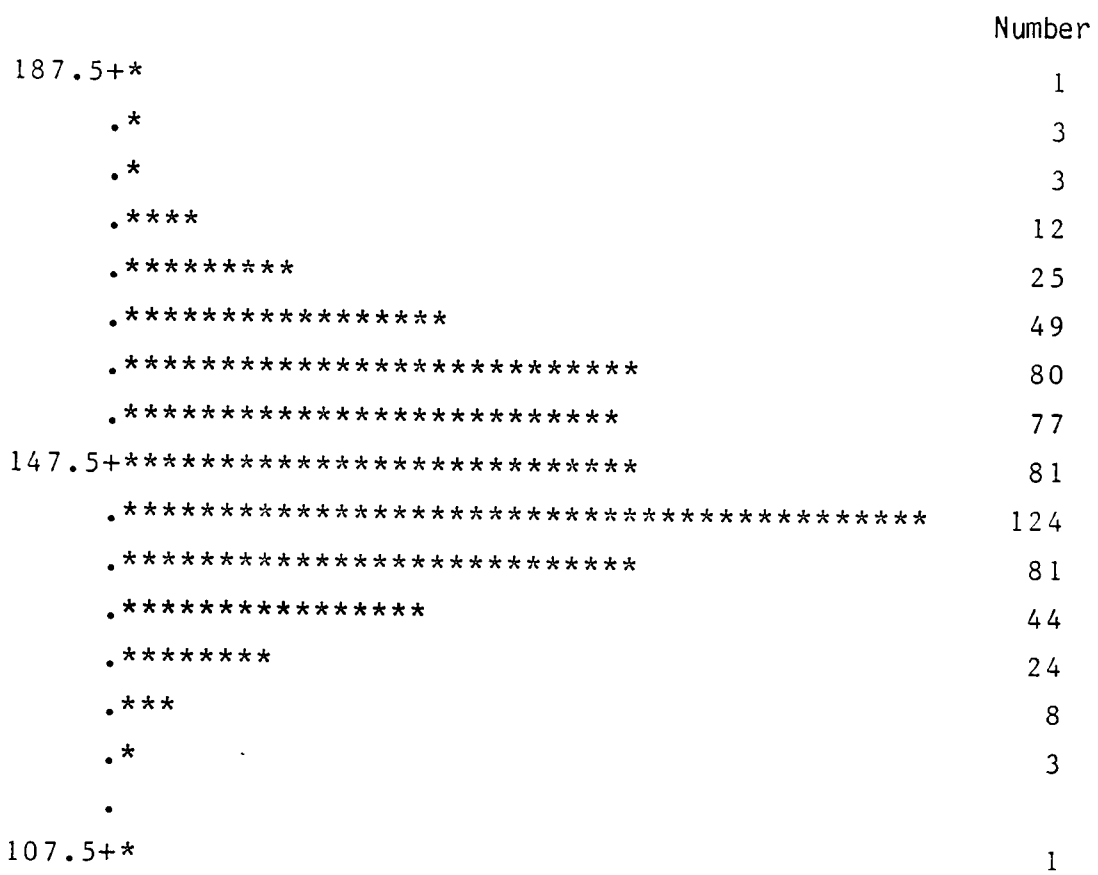


FIGURE 5

BAR CHART OF DISTRIBUTION OF SCALE SCORES:
MATH, THIRD GRADE, 1981

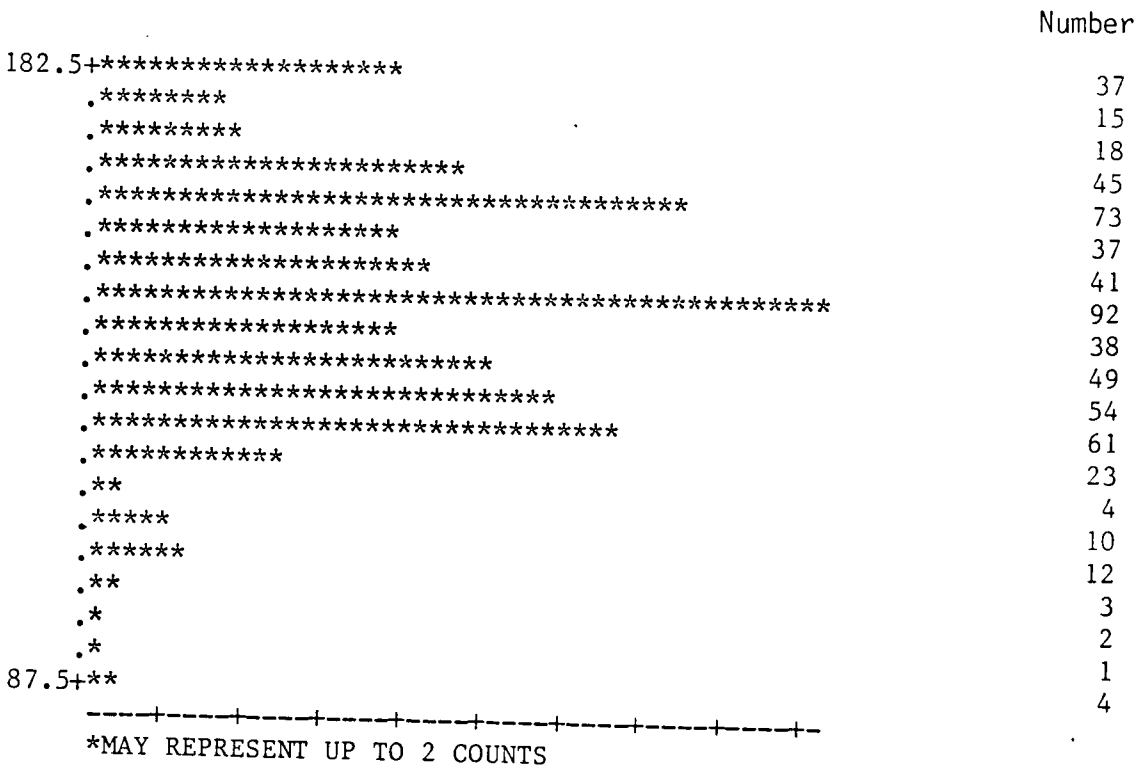


FIGURE 6

BAR CHART OF DISTRIBUTION OF SCALE SCORES:
LANGUAGE ARTS, THIRD GRADE, 1981

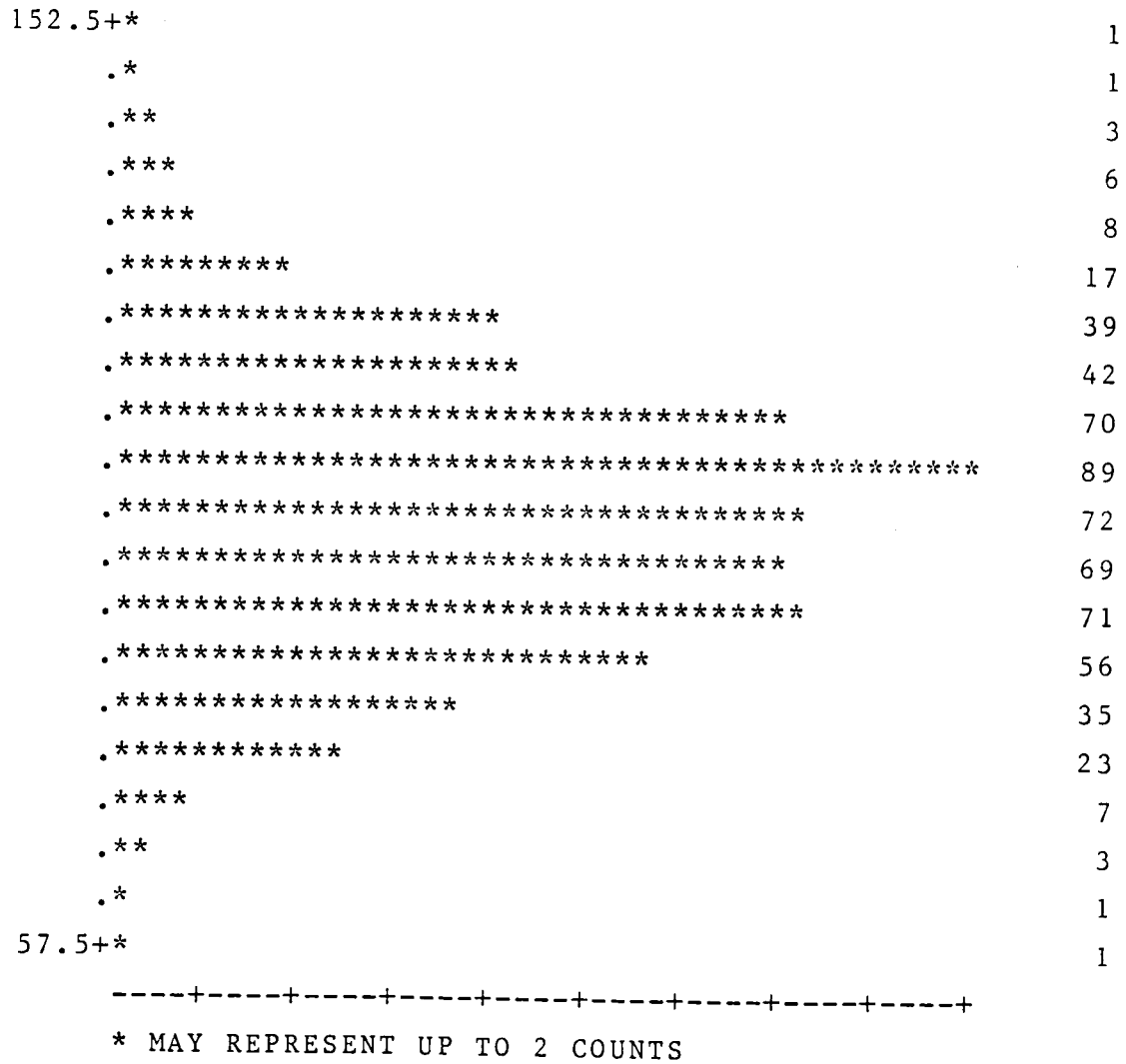


FIGURE 7

BAR CHART OF DISTRIBUTION OF SCALE SCORES:
SCHOOL ABILITY, FOURTH GRADE, 1982

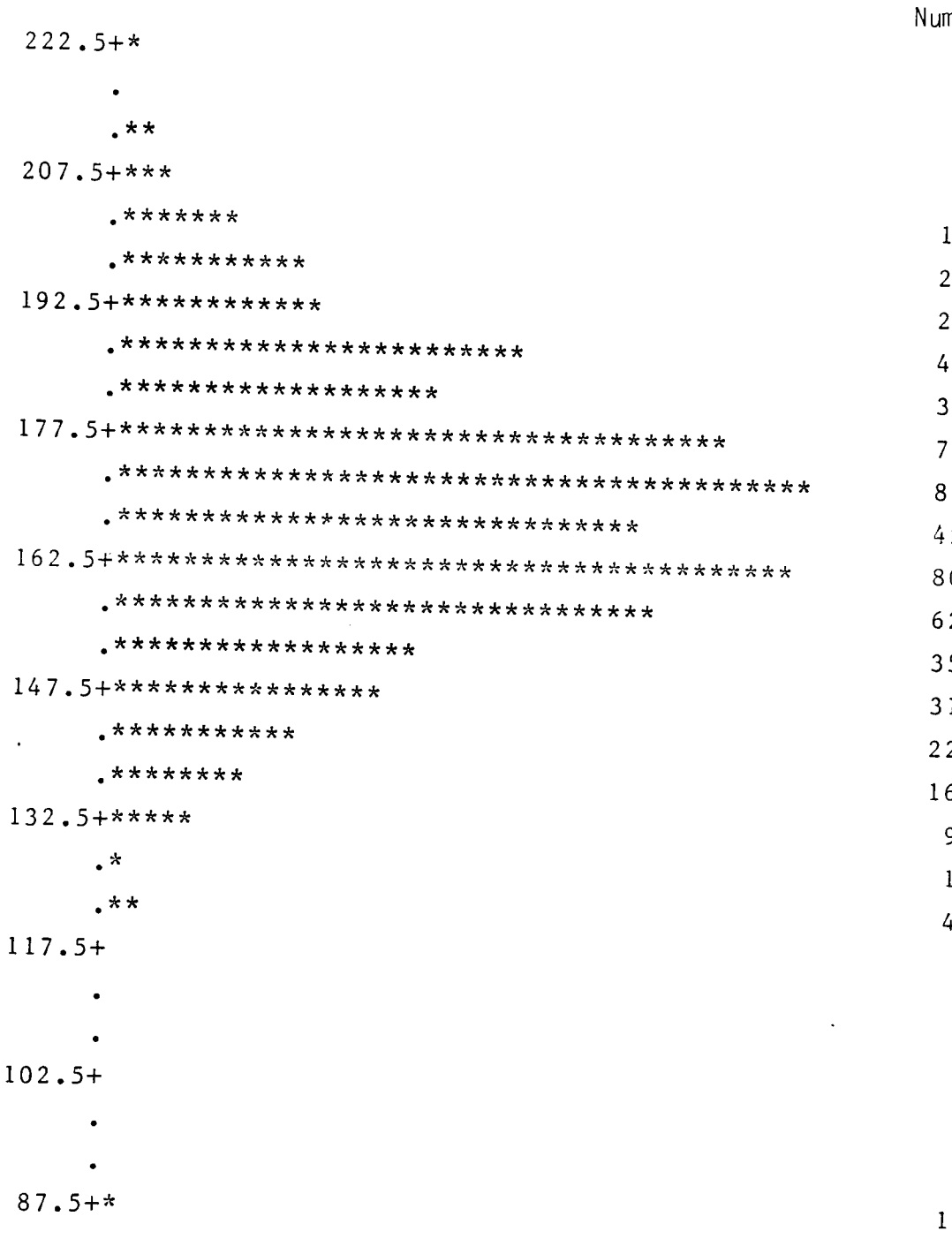


FIGURE 8

BAR CHART OF DISTRIBUTION OF SCALE SCORES:
READING, FIFTH GRADE, 1983

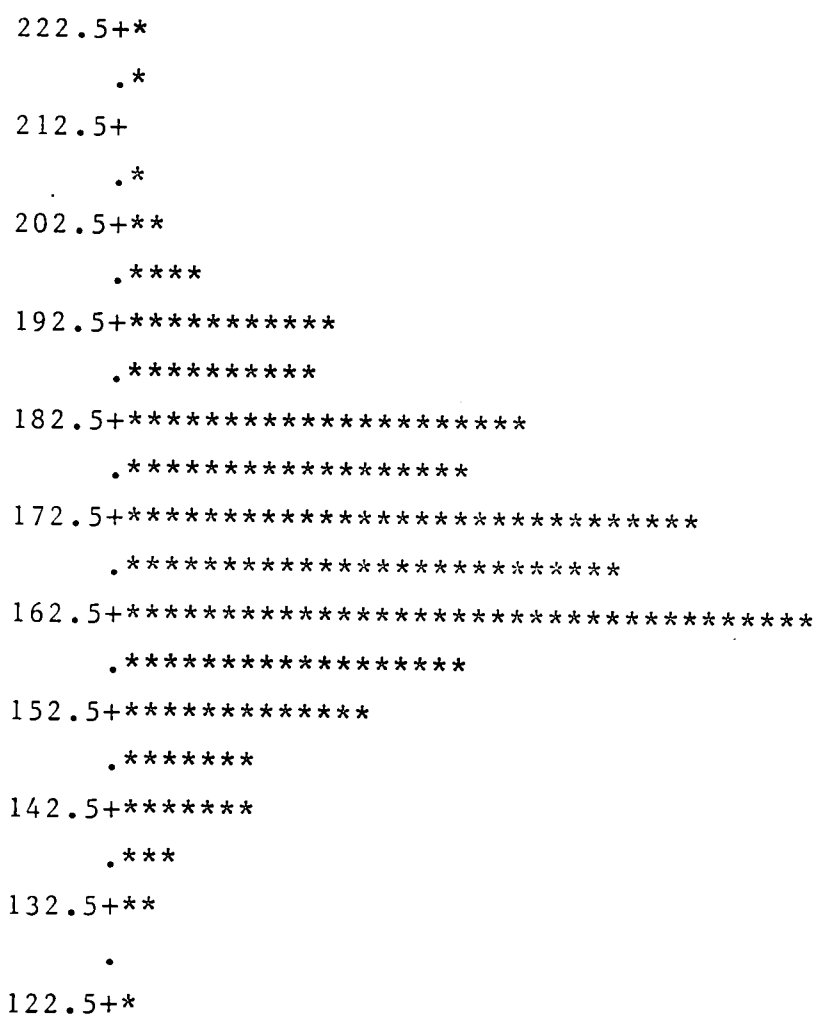


FIGURE 9

BAR CHART OF DISTRIBUTION OF SCALE SCORES:
MATH, FIFTH GRADE, 1983

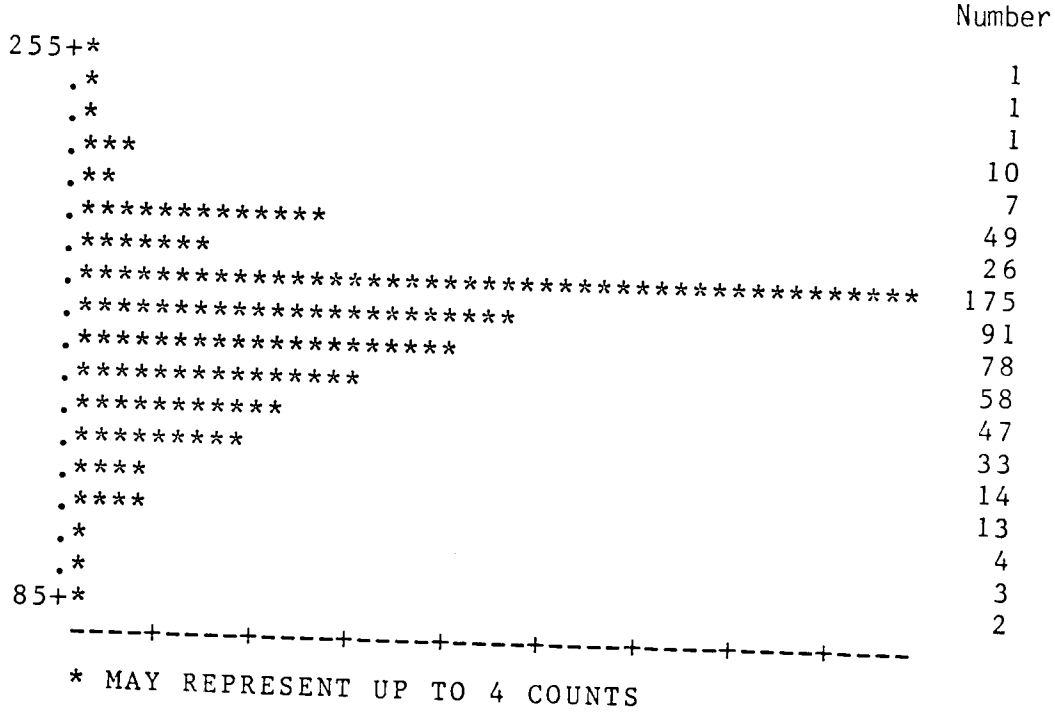


FIGURE 10

BAR CHART OF DISTRIBUTION OF SCALE SCORES:
LANGUAGE ARTS, FIFTH GRADE, 1983

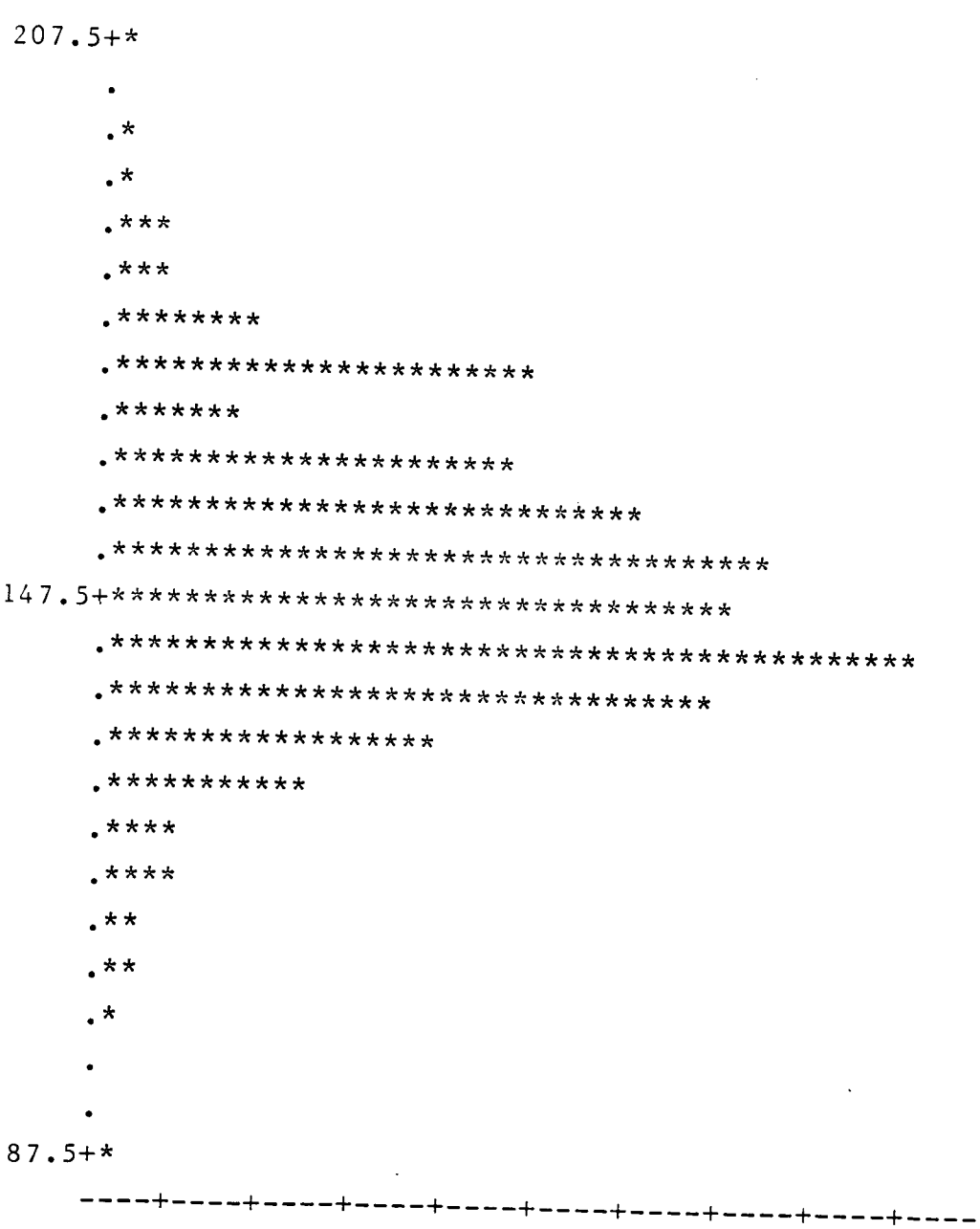


FIGURE 11

BAR CHART OF DISTRIBUTION OF SCALE SCORES:
READING, THIRD GRADE, 1982

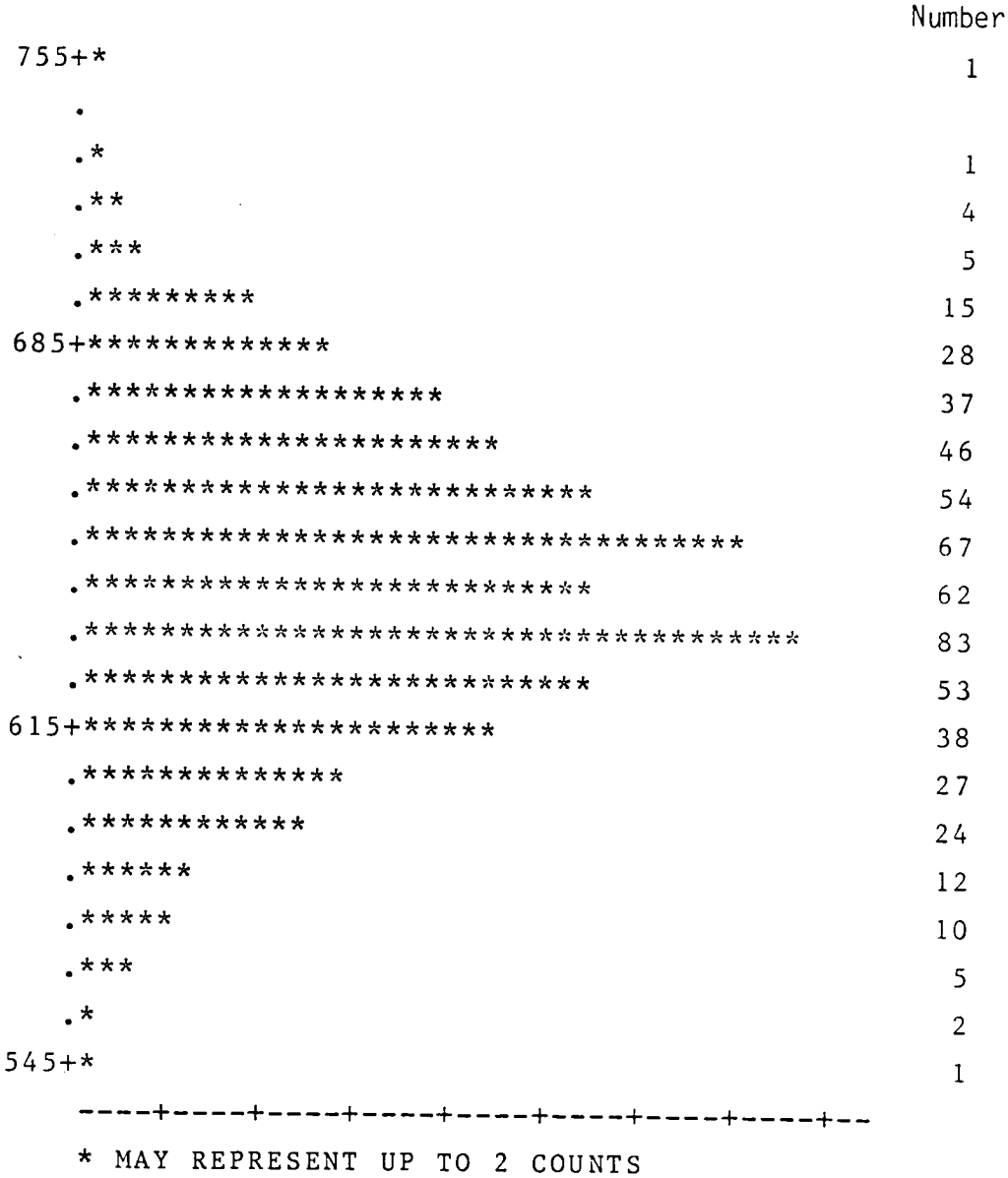


FIGURE 12

BAR CHART OF DISTRIBUTION OF SCALE SCORES:
READING, FIFTH GRADE, 1984

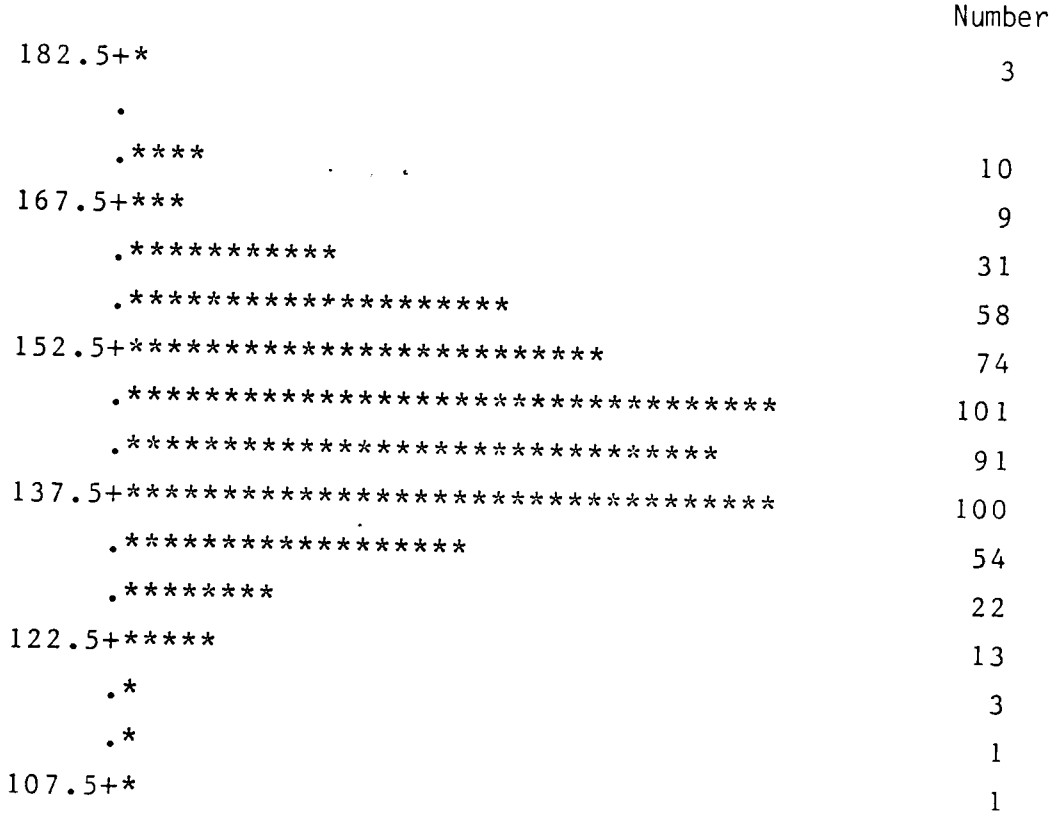


FIGURE 13

BAR CHART OF DISTRIBUTION OF SCALE SCORES:
MATH, THIRD GRADE, 1982

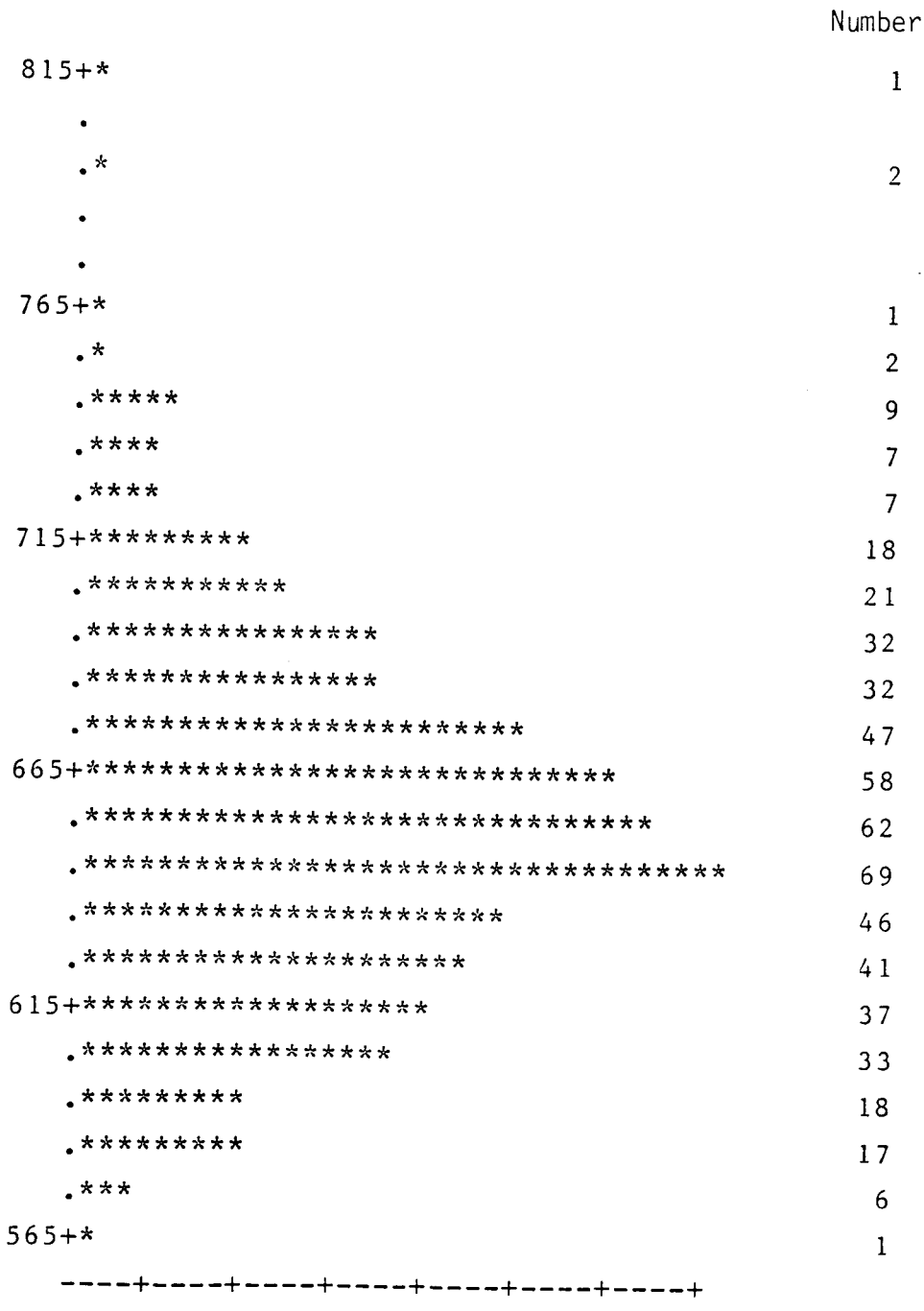


FIGURE 14

BAR CHART OF DISTRIBUTION OF SCALE SCORES:
MATH, FIFTH GRADE, 1984

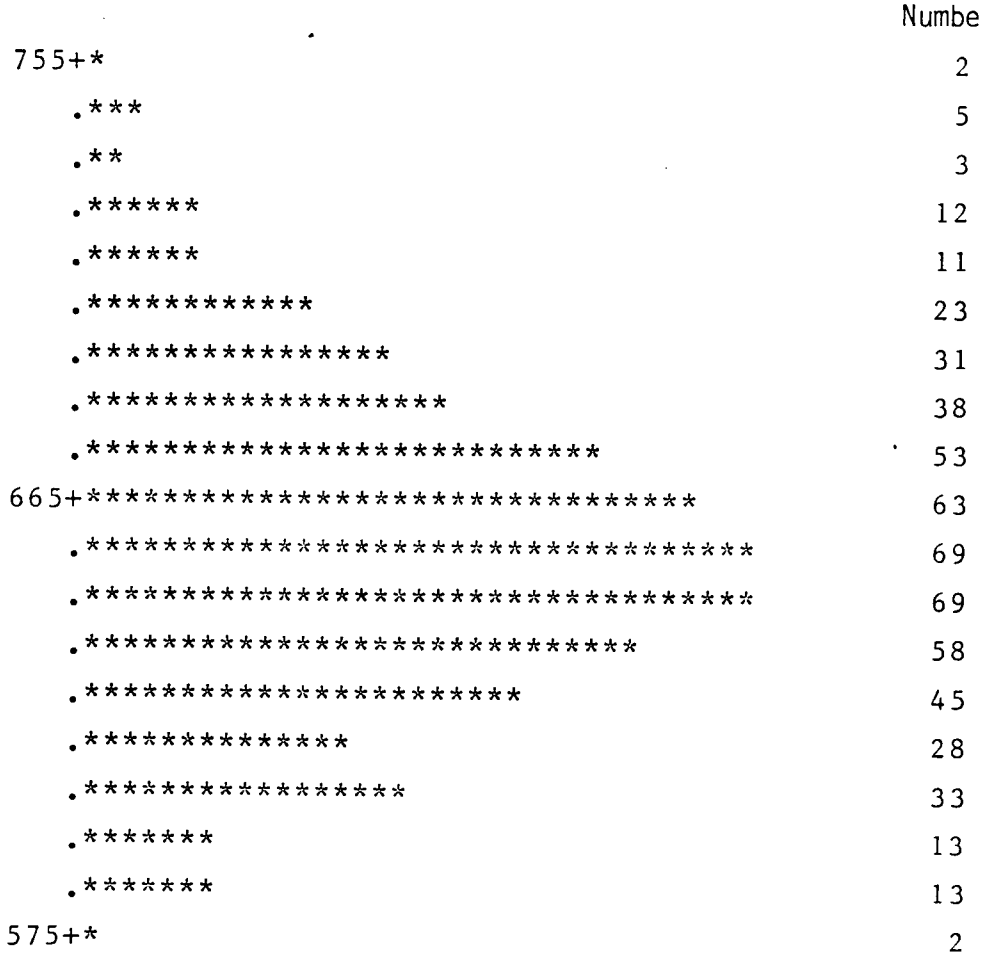


FIGURE 15

BAR CHART OF DISTRIBUTION OF SCALE SCORES:
LANGUAGE ARTS, FIFTH GRADE, 1984

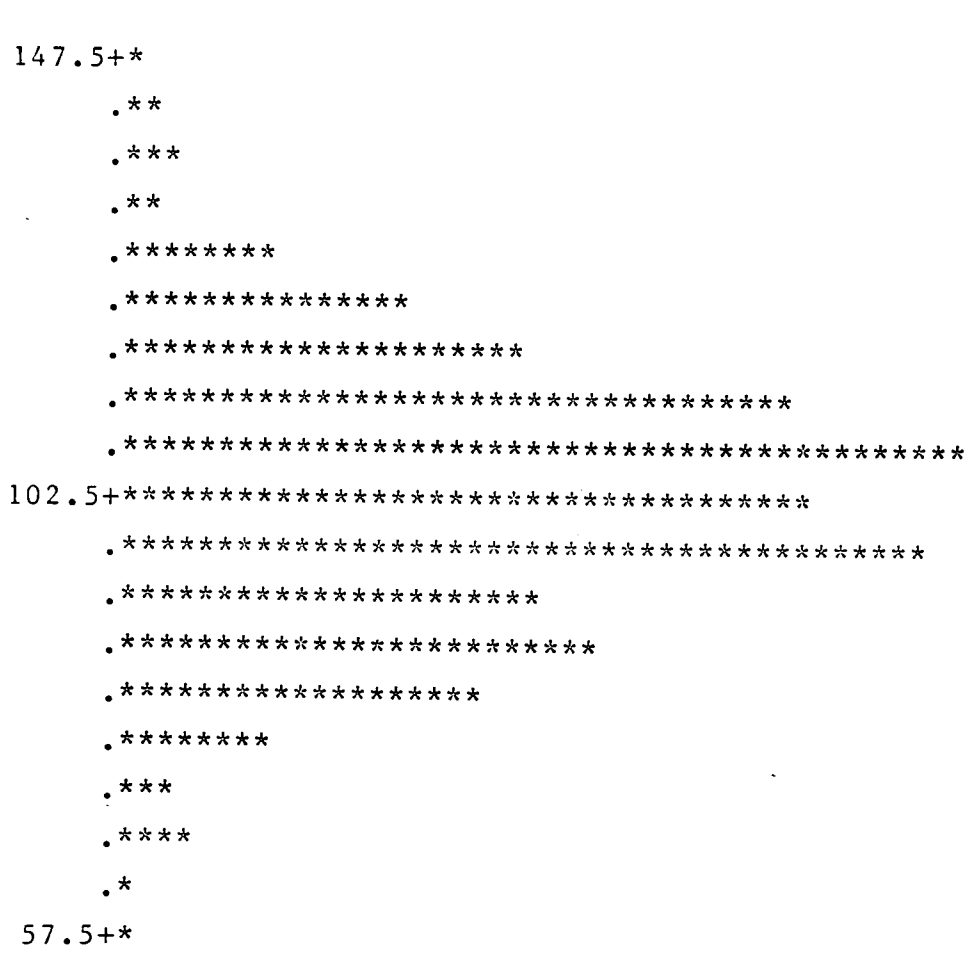


FIGURE 16

BAR CHART OF DISTRIBUTION OF SCALE SCORES:
SCHOOL ABILITY, FOURTH GRADE, 1983

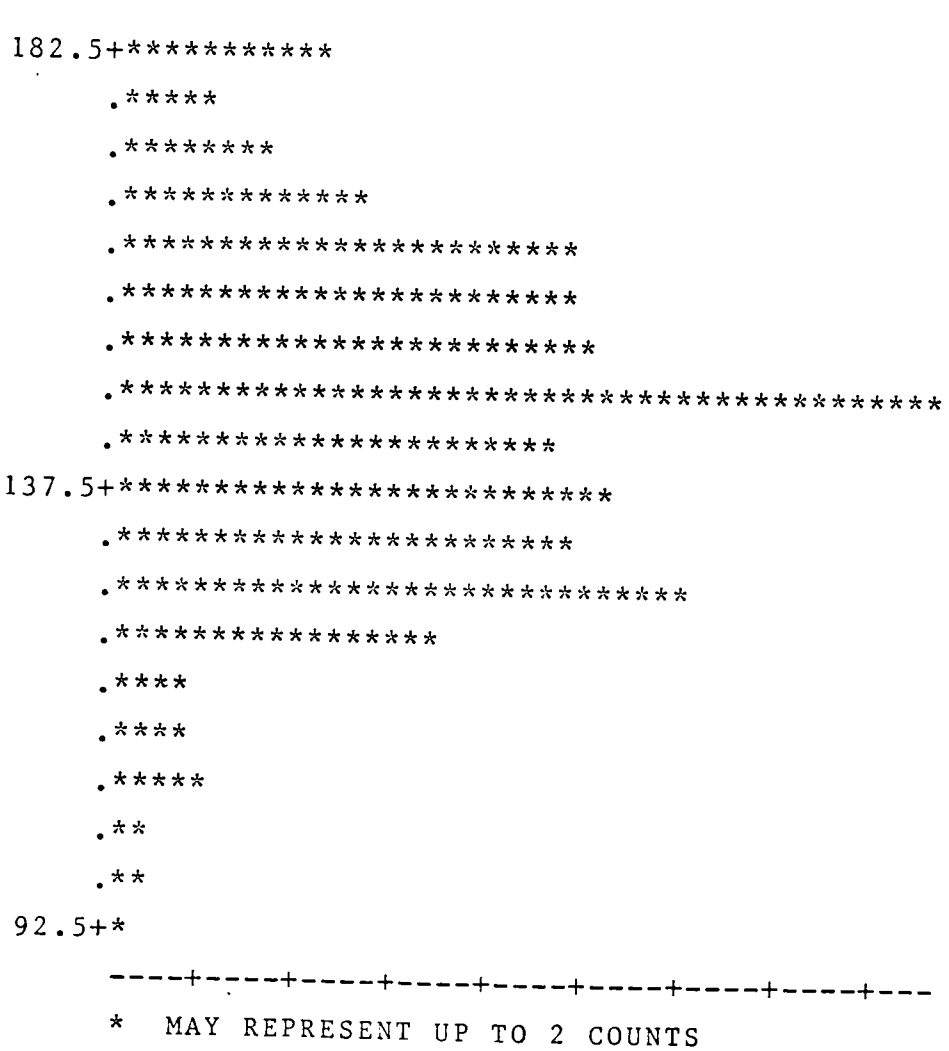


FIGURE 17

BAR CHART OF DISTRIBUTION OF SCALE SCORES:
LANGUAGE ARTS, THIRD GRADE, 1982

For further clarification and reverification of the distribution patterns of the various data sets, frequency tables were also developed of the distribution patterns. The first set of tables provided frequency distribution for third grade achievement 1981, school ability test data 1982, and fifth grade achievement 1983. An examination of the lower and upper distribution of each table demonstrated no serious problem with either ceiling or floor effect, with only one exception. Normal distributions were found in the following: reading third grade, 1981 (Table III); reading fifth grade, 1983 (Table IV); math third grade, 1981 (Table V); math fifth grade, 1983 (Table VI); language arts fifth grade, 1983 (Table VII); and school ability fourth grade, 1982 (Table VIII). The frequency count in language arts third grade (Table IX) exhibited a heavier number of students at the top end of the distribution. Approximately 4 percent of the population appeared at the top of the distribution, evidence of a slight ceiling effect.

Tables X-XVI provided frequency distribution for third grade achievement in 1982, school ability test data in 1983, and fifth grade achievement in 1984. An examination of these frequency tables results in the same conclusion as that of the first set of data; language arts third grade exhibits some evidence of a slight ceiling effect (Table XIV). Language arts third grade has approximately 3 percent of the population located at the top of the distribution. Also an examination of the count column shows a heavier number of students clustering at the upper portion of the distribution. The ceiling effect was not as

TABLE III

FREQUENCY TABLE OF SCALE SCORES:
READING, THIRD GRADE, 1981

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
103	1	0.2	0.2
107	3	0.5	0.7
109	1	0.2	0.8
112	2	0.3	1.1
115	1	0.2	1.3
116	4	0.7	2.0
118	2	0.3	2.3
119	5	0.8	3.1
120	1	0.2	3.3
124	5	0.8	4.1
125	3	0.5	4.6
126	1	0.2	4.7
127	6	1.0	5.7
128	2	0.3	6.0
129	7	1.1	7.2
131	6	1.0	8.1
132	6	1.0	9.1
133	18	2.9	12.0
134	24	3.9	15.9
135	4	0.7	16.6
137	9	1.5	18.0
138	39	6.3	24.4
139	22	3.6	28.0
140	11	1.8	29.8
141	8	1.3	31.1
142	20	3.3	34.3
143	24	3.9	38.2
144	15	2.4	40.7
145	10	1.6	42.3
146	22	3.6	45.9
148	20	3.3	49.1
149	13	2.1	51.2
150	13	2.1	53.3
151	22	3.6	56.9
153	29	4.7	61.6
154	9	1.5	63.1
155	5	0.8	63.9
156	26	4.2	68.1
158	11	1.8	69.9

TABLE III (Continued)

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
160	12	2.0	74.8
162	16	2.6	77.4
164	35	5.7	83.1
166	8	1.3	84.4
167	11	1.8	86.2
170	18	2.9	89.1
172	26	4.2	93.3
174	9	1.5	94.8
176	13	2.1	96.9
179	6	1.0	97.9
181	6	1.0	98.9
185	1	0.2	99.0
192	3	0.5	99.5
198	2	0.3	99.8
233	1	0.2	100.0

TABLE IV

FREQUENCY TABLE OF SCALE SCORES:
READING, FIFTH GRADE, 1983

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
89	1	0.2	0.2
120	1	0.2	0.3
124	3	0.5	0.8
126	1	0.2	1.0
131	2	0.3	1.3
133	5	0.8	2.1
134	2	0.3	2.5
137	6	1.0	3.4
138	10	1.6	5.1
140	3	0.5	5.6
141	5	0.8	6.4
142	6	1.0	7.4
143	3	0.5	7.8
144	5	0.8	8.7
145	2	0.3	9.0
146	6	1.0	10.0
148	11	1.8	11.8
149	12	2.0	13.7
150	1	0.2	13.9
151	17	2.8	16.7
153	12	2.0	18.6
154	5	0.8	19.4
155	6	1.0	20.4
156	15	2.5	22.9
157	8	1.3	24.2
158	11	1.8	26.0
159	22	3.6	29.6
160	19	3.1	32.7
161	8	1.3	34.0
162	17	2.8	36.8
163	9	1.9	38.2
164	27	4.4	42.6
166	14	2.3	44.9
167	11	1.8	46.7
168	7	1.1	47.9
169	10	1.6	49.5
170	17	2.8	52.3
171	12	2.0	54.2
172	29	4.7	59.0

TABLE IV (Continued)

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
173	12	2.0	60.9
174	11	1.8	62.7
175	13	2.1	64.9
176	20	3.3	68.1
177	11	1.8	69.9
178	10	1.6	71.6
179	18	2.9	74.5
180	11	1.8	76.3
181	9	1.5	77.8
183	7	1.1	78.9
184	11	1.8	80.7
185	21	3.4	84.2
187	9	1.5	85.6
188	18	2.9	88.6
190	5	0.8	89.4
192	7	1.1	90.5
193	7	1.1	91.7
194	5	0.8	92.5
195	5	0.8	93.3
197	15	2.5	95.8
198	2	0.3	96.1
201	7	1.1	97.2
204	7	1.1	98.4
208	4	0.7	99.0
209	2	0.3	99.3
211	1	0.2	99.5
212	2	0.3	99.8
221	1	0.2	100.0

TABLE V

FREQUENCY TABLE OF SCALE SCORES:
MATH, THIRD GRADE, 1981

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
106	1	0.2	0.2
118	2	0.3	0.5
119	1	0.2	0.6
120	1	0.2	0.8
121	2	0.3	1.1
122	4	0.6	1.8
124	1	0.2	1.9
125	4	0.6	2.6
126	4	0.6	3.2
127	3	0.5	3.7
129	13	2.1	5.8
130	3	0.5	6.3
131	3	0.5	6.8
132	2	0.3	7.1
133	18	2.9	10.1
134	18	2.9	13.0
135	7	1.1	14.1
136	23	3.7	17.9
137	12	1.9	19.8
138	20	3.2	23.1
139	19	3.1	26.1
140	12	1.9	28.1
141	27	4.4	32.5
142	13	2.1	34.6
143	31	5.0	39.6
144	41	6.7	46.3
145	18	2.9	49.2
146	10	1.6	50.8
147	9	1.5	52.3
148	21	3.4	55.7
149	23	3.7	59.4
150	22	3.6	63.0
151	12	1.9	64.9
152	24	3.9	68.8
153	11	1.8	70.6
154	8	1.3	71.9
155	23	3.7	75.6
156	24	3.9	79.5
157	12	1.9	81.5

TABLE V (Continued)

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
158	8	1.3	82.8
159	13	2.1	84.9
160	13	2.1	87.0
161	20	3.2	90.3
163	9	1.5	91.7
164	7	1.1	92.9
166	7	1.1	94.0
167	7	1.1	95.1
168	11	1.8	96.9
170	8	1.3	98.2
172	2	0.3	98.5
173	2	0.3	98.9
176	3	0.5	99.4
180	2	0.3	99.7
183	1	0.2	99.8
187	1	0.2	100.0

TABLE VI

FREQUENCY TABLE OF SCALE SCORES:
MATH, FIFTH GRADE, 1983

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
123	2	0.3	0.3
130	2	0.3	0.7
131	1	0.2	0.8
134	1	0.2	1.0
135	1	0.2	1.1
137	1	0.2	1.3
138	4	0.7	2.0
139	2	0.3	2.3
140	3	0.5	2.8
141	8	1.3	4.1
142	3	0.5	4.6
144	6	1.0	5.5
145	8	1.3	6.8
147	8	1.3	8.1
148	1	0.2	8.3
149	4	0.7	9.0
150	9	1.5	10.4
151	5	0.8	11.2
152	4	0.7	11.9
153	5	0.8	12.7
154	14	2.3	15.0
155	10	1.6	16.6
156	25	4.1	20.7
157	10	1.6	22.3
159	9	1.5	23.8
160	15	2.4	26.2
161	46	7.5	33.7
162	13	2.1	35.8
163	6	1.0	36.8
164	27	4.4	41.2
165	4	0.7	41.9
166	19	3.1	45.0
167	28	4.6	49.5
168	19	3.1	52.6
169	8	1.3	53.9
170	14	2.3	56.2
171	24	3.9	60.1
172	30	4.9	65.0
173	22	3.6	68.6

TABLE VI (Continued)

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
175	13	2.1	70.7
176	12	2.0	72.6
178	16	2.6	75.2
179	11	1.8	77.0
180	21	3.4	80.5
181	16	2.6	83.1
182	6	1.0	84.0
183	5	0.8	84.9
184	13	2.1	87.0
186	9	1.5	88.4
187	5	0.8	89.3
188	14	2.3	91.5
190	8	1.3	92.8
191	6	1.0	93.8
192	7	1.1	95.0
193	6	1.0	95.9
194	6	1.0	96.9
196	7	1.1	98.0
198	1	0.2	98.2
199	2	0.3	98.5
201	1	0.2	98.7
203	3	0.5	99.2
208	1	0.2	99.3
209	2	0.3	99.7
215	1	0.2	99.8
220	1	0.2	100.0

TABLE VII

FREQUENCY TABLE OF SCALE SCORES: LANGUAGE ARTS, FIFTH GRADE, 1983

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
83	1	0.2	0.2
88	1	0.2	0.3
92	1	0.2	0.5
95	1	0.2	0.7
97	1	0.2	0.8
101	2	0.3	1.1
109	2	0.3	1.5
112	1	0.2	1.6
114	5	0.8	2.4
118	7	1.1	3.6
124	6	1.0	4.6
126	2	0.3	4.9
129	6	1.0	5.9
133	5	0.8	6.7
137	14	2.3	9.0
139	14	2.3	11.3
145	11	1.8	13.1
147	14	2.3	15.3
148	10	1.6	17.0
149	12	2.0	18.9
153	11	1.8	20.7
154	16	2.6	23.3
156	13	2.1	25.4
159	18	2.9	28.4
160	18	2.9	31.3
162	9	1.5	32.8
164	13	2.1	34.9
165	12	2.0	36.9
168	15	2.4	39.3
169	11	1.8	41.1
170	12	2.0	43.1
171	12	2.0	45.0
172	17	2.8	47.8
174	13	2.1	49.9
176	11	1.8	51.7
177	16	2.6	54.3
178	10	1.6	56.0
180	10	1.6	57.6

TABLE VII (Continued)

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
181	9	1.5	59.1
183	25	4.1	63.1
184	8	1.3	64.4
186	45	7.3	71.8
187	30	4.9	76.7
188	20	3.3	79.9
189	28	4.6	84.5
190	14	2.3	86.8
194	7	1.1	87.9
197	5	0.8	88.7
201	10	1.6	90.4
202	10	1.6	92.0
204	6	1.0	93.0
205	14	2.3	95.3
208	4	0.7	95.9
209	5	0.8	96.7
212	5	0.8	97.6
219	2	0.3	97.9
220	4	0.7	98.5
222	1	0.2	98.7
227	4	0.7	99.3
229	1	0.2	99.5
237	1	0.2	99.7
240	1	0.2	99.8
257	1	0.2	100.0

TABLE VIII

FREQUENCY TABLE OF SCALE SCORES: SCHOOL ABILITY, FOURTH GRADE, 1982

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
58	1	0.2	0.2
64	1	0.2	0.3
66	1	0.2	0.5
67	1	0.2	0.7
68	1	0.2	0.8
70	1	0.2	1.0
72	2	0.3	1.3
73	1	0.2	1.5
74	3	0.5	2.0
75	4	0.7	2.6
77	6	1.0	3.6
78	6	1.0	4.6
79	7	1.1	5.7
80	2	0.3	6.0
81	8	1.3	7.3
82	3	0.5	7.8
83	9	1.5	9.3
84	13	2.1	11.4
85	8	1.3	12.7
86	8	1.3	14.0
87	9	1.5	15.5
88	17	2.8	18.2
89	14	2.3	20.5
90	15	2.4	23.0
91	16	2.6	25.6
92	16	2.6	28.2
93	11	1.8	30.0
94	13	2.1	32.1
95	13	2.1	34.2
96	16	2.6	36.8
97	9	1.5	38.3
98	13	2.1	40.4
99	18	2.9	43.3
100	16	2.6	45.9
101	10	1.6	47.6
102	16	2.6	50.2
103	19	3.1	53.3
104	11	1.8	55.0

TABLE VIII (Continued)

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
105	22	3.6	58.6
106	19	3.1	61.7
107	18	2.9	64.7
108	11	1.8	66.4
109	19	3.1	69.5
110	14	2.3	71.8
111	14	2.3	74.1
112	16	2.6	76.7
113	20	3.3	80.0
114	6	1.0	80.9
115	5	0.8	81.8
116	10	1.6	83.4
117	12	2.0	85.3
118	11	1.8	87.1
119	4	0.7	87.8
120	8	1.3	89.1
121	8	1.3	90.4
122	9	1.5	91.9
123	4	0.7	92.5
124	10	1.6	94.1
125	2	0.3	94.5
127	6	1.0	95.4
129	9	1.5	96.9
131	2	0.3	97.2
132	4	0.7	97.9
133	1	0.2	98.0
134	1	0.2	98.2
135	1	0.2	98.4
137	2	0.3	98.7
138	2	0.3	99.0
139	1	0.2	99.2
140	1	0.2	99.3
141	2	0.3	99.7
145	1	0.2	99.8
150	1	0.2	100.0

TABLE IX

FREQUENCY TABLE OF SCALE SCORES: LANGUAGE ARTS, THIRD GRADE, 1981

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
88	4	0.6	0.6
92	1	0.2	0.8
97	2	0.3	1.1
101	3	0.5	1.6
105	6	1.0	2.6
109	6	1.0	3.6
112	4	0.6	4.2
114	6	1.0	5.2
118	4	0.6	5.8
120	7	1.1	6.9
122	7	1.1	8.1
124	9	1.5	9.5
125	15	2.4	12.0
126	16	2.6	14.5
128	15	2.4	17.0
129	15	2.4	19.4
131	14	2.3	21.6
133	25	4.0	25.7
134	15	2.4	28.1
136	12	1.9	30.0
137	13	2.1	32.1
139	24	3.9	36.0
141	16	2.6	38.6
143	22	3.6	42.2
145	24	3.9	46.0
147	23	3.7	49.8
148	21	3.4	53.2
149	24	3.9	57.0
151	21	3.4	60.4
153	20	3.2	63.7
155	17	2.7	66.4
158	20	3.2	69.6
160	26	4.2	73.8
162	26	4.2	78.0
164	21	3.4	81.4
167	26	4.2	85.6
169	19	3.1	88.7
172	18	2.9	91.6

TABLE IX (Continued)

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
176	15	2.4	94.0
180	12	1.9	96.0
182	25	4.0	100.0

TABLE X

FREQUENCY TABLE OF SCALE SCORES:
READING, THIRD GRADE, 1982

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
87	1	0.2	0.2
104	1	0.2	0.3
107	1	0.2	0.5
109	2	0.3	0.9
111	1	0.2	1.0
112	1	0.2	1.2
114	1	0.2	1.4
116	1	0.2	1.6
117	3	0.5	2.1
118	1	0.2	2.3
119	2	0.3	2.6
120	2	0.3	3.0
124	6	1.0	4.0
125	2	0.3	4.3
126	3	0.5	4.9
127	6	1.0	5.9
128	3	0.5	6.4
129	7	1.2	7.7
131	1	0.2	7.8
132	7	1.2	9.0
133	15	2.6	11.7
134	12	2.1	13.7
135	3	0.5	14.3
137	12	2.1	16.3
138	35	6.1	22.4
139	16	2.8	25.2
140	17	3.0	28.2
141	5	0.9	29.0
142	23	4.0	33.0
143	25	4.3	47.4
144	17	3.0	40.3
145	14	2.4	42.8
146	26	4.5	47.3
148	20	3.5	50.8
149	7	1.2	52.0
150	18	3.1	55.1
151	25	4.3	59.5
153	31	5.4	64.9
154	10	1.7	66.6

TABLE X (Continued)

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
155	8	1.4	68.0
156	32	5.6	73.6
158	8	1.4	75.0
159	10	1.7	76.7
160	12	2.1	78.8
162	10	1.7	80.5
164	22	3.8	84.3
166	5	0.9	85.2
167	9	1.6	86.8
170	9	1.6	88.3
172	24	4.2	92.5
174	13	2.3	94.8
176	10	1.7	96.5
179	6	1.0	97.6
181	6	1.0	98.6
185	5	0.9	99.5
192	1	0.2	99.7
198	1	0.2	99.8
208	1	0.2	100.0

TABLE XI

FREQUENCY TABLE OF SCALE SCORES:
READING, FIFTH GRADE, 1984

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
546	1	0.2	0.2
557	1	0.2	0.3
559	1	0.2	0.5
561	2	0.3	0.9
563	1	0.2	1.0
565	1	0.2	1.2
567	1	0.2	1.4
571	1	0.2	1.6
573	1	0.2	1.7
576	2	0.3	2.1
578	3	0.5	2.6
579	3	0.5	3.1
581	3	0.5	3.7
583	1	0.2	3.8
584	4	0.7	4.5
586	3	0.5	5.0
587	1	0.2	5.2
590	2	0.3	5.6
592	5	0.9	6.4
593	4	0.7	7.1
595	2	0.3	7.5
596	2	0.3	7.8
597	1	0.2	8.0
598	7	1.2	9.2
599	1	0.2	9.4
601	7	1.2	10.6
602	4	0.7	11.3
603	3	0.5	11.8
605	2	0.3	12.2
606	6	1.0	13.2
608	3	0.5	13.7
609	2	0.3	14.1
612	4	0.7	14.8
613	6	1.0	15.8
615	7	1.2	17.0
616	6	1.0	18.1
617	5	0.9	19.0
619	10	1.7	20.7
620	7	1.2	21.9

TABLE XI (Continued)

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
622	8	1.4	23.3
623	10	1.7	25.0
624	9	1.6	26.6
626	4	0.7	27.3
627	5	0.9	28.2
629	10	1.7	29.9
630	13	2.3	32.2
632	6	1.0	33.2
633	14	2.4	35.7
634	10	1.7	37.4
636	13	2.3	39.7
637	14	2.4	42.1
639	13	2.3	44.3
640	13	2.3	46.6
642	10	1.7	48.3
643	12	2.1	50.4
645	8	1.4	51.8
646	11	1.9	53.7
648	8	1.4	55.1
650	11	1.9	57.0
651	11	1.9	59.0
653	14	2.4	61.4
655	11	1.9	63.3
656	9	1.6	64.9
658	11	1.9	66.8
660	9	1.6	68.3
661	11	1.9	70.3
663	8	1.4	71.7
664	8	1.4	73.0
667	14	2.4	75.5
669	4	0.7	76.2
673	12	2.1	80.0
675	15	2.6	82.6
677	9	1.6	84.2
680	3	0.5	84.7
682	11	1.9	86.6
684	8	1.4	88.0
687	5	0.9	88.9
689	10	1.7	90.6
692	7	1.2	91.8
695	12	2.1	93.9

TABLE XI (Continued)

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
698	9	1.6	95.5
701	7	1.2	96.7
704	3	0.5	97.2
708	5	0.9	98.1
712	4	0.7	98.8
716	1	0.2	99.0
721	3	0.5	99.5
726	1	0.2	99.7
739	1	0.2	99.8
758	1	0.2	100.0

TABLE XII

FREQUENCY TABLE OF SCALE SCORES:
MATH, THIRD GRADE, 1982

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
106	1	0.2	0.2
114	1	0.2	0.4
118	2	0.4	0.7
119	1	0.2	0.9
120	4	0.7	1.6
121	3	0.5	2.1
124	6	1.1	3.2
125	3	0.5	3.7
126	3	0.5	4.2
127	3	0.5	4.7
129	13	2.3	7.0
130	6	1.1	8.1
131	6	1.1	9.1
132	10	1.8	10.9
133	18	3.2	14.0
134	14	2.5	16.5
135	10	1.8	18.2
136	17	3.0	21.2
137	23	4.0	25.2
138	14	2.5	27.7
139	36	6.3	34.0
140	10	1.8	35.7
141	15	2.6	38.4
142	6	1.1	39.4
143	24	4.2	43.6
144	36	6.3	49.9
145	27	4.7	54.6
146	14	2.5	57.1
147	19	3.3	60.4
148	14	2.5	62.9
149	27	4.7	67.6
150	17	3.0	70.6
151	10	1.8	72.3
152	21	3.7	76.0
153	12	2.1	78.1
154	14	2.5	80.6
155	20	3.5	84.1
156	19	3.3	87.4
157	7	1.2	88.6

TABLE XII (Continued)

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
158	5	0.9	89.5
159	7	1.2	90.7
160	9	1.6	92.3
161	14	2.5	94.7
163	5	0.9	95.6
164	3	0.5	96.1
166	1	0.2	96.3
167	4	0.7	97.0
168	4	0.7	97.7
170	5	0.9	98.6
172	1	0.2	98.8
173	4	0.7	99.5
180	2	0.4	99.8
183	1	0.2	100.0

TABLE XIII

FREQUENCY TABLE OF SCALE SCORES:
MATH, FIFTH GRADE, 1984

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
564	1	0.2	0.2
575	3	0.5	0.7
578	3	0.5	1.2
580	3	0.5	1.8
582	5	0.9	2.6
585	1	0.2	2.8
587	4	0.7	3.5
589	4	0.7	4.2
591	4	0.7	4.9
592	2	0.4	5.3
594	3	0.5	5.8
596	2	0.4	6.2
597	5	0.9	7.1
599	2	0.4	7.4
601	3	0.5	7.9
602	8	1.4	9.3
604	9	1.6	10.9
605	7	1.2	12.2
607	2	0.4	12.5
608	4	0.7	13.2
610	5	0.9	14.1
611	5	0.9	15.0
613	5	0.9	15.9
614	6	1.1	16.9
616	2	0.4	17.3
617	10	1.8	19.0
619	4	0.7	19.8
620	4	0.7	20.5
621	8	1.4	21.9
623	6	1.1	22.9
624	6	1.1	24.0
626	4	0.7	24.7
627	5	0.9	25.0
628	8	1.4	27.8
630	6	1.1	28.0
631	9	1.6	29.6
633	9	1.6	31.2
634	2	0.4	31.6
635	6	1.1	32.6

TABLE XIII (Continued)

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
637	6	1.1	33.7
638	8	1.4	35.1
640	6	1.1	36.2
641	12	2.1	38.3
642	9	1.6	39.9
644	10	1.8	41.6
645	13	2.3	43.9
647	10	1.8	45.7
648	9	1.6	47.3
650	4	0.7	48.0
651	13	2.3	50.3
653	3	0.5	50.8
654	17	3.0	53.8
656	10	1.8	55.6
657	9	1.6	57.1
659	6	1.1	58.2
660	4	0.7	58.9
662	9	1.6	60.5
663	13	2.3	62.8
665	11	1.9	64.7
667	13	2.3	67.0
668	8	1.4	68.4
670	5	0.9	69.3
672	14	2.5	71.8
673	9	1.6	73.4
675	8	1.4	74.8
677	4	0.7	75.5
679	7	1.2	76.7
681	4	0.7	77.4
682	10	1.8	79.2
684	6	1.1	80.2
686	8	1.4	81.7
689	4	0.7	82.4
691	7	1.2	83.6
693	6	1.1	84.7
695	9	1.6	86.2
697	10	1.8	88.0
700	7	1.2	89.2
702	4	0.7	89.9
705	7	1.2	91.2
708	3	0.5	91.7
711	7	1.2	92.9

TABLE XIII (Continued)

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
714	6	1.1	94.0
717	5	0.9	94.9
720	2	0.4	95.2
724	4	0.7	95.9
728	1	0.2	96.1
732	5	0.9	97.0
737	2	0.4	97.4
743	5	0.9	98.2
749	4	0.7	98.9
756	2	0.4	99.3
764	1	0.2	99.5
790	2	0.4	99.8
815	1	0.2	100.0

TABLE XIV

FREQUENCY TABLE OF SCALE SCORES: LANGUAGE ARTS, THIRD GRADE, 1982

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
92	1	0.2	0.2
97	3	0.5	0.7
101	4	0.7	1.4
105	5	0.9	2.3
109	5	0.9	3.2
112	2	0.4	3.5
114	5	0.9	4.4
118	7	1.2	5.6
120	7	1.2	6.8
122	14	2.5	9.3
124	13	2.3	11.6
125	14	2.5	14.0
126	14	2.5	16.5
128	14	2.5	18.9
129	17	3.0	21.9
131	16	2.8	24.7
133	15	2.6	27.3
134	16	2.8	30.1
136	12	2.1	32.2
137	16	2.8	35.0
139	23	4.0	39.1
141	16	2.8	41.9
143	29	5.1	46.9
145	19	3.3	50.3
147	20	3.5	53.8
148	22	3.9	57.6
149	24	4.2	61.8
151	27	4.7	66.5
153	23	4.0	70.6
155	24	4.2	74.8
158	23	4.0	78.8
160	21	3.7	82.5
162	12	2.1	84.6
164	15	2.6	87.2
167	14	2.5	89.7
169	12	2.1	91.8
172	15	2.6	94.4
176	10	1.8	96.1
180	8	1.4	97.5
182	14	2.5	100.0

TABLE XV

FREQUENCY TABLE OF SCALE SCORES: LANGUAGE ARTS, FIFTH GRADE, 1984

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
575	2	0.4	0.4
580	1	0.2	0.5
582	2	0.4	8.9
584	2	0.4	1.2
586	4	0.7	1.9
588	4	0.7	2.6
590	3	0.5	3.2
592	4	0.7	3.9
594	1	0.2	4.0
595	2	0.4	4.4
597	1	0.2	4.6
599	2	0.4	4.9
601	4	0.7	5.6
603	6	1.1	6.7
604	7	1.2	7.9
606	6	1.1	8.9
608	4	0.7	9.6
609	6	1.1	10.7
611	6	1.1	11.7
613	6	1.1	12.8
614	2	0.4	13.1
616	7	1.2	14.4
617	4	0.7	15.1
619	3	0.5	15.6
620	3	0.5	16.1
622	8	1.4	17.5
624	6	1.1	18.6
625	11	1.9	20.5
627	6	1.1	21.5
628	11	1.9	23.5
630	7	1.2	24.7
631	7	1.2	25.9
633	8	1.4	27.3
635	6	1.1	28.4
636	15	2.6	31.0
638	7	1.2	32.2
639	8	1.4	33.6
641	11	1.9	35.6
642	14	2.5	38.0

TABLE XV (Continued)

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
644	6	1.1	39.1
646	12	2.1	41.2
647	14	2.5	43.6
649	12	2.1	45.7
651	14	2.5	48.2
652	10	1.8	49.9
654	8	1.4	51.3
656	11	1.9	53.2
657	12	2.1	55.3
659	14	2.5	57.8
661	8	1.4	59.2
663	16	2.8	62.0
665	12	2.1	64.1
666	14	2.5	66.5
668	13	2.3	68.8
670	13	2.3	71.1
672	12	2.1	73.2
674	9	1.6	74.8
677	10	1.8	76.5
679	9	1.6	78.1
681	11	1.9	80.0
683	6	1.1	81.1
686	10	1.8	82.8
688	11	1.9	84.8
691	7	1.2	86.0
694	13	2.3	88.3
696	4	0.7	89.0
699	7	1.2	90.2
703	6	1.1	91.2
706	8	1.4	92.6
709	9	1.6	94.2
713	7	1.2	95.4
718	4	0.7	96.1
722	4	0.7	96.8
728	8	1.4	98.2
733	3	0.5	98.8
740	1	0.2	98.9
749	4	0.7	99.6
759	3	0.4	100.0

TABLE XVI

FREQUENCY TABLE OF SCALE SCORES: SCHOOL ABILITY, FOURTH GRADE, 1983

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
58	2	0.3	0.3
63	1	0.2	0.5
66	2	0.3	0.9
67	1	0.2	1.0
68	2	0.3	1.4
69	3	0.5	1.9
70	2	0.3	2.3
71	1	0.2	2.4
72	1	0.2	2.6
73	1	0.2	2.8
75	3	0.5	3.3
76	4	0.7	4.0
77	3	0.5	4.5
78	2	0.3	4.9
79	4	0.7	5.6
80	8	1.4	7.0
81	6	1.0	8.0
82	6	1.0	9.0
83	6	1.0	10.1
84	11	1.9	12.0
85	6	1.0	13.0
86	12	2.1	15.1
87	7	1.2	16.3
88	10	1.7	18.1
89	14	2.4	20.5
90	10	1.7	22.3
91	15	2.6	24.9
92	6	1.0	25.9
93	7	1.2	27.1
94	6	1.0	28.2
95	10	1.7	29.9
96	19	3.3	33.2
97	15	2.6	35.8
98	25	4.3	40.2
99	15	2.6	42.8
100	11	1.9	44.7
101	15	2.6	47.3
102	17	3.0	50.3
103	14	2.4	52.7

TABLE XVI (Continued)

VALUE (Scale Score)	COUNT (Frequency)	PERCENTS	
		Individual	Cumulative
104	15	2.6	55.3
105	12	2.1	57.4
106	16	2.8	60.2
107	19	3.3	63.5
108	17	3.0	66.4
109	23	4.0	70.4
110	10	1.7	72.2
111	12	2.1	74.3
112	19	3.3	77.6
113	18	3.1	80.7
114	11	1.9	82.6
115	15	2.6	85.2
116	9	1.6	86.8
117	6	1.0	87.8
118	10	1.7	89.6
119	2	0.3	89.9
120	6	1.0	91.0
121	8	1.4	92.3
122	10	1.7	94.1
123	2	0.3	94.4
124	4	0.7	95.1
125	3	0.5	95.7
126	3	0.5	96.2
127	7	1.2	97.4
128	2	0.3	97.7
131	1	0.2	97.9
132	2	0.3	98.3
133	1	0.2	98.4
135	3	0.5	99.0
136	2	0.3	99.3
141	1	0.2	99.5
142	1	0.2	99.7
144	1	0.2	99.8
148	1	0.2	100.0

pronounced in this case as that demonstrated by language--third grade in 1981. There was evidence of a slight ceiling effect with language--third grade in both data sets.

This section on distribution diagnostics was undertaken to determine if the test data sets within this study exhibited the property of normal distribution. Normal distribution of test scores would be desired if one wished to evaluate teachers from standardized test scores. Normal distribution patterns were found in all test score data distributions with the exception of the two language arts--third grade data groups (1981 and 1982). A slight ceiling effect was identified within both data sets. Because the ceiling effect was identified in only adjustment data and was of such small magnitude, however, it was not considered significant by this researcher.

Data Adjustments

One of the major criticisms of using standardized achievement test data in the evaluation of teachers has been the fact that students enter a classroom with different abilities and with different achievement levels due to the non-random grouping characteristic of most public school classrooms. In designing a teacher evaluation system based upon achievement test data, it was necessary to make an adjustment of scores due to this nonrandom grouping aspect. The purpose of this section was to present the unadjusted and adjusted mean scores for fifth grade reading, math, and language arts per teacher for 1983 and 1984 school years. The analysis of covariance was selected as

the statistical control measure. Covariance is a form of analysis of variance and is a statistical procedure used to equate groups on one or more variables. By using this technique, an attempt was made to reduce variation in fifth grade achievement scores (reading/math/language arts) which is attributable to prior knowledge or ability, in this study as a measure of fourth grade school ability and third grade achievement (reading/math/language arts). The particular model for this analysis was presented and discussed in Chapter III, page 62.

A series of tables was created to illustrate the unadjusted and adjusted means for the teachers in this study. (Teachers are identified by numerical designation only.) The following tables provided the unadjusted and adjusted means for the three subject areas per teacher: Table XVII, reading--fifth grade, 1983; Table XVIII, math--fifth grade, 1983; Table XIX, language arts--fifth grade, 1983; Table XX, reading--fifth grade, 1984; Table XXI, math--fifth grade, 1984; and Table XXII, language arts--fifth grade, 1984. A quick comparison of the prior means and adjusted mean can be obtained by comparing the prior mean column (unadjusted means) with the mean in the last column (adjusted means). (For example, in the case of reading--fifth grade, 1983, Table XVII gives teacher 1 a mean score of 171.89. In comparison after adjustment teacher 1 received a mean score of 173.80).

TABLE XVII

READING, FIFTH GRADE, UNADJUSTED AND ADJUSTED MEAN SCORES: 1983

Teacher Identification Number	Number of Students	Unadjusted Prior Mean Reading 5	Reading 3	School Ability 4	Adjusted Mean
1	18	171.888889	149.333333	98.500000	173.799669
2	18	170.388889	151.000000	99.000000	171.172059
3	21	152.095238	139.523810	94.142857	161.135317
4	17	172.176471	151.235294	103.176471	171.370624
9	51	170.392157	150.196078	102.941176	170.262778
11	21	172.857143	145.523810	101.142857	176.026987
12	11	167.636364	140.545455	95.090909	175.761817
14	22	167.545455	150.818182	105.045455	166.327384
20	19	165.052632	152.263158	100.947368	166.327384
21	22	182.090909	170.272727	115.318182	166.165644
22	19	154.526316	143.473684	99.052632	159.596922
24	39	165.871795	149.282051	101.000000	166.941318
25	19	176.578947	154.894737	107.473684	172.183100
27	19	180.052632	157.684211	106.526316	174.390851
28	22	177.500000	156.681818	108.318182	171.787687
31	9	160.111111	139.000000	94.222222	159.423220
32	22	162.545455	144.045455	100.590909	166.753267
33	18	165.888889	145.388889	100.777770	169.263055
34	22	175.227273	150.090909	101.136364	175.786566
35	25	164.720000	145.120000	100.360000	168.393482
36	22	168.636364	155.681818	106.272727	164.208450
37	10	160.800000	136.200000	89.700000	173.288784
38	11	175.545455	161.636364	113.545455	165.178337
39	22	178.227273	153.818182	104.590909	175.451209

TABLE XVIII

MATH, FIFTH GRADE, UNADJUSTED AND ADJUSTED MEAN SCORES: 1983

Teacher Identification Number	Number of Students	Unadjusted Prior Mean Math 5	Math 5	School Ability 4	Adjusted Mean
1	18	174.166667	140.555556	98.500000	175.881267
2	18	170.888889	149.500000	99.000000	170.497761
3	21	158.142857	137.333333	94.142857	167.360123
4	17	176.176471	151.705882	103.176471	172.971834
8	51	165.078431	148.745098	103.294118	163.789514
11	21	167.761905	146.190476	101.142857	168.858265
12	49	162.551020	145.000000	98.612245	165.255717
16	23	173.608696	147.478261	103.739130	173.011155
17	70	170.000000	147.057143	105.557143	169.089352
20	14	154.571429	143.142857	95.000000	159.676058
21	14	157.428571	143.428571	99.714286	160.812064
22	23	180.869565	160.130435	113.347826	168.798526
23	40	165.775000	142.175000	101.150000	169.519045
26	60	175.500000	151.116667	107.483333	171.284086
32	21	166.523810	146.333333	101.619048	167.371085
33	18	166.500000	146.222222	100.777778	167.694094
34	22	170.636364	144.318182	101.136364	172.970452
35	26	167.692308	145.615385	99.730769	169.627258
36	23	169.304348	150.478261	105.304348	166.218121
37	11	164.363636	141.909091	93.272727	170.844011
39	21	173.761905	150.047619	104.666667	171.167185

TABLE XIX

LANGUAGE ARTS, FIFTH GRADE, UNADJUSTED AND ADJUSTED MEAN SCORES: 1983

Teacher Identification Number	Number of Students	Unadjusted Prior Mean Language 5	Language 3	School Ability 4	Adjusted Mean
1	17	179.000000	137.823529	97.176471	187.038956
2	18	165.166667	141.444444	99.000000	170.061267
3	19	151.473684	132.631579	92.736842	164.721368
4	17	183.117647	148.941176	103.176471	181.348328
9	51	180.333333	147.627451	102.764706	179.608988
11	21	167.666667	142.523810	101.142857	171.010417
15	23	177.086957	146.869565	103.739130	176.500496
17	22	167.045455	152.545455	102.772727	162.997479
18	17	178.411765	153.058824	106.470588	172.598531
19	18	180.055556	149.888889	105.611111	176.712336
24	40	171.125000	145.600000	101.150000	172.389018
26	18	174.777778	154.277778	108.444444	167.384280
27	18	183.833333	153.166667	108.888889	177.019534
28	22	172.909091	157.500000	108.318182	163.388443
32	22	173.318182	141.227273	100.590909	177.749080
33	18	165.944444	143.444444	100.777778	168.806656
34	22	178.727273	142.772727	101.136364	181.905449
35	26	165.153846	143.230769	99.730769	168.561999
36	23	167.521739	152.130435	105.304348	162.782752
37	11	160.545455	134.818182	93.272727	172.111195
38	51	164.921569	139.823529	98.019608	171.286699
39	21	180.333333	150.428571	104.666667	176.988054
40	71	168.211268	154.239437	103.323944	162.808076

TABLE XX

READING, FIFTH GRADE, UNADJUSTED AND ADJUSTED MEAN SCORES: 1984

Teacher Identification Number	Number of Students	Unadjusted Prior Mean Reading 5	Reading 3	School Ability 4	Adjusted Mean
1	19	650.789474	143.736842	100.789474	658.378828
2	18	664.333333	154.000000	106.833333	655.798494
3	15	640.466667	145.800000	98.866667	647.455194
4	22	630.954545	141.909091	98.545455	642.398075
5	21	637.333333	147.904762	102.523810	638.980566
6	20	635.800000	148.650000	102.600000	636.581608
9	46	655.217391	153.717391	104.065217	649.315964
10	33	642.333333	151.333333	100.121212	642.314990
11	25	639.320000	141.640000	96.600000	652.690106
14	24	638.333333	146.458333	97.291667	645.939237
17	13	612.615385	137.000000	86.538462	639.442967
18	23	647.217391	163.260870	113.000000	650.532977
19	22	637.909091	147.227273	98.045455	644.053658
20	20	634.750000	150.350000	103.700000	632.777562
21	22	663.000000	166.909091	116.863636	632.140522
22	13	605.538462	135.615385	88.384615	632.301602
24	53	651.396226	153.792453	104.301887	645.214905
26	16	665.250000	156.812500	114.375000	647.343728
27	19	655.052632	149.578947	103.157895	654.365665
28	18	656.944444	150.222222	103.388889	655.371243
29	6	616.333333	132.166667	87.833333	647.269510
30	10	671.400000	164.000000	111,500000	648.182950
31	10	664.200000	151.600000	106.600000	658.442730

TABLE XX (Continued)

Teacher Identification Number	Number of Students	Unadjusted Prior Mean Reading 5	Reading 3	School Ability 4	Adjusted Mean
32	6	623.000000	132.666667	95.833333	646.666146
33	22	625.772727	143.500000	93.454545	639.789403
34	26	639.961538	147.923077	99.961538	643.745313

TABLE XXI

MATH, FIFTH GRADE, UNADJUSTED AND ADJUSTED MEAN SCORES: 1984

Teacher Identification Number	Number of Students	Unadjusted Prior Mean Math 5	Math 3	School Ability 4	Adjusted Mean
1	18	665.500000	141.666667	100.944444	672.481083
2	18	695.888889	142.722222	106.444444	695.412750
3	14	639.714286	.39.642857	98.285714	652.824546
4	22	638.090909	144.590909	98.545455	642.679711
5	21	646.095238	143.571429	102.523810	648.263670
6	20	635.450000	142.600000	102.600000	639.159677
8	44	653.022727	146.909091	103.863636	648.236805
10	33	655.757576	144.181818	100.121212	659.396537
11	25	642.320000	141.240000	96.600000	654.512542
12	47	650.957447	144.872340	102.553191	650.925767
16	23	661.565217	143.000000	98.260970	669.102124
17	57	652.596491	144.192982	101.842105	654.434378
20	18	640.111111	151.611111	102.555556	628.837972
22	18	686.222222	161.388889	115.555556	645.176525
23	53	661.207547	147.283019	104.301887	655.344055
25	18	629.388889	145.166667	97.833333	633.755031
26	54	674.555556	147.555556	106.000000	666.478663
33	21	626.666667	138.095238	93.238095	647.586255
34	26	647.230769	142.807692	99.961538	653.326902

TABLE XXII

LANGUAGE ARTS, FIFTH GRADE, UNADJUSTED AND ADJUSTED MEAN SCORES: 1984

Teacher Identification Number	Number of Students	Unadjusted Prior Mean Language 5	Language 3	School Ability 4	Adjusted Mean
1	17	660.529412	136.823529	101.176471	668.636635
2	17	676.117647	147.529412	107.352941	668.504916
3	14	652.214286	141.428571	98.285714	659.075423
4	23	642.782609	140.521739	98.130435	650.609507
5	22	642.636364	143.090909	101.772727	644.543129
6	19	640.526316	142.684211	102.000000	642.570997
9	45	661.200000	148.666667	103.822222	656.080214
10	34	651.852941	143.617647	100.029412	655.021742
11	25	651.720000	135.800000	96.600000	665.293247
13	49	650.795918	140.836735	102.918367	653.580490
15	24	659.083333	143.833333	97.041667	665.029773
17	17	661.294118	153.529412	104.941176	650.710942
18	21	653.000000	146.190476	101.619048	652.286286
19	19	656.736842	146.684211	97.315789	659.859962
21	53	653.641509	153.622642	105.320755	642.597516
24	53	662.245283	148.452830	104.301887	656.839908
26	16	670.687500	148.125000	114.375000	655.560066
27	19	660.473684	147.789474	103.157895	656.799270
28	18	663.833333	147.388889	103.388889	660.287665
33	22	633.454545	135.363636	93.454545	650.545647
34	26	651.846154	140.961538	99.961538	657.458941

The analysis of covariance provides a statistical method of adjusting the fifth grade teacher mean scores in reading, math and language arts for prior differences in student ability.

The analysis of covariance with its accompanying F statistic also provided valuable information concerning the significance of differences in teacher effectiveness. This portion of the analysis directly answers one of this study's primary research questions--Are there significant differences in teachers' effectiveness? Table XXIII, Teacher Effect After Adjustment, was created to provide data on these levels of difference. In each subject area for both school years, this study found significant differences in teachers' effectiveness at the .01 level.

This section presents the adjusted achievement means for the teachers within this study in reading, math and language arts. The analysis of covariance also found statistically significant differences in teacher effects.

Development of Best Linear Unbiased Prediction Estimates

Many previous studies dealing with the concept of evaluating teachers on the basis of student achievement data have concluded with effectiveness measures calculated through the analysis of covariance. Statisticians and many educators have been critical of such methods due to the statistical problem of the regression to the mean. This term refers to the tendency of students who have scored either extremely high or low on prior tests, due partly to testing process errors, to move toward the mean of all students on subsequent tests. Nonrandomed

TABLE XXIII

TEACHER EFFECT AFTER ADJUSTMENT

Source	Degrees of Freedom	Sums of Squares	Mean Square	F-Value	Expected Mean Square	Variance Component
School Year 1982-83						
Reading 5	23	10,687.72	464.68	*[5.02]*	20.46	18.19
Scale Score Reading 3	1	17,125.25	17,125.25	185.08		
School Ability 4	1	5,824.87	5,824.87	62.95		
Math 5	20	6,784.08	339.20	*[5.77]*	26.92	10.42
Scale Score Math 5	1	12,754.92	12,754.92	217.00		
School Ability 4	1	5,207.22	5,207.22	88.59		
Language 5	22	25,685.50	1,167.52	*[4.37]*	24.92	36.14
Scale Score Language 3	1	35,504.45	35,504.45	132.99		
School Ability 4	1	6,990.04	6,990.04	26.18		
School Year 1983-84						
Reading 5	25	26,387.10	1,055.48	*[3.21]*	20.41	35.60
Scale Score Reading 3	1	53,558.20	53,558.20	162.90		
School Ability 4	1	29,910.88	29,910.88	90.97		
Math 5	18	79,993.38	4,444.08	*[9.94]*	28.34	141.04
Scale Score Math 3	1	74,118.98	74,118.98	165.95		
School Ability 4	1	51,620.15	51,620.15	115.46		
Language 5	20	27,575.36	1,378.77	*[3.52]*	25.81	38.22
Scale Score Language 3	1	62,091.97	62,091.97	158.31		
School Ability 4	1	51,127.09	51,127.09	130.35		

*The critical value of F at the 0.01 level of significance is approximately 2.04 where the degrees of freedom among teachers is greater than 16 and degree of freedom within teachers is greater than 400 in all cases.

student assignment to teachers is common practice in most schools. Therefore, teacher effects calculated without consideration of the regression to the mean problem would be biased and unfair. The purpose of this section was to present the Best Linear Unbiased Prediction estimates of teacher effects in reading, math, and language arts which control for the problem of the regression to the mean.

Best Linear Unbiased Prediction estimates were calculated for teacher effects in reading, math, and language arts for the two years of this study. A detailed description of this model and an example was provided in Chapter III, page 64. The Henderson's mixed model solution procedure provided the best linear unbiased prediction (BLUP) of the random teacher effects. The BLUP estimates are presented on Table XXIV. In this teacher evaluation plan BLUP estimates of teacher effects provided a solution to the regression to the mean problem which has often been criticized in past evaluation proposals.

The table of BLUP estimates contains six columns, BLUP Reading 1, BLUP Reading 2, BLUP Math 1, BLUP Math 2, BLUP Language Arts 1, and PLUP Language Arts 2. Teacher numerical identification was also included in the table. This table provided a quick visual scale of teacher effects per year for each of the three subject areas of the study. Positive estimates can be interpreted as above average performance, negative estimates as below average, and zero estimates as average performance. Estimates represent deviations from the mean. These yearly BLUP estimates will be used in further analysis to check for consistency between years and in calculating a composite estimate of teacher effects for the two years of the study.

TABLE XXIV

BLUP (BEST LINEAR UNBIASED PREDICTION) OF TEACHER EFFECTS*

Observation	BLUP Estimates Reading 1	Teacher	BLUP Estimates Reading 2	BLUP Estimates Math 1	BLUP Estimates Math 2	BLUP Estimates Language Arts 1	BLUP Estimates Language Arts 2
1	3.6149	1	8.6142	5.4565	15.913	10.2117	8.081
2	1.5612	2	6.7387	1.3457	35.312	- 1.6271	8.060
3	-6.4344	3	1.2031	-1.0289	- 0.779	- 5.5361	2.005
4	1.7028	4	-2.2186	3.1772	- 9.505	6.3108	- 3.459
5	.	5	-4.5787	.	- 4.743	.	- 7.541
6	1.7028	6	-6.1467	.	-12.612	.	- 8.465
7	.	8	.	-4.4425	- 5.008	.	.
8	1.0082	9	3.1424	.	.	6.3756	0.438
9	.	10	-2.4721	.	5.269	.	- 0.431
10	5.5467	11	5.2466	0.1164	0.762	- 1.0046	6.828
11	4.5343	12	.	-3.0935	- 2.564	.	.
12	.	13	.	.	.	.	- 1.701
13	-2.2931	14	0.3403	.	.	.	.
14	.	15	.	.	.	3.1526	6.633
15	.	16	.	3.4487	13.606	.	.
16	.	17	-3.4401	0.3483	0.721	- 6.9127	- 2.978
17	.	18	3.4901	.	.	0.2260	- 2.199
18	.	19	-1.0025	.	.	3.1191	2.814
19	-3.7350	20	-8.7536	-6.4295	-20.836	.	.
20	-2.4762	21	-9.5443	-5.6192	.	.	-10.799
21	-7.5220	22	-7.6498	0.0140	- 6.887	.	.
22	.	23	.	0.7279	1.644	.	.
23	-1.9586	24	-0.2832	.	.	0.0584	1.082
24	.	25	.	.	-16.793	.	.
25	2.3821	26	1.0135	2.3305	12.127	- 3.4723	- 0.034
26	4.1106	27	5.9206	.	.	3.3420	0.812
27	2.1266	28	6.4808	.	.	- 6.6179	3.014
28	.	29	0.7325	.	.	.	.
29	.	30	1.3434	.	.	.	.
30	-6.1987	31	6.6648	.	.	.	.
31	-1.9313	32	0.4245	-1.0562	.	4.0181	.
32	0.0982	33	-3.9611	-0.7742	- 5.312	- 2.5114	- 3.480
33	5.3830	34	-1.3046	3.3985	- 0.316	7.1422	1.321
34	-0.6178	35	.	0.7551	.	- 2.9404	.
35	-4.0303	36	.	-2.0187	.	- 7.1828	.
36	2.7594	37	.	1.4228	.	- 0.1626	.
37	-2.7347	38	.	.	.	- 0.9450	.
38	5.1043	39	.	1.9210	.	3.4743	.
39	.	40	.	.	.	- 8.5180	.

*Positive estimates can be interpreted as above average performance, negative estimates as below average, and zero estimates as average performance. Estimates represent deviations from the mean.

The BLUP estimates demonstrated that teachers do vary in their effectiveness even after adjusting for the regression to the mean. The group of teachers involved in this study differed in effectiveness as indicated by standardized achievement test scores in reading, math, and language arts.

Correlation Study of Consecutive Year Estimates

The purpose of this section was to present the correlations between teacher effects in consecutive years. The original independent BLUP estimates for 1983 and 1984 were used in this analysis. The relationship of effects between the two consecutive years was presented graphically in Figures 18-20.

A simple coefficient of correlation between the two effects or estimates was presented on the figures. The product moment correlation (Pearson r) was used in these calculations. The Pearson r is appropriate when both variables to be correlated are expressed as ratio or interval data. Achievement measures are expressed in the form of interval data. The values of these correlation coefficients for reading, math, and language arts were .45, .58, and .20, respectively. The total correlation in reading (.45) was significant at the 0.06 level. Math (.58) was significant at the 0.02 level. Both levels of significance are considered adequate for the purposes of this study, ranking or evaluating teachers on the basis of achievement test data. Language arts .20 was significant at the 0.45 level, a lower correlation of consistency.

READING

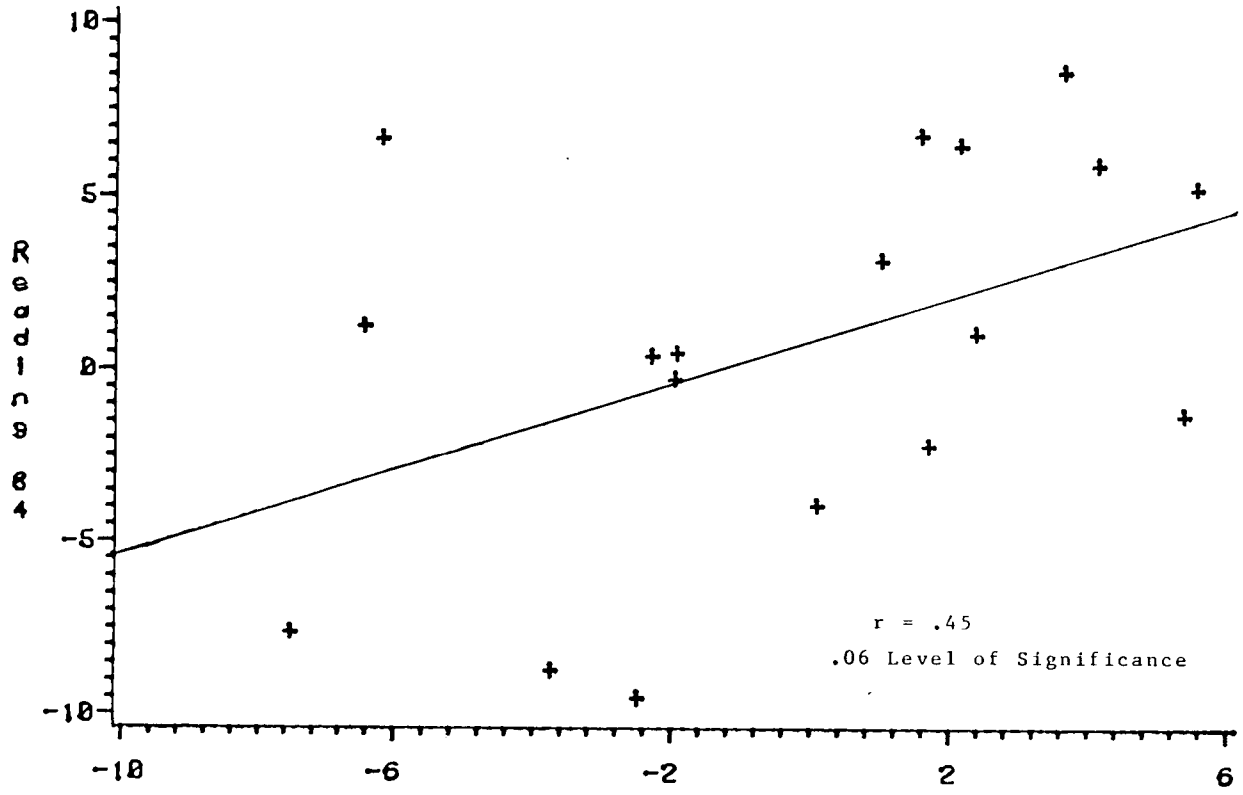


FIGURE 18

RELATIONSHIP BETWEEN TEACHER EFFECTS ESTIMATED IN CONSECUTIVE YEARS, READING, 1983

MATH

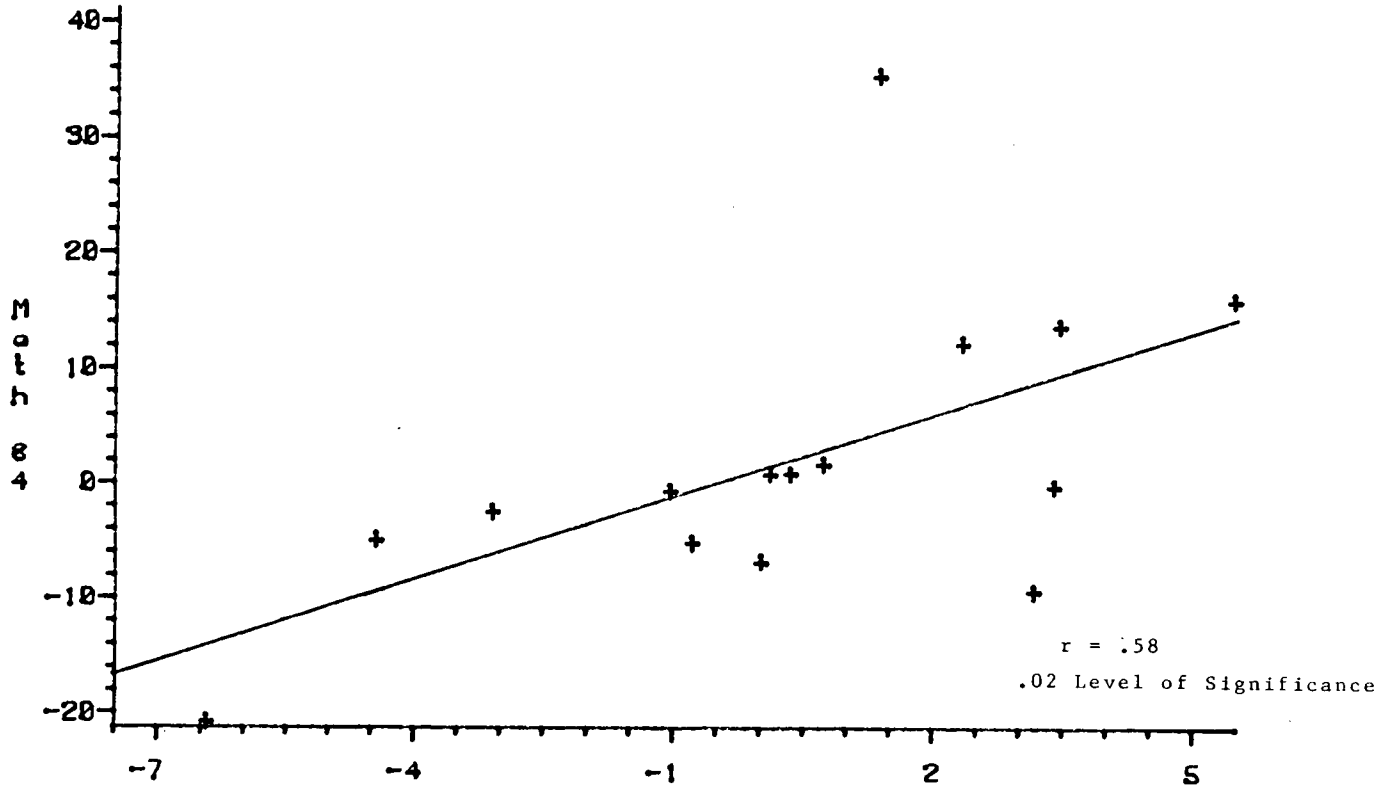


FIGURE 19

RELATIONSHIP BETWEEN TEACHER EFFECTS ESTIMATED IN CONSECUTIVE YEARS, MATH, 1983

LANGUAGE

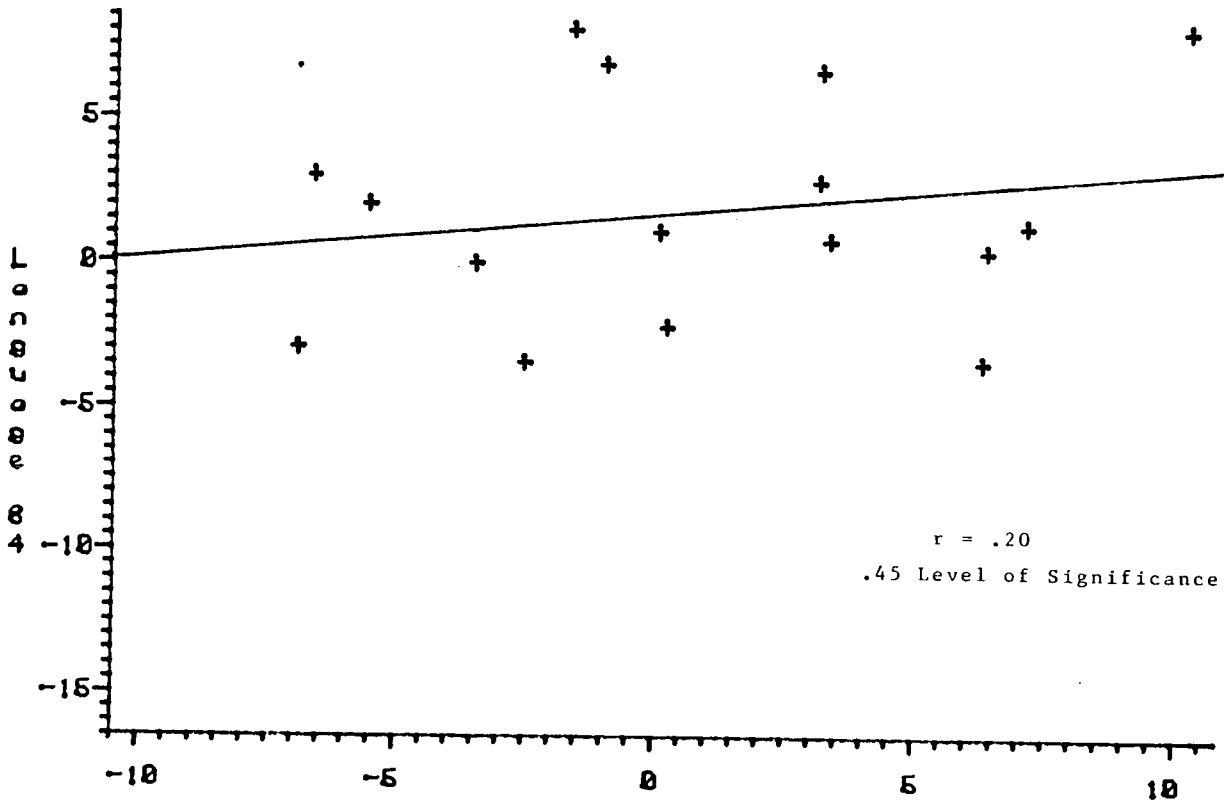


FIGURE 20

RELATIONSHIP BETWEEN TEACHER EFFECTS ESTIMATED IN CONSECUTIVE YEARS, LANGUAGE, 1983

An important factor that needs to be considered in the interpretation of the correlations is the small sample size of instructors. The consistency correlation in reading is based upon a sample of 18 teachers, math--15 teachers, and language arts--16 teachers.

The correlation calculated for total teacher effect (reading, math and language) between the two years was 0.36 which was significant at the 0.01 level. This correlation is based upon a total sample size of 49.

Significant correlations were found in teacher effects between the two consecutive years of the study for reading and math. The correlation between teacher effects in language arts, however, was somewhat low.

Composite Teacher Rankings

Two sets of rankings or evaluations can be derived from the BLUP estimates of teacher effects. A ranking of teachers based upon student achievement data for a single school year or a composite ranking for a two year or more period can be developed. The single year rating or ranking of teachers for this study will be presented and discussed first. Table XXV gives a ranking of fifth grade reading teachers for the Blount County System for the 1982-83 school year. This table can be interpreted as follows: teacher 11 had the highest mean adjusted achievement score, 34 second highest, and so forth. The following tables provide the other single year rankings developed during this study: Table XXV, Reading Rank,

TABLE XXV

SINGLE YEAR TEACHER RANK: READING,
1983, FIFTH GRADE

Rank	*BLUP Estimates Reading 1	Teacher
1	5.5467	11
2	5.3830	34
3	5.1043	39
4	4.5343	12
5	4.1106	27
6	3.6149	1
7	2.7594	37
8	2.3821	26
9	2.1266	28
10	1.7028	4
11	1.5612	2
12	1.0082	9
13	0.0982	33
14	-0.6178	35
15	-1.9313	32
16	-1.9586	24
17	-2.2931	14
18	-2.4762	21
19	-2.7347	38
20	-3.7350	20
21	-4.0303	36
22	-6.1987	31
23	-6.4344	3
24	-7.5220	22

*Best Linear Unbiased Prediction.

1983; Table XXVI, Math Rank, 1983; Table XXVII, Language Rank, 1983; Table XXVIII, Reading Rank, 1984; Table XXIX, Language Rank, 1984; and Table XXX, Math Rank, 1984.

An evaluation system of teaching based upon student achievement data for fairness and completeness should be based upon more than one year's data. Some form of cumulative or composite score would need to be developed. The next step in this analysis involved the calculation of a composite score for each teacher (teachers in the study for both consecutive years).

A statistical problem encountered in this study that had to be solved before a composite ranking could be calculated involved a change in editions of the Stanford Achievement test in 1984. In 1984, the fifth grade Stanford Achievement Test was changed from the 1973 edition to the 1983 edition within the Blount County System. The scale scores from the 1983 edition were not comparable to the other data within the study. To solve this problem and provide comparable scores, scale scores had to be standardized. This standardization process was explained in detail in Chapter III, page 70.

After the standardization process was completed, composite scores were calculated in order to rank the fifth grade teachers in reading, math, and language arts for a two year period of time (1982-83 and 1983-84). These composite scores were also calculated by using the Henderson Best Linear Unbiased Prediction (BLUP) procedure. The concept underlying this procedure was previously discussed in Chapter III, page 64. Tables XXXI-XXXIII were developed to provide the composite rankings in reading, math, and language arts for fifth grade

TABLE XXVI

SINGLE YEAR TEACHER RANK: MATH, 1983, FIFTH GRADE

Rank	*BLUP Estimates Math 1	Teacher
1	5.4565	1
2	3.4487	16
3	3.3985	34
4	3.1772	4
5	2.3305	26
6	1.9210	39
7	1.4228	37
8	1.3457	2
9	0.7551	35
10	0.7279	23
11	0.3483	17
12	0.1164	11
13	0.0140	22
14	-0.7742	33
15	-1.0289	3
16	-1.0562	32
17	-2.0187	36
18	-3.0935	12
19	-4.4425	8
20	-5.6192	21
21	-6.4295	20

*Best Linear Unbiased Prediction.

TABLE XXVII

SINGLE YEAR TEACHER RANK: LANGUAGE ARTS, 1983, FIFTH GRADE

Rank	*BLUP Estimates Language Arts 1	Teacher
1	10.2117	1
2	7.1422	34
3	6.3756	9
4	6.3108	4
5	4.0181	32
6	3.4743	39
7	3.3420	27
8	3.1526	15
9	3.1191	19
10	0.0584	24
11	0.2260	18
12	-0.1626	37
13	-0.9450	38
14	-1.0046	11
15	-1.6271	2
16	-2.5114	33
17	-2.9404	35
18	-3.4723	26
19	-5.5361	3
20	-6.6179	28
21	-6.9127	17
22	-7.1828	36
23	-8.5180	40

*Best Linear Unbiased Prediction.

TABLE XXVIII

SINGLE YEAR TEACHER RANK: READING,
1984, FIFTH GRADE

Rank	*BLUP Estimates Reading 2	Teacher
1	8.6142	1
2	6.7387	2
3	6.6648	31
4	6.4808	28
5	5.9206	27
6	5.2466	11
7	3.4901	18
8	3.1424	9
9	1.3434	30
10	1.2031	3
11	1.0135	26
12	0.7325	29
13	0.4245	32
14	0.3403	14
15	-0.2832	24
16	-1.0025	19
17	-1.3046	34
18	-2.2186	4
19	-2.4721	10
20	-3.4401	17
21	-3.9611	33
22	-4.5787	5
23	-6.1467	6
24	-7.6498	22
25	-8.7536	20
26	-9.5443	21

*Best Linear Unbiased Prediction.

TABLE XXIX

SINGLE YEAR TEACHER RANK: LANGUAGE ARTS, 1984, FIFTH GRADE

Rank	*BLUP Estimates Language Arts 2	Teacher
1	8.081	1
2	8.060	2
3	6.828	11
4	6.633	15
5	3.014	28
6	2.814	19
7	2.005	3
8	1.321	34
9	1.082	24
10	0.812	27
11	0.438	9
12	- 0.034	26
13	- 0.431	10
14	- 1.701	13
15	- 2.199	18
16	- 2.978	17
17	- 3.459	4
18	- 3.480	33
19	- 7.541	5
20	- 8.465	6
21	-10.799	21

*Best Linear Unbiased Prediction.

TABLE XXX

SINGLE YEAR TEACHER RANK: MATH, 1984, FIFTH GRADE

Rank	*BLUP Estimates Math 2	Teacher
1	35.312	2
2	15.913	1
3	13.606	16
4	12.127	26
5	5.269	10
6	1.644	23
7	0.762	11
8	0.721	17
9	- 0.316	34
10	- 0.779	3
11	- 2.564	12
12	- 4.743	5
13	- 5.008	8
14	- 5.312	33
15	- 6.887	22
16	- 9.505	4
17	-12.612	6
18	-16.793	25
19	-20.836	20

*Best Linear Unbiased Prediction.

TABLE XXXI

READING COMPOSITE BASED UPON TWO YEAR ACHIEVEMENT DATA: 1983-1984

Rank	**BLUP Estimates Reading	Teacher
1	2.8046	1*
2	2.7099	39
3	2.5647	11*
4	2.3756	12
5	2.3607	27*
6	1.8533	2*
7	1.8026	28*
8	1.5672	37
9	1.2883	34*
10	1.1891	18
11	1.0558	26*
12	0.8598	9*
13	0.4706	30
14	0.1550	29
15	0.0585	4*
16	-0.1934	35
17	-0.2793	31*
18	-0.3911	19
19	-0.4817	24*
20	-0.6215	14*
21	-0.6714	32*
22	-0.7783	33*
23	-0.8490	10
24	-1.2959	17
25	-1.4002	38
26	-1.4119	5

TABLE XXXI (Continued)

Rank	**BLUP Estimates Reading	Teacher
27	-1.8512	3*
28	-1.9140	6
29	-2.0454	36
30	-2.4331	21*
31	-2.6427	20*
32	-3.8557	22*

*Two year data estimates.

**Best Linear Unbiased Prediction.

TABLE XXXII

MATH COMPOSITE BASED UPON TWO YEAR
ACHIEVEMENT DATA: 1983-1984

Rank	**BLUP Estimates Math	Teacher
1	4.9221	2*
2	4.5234	1*
3	3.2967	16*
4	2.6072	26*
5	1.7173	39
6	1.3443	37
7	1.3432	34*
8	1.3168	10
9	0.8912	35
10	0.5948	23*
11	0.4126	17*
12	0.3446	11*
13	-0.0346	4*
14	-0.3764	3*
15	-0.4156	32
16	-.5292	22*
17	-0.8221	33*
18	-1.0361	36
19	-1.0432	5
20	-1.2241	12*
21	-2.0779	8*
22	-2.8770	6
23	-3.7148	21
24	-3.9378	25
25	-5.2252	20*

TABLE XXXIII

LANGUAGE ARTS COMPOSITE BASED UPON TWO YEAR ACHIEVEMENT DATA: 1983-1984

Rank	**BLUP Estimates Language Arts	Teacher
1	3.9464	1*
2	1.8954	34*
3	1.8934	15*
4	1.7806	9*
5	1.6097	32
6	1.3718	39
7	1.2271	19*
8	1.1586	11*
9	1.0019	27*
10	0.7466	4*
11	0.7429	2*
12	0.5461	24*
13	0.2315	37
14	0.0678	10
15	-0.0482	38
16	-0.1122	13
17	-0.1759	18*
18	-0.4235	26*
19	-0.7339	3*
20	-0.8313	35
21	-0.8811	28*
22	-1.0526	33*
23	-1.8132	17*
24	-1.8635	5
25	-2.1132	6

TABLE XXXIII (Continued)

Rank	**BLUP Estimates Language Arts	Teacher
26	-2.4719	36
27	-2.6995	21
28	-2.9996	40

*Two year data estimates.

**Best Linear Unbiased Prediction.

teachers in Blount County for two consecutive school years (1982-83 and 1983-84). This same data was used in developing Figures 21-23 in order to provide a graphic presentation of the rankings.

The final ranking developed from these composite scores include teachers having taught two consecutive years in a subject, as well as teachers having taught only one year. The BLUP procedure allows this calculation. However, the estimates based upon single year data would be interpreted as estimates with smaller confidence limits. An asterisk identifies those teachers with achievement data available for both years of the study.

Both single year rankings and composite rankings for fifth grade teachers in reading, math, and language arts were developed through using the Henderson Best Linear Unbiased Prediction (BLUP) procedure.

Two-Level Independent Corroboration of Teacher Rankings

As a secondary form of independent corroboration of the results obtained from the ranking of teachers based upon achievement data, both elementary principals and instructional supervisors were involved in this study. When the various school principals were asked to provide roll books for teacher identification in this study, they were simultaneously asked to subjectively rank their fifth grade teacher or teachers as to where they felt they would fall in the final ranking (from achievement data) in reading/math/language arts in comparison with the entire population of fifth grade teachers within the Blount County Schools. The principal ranking form found

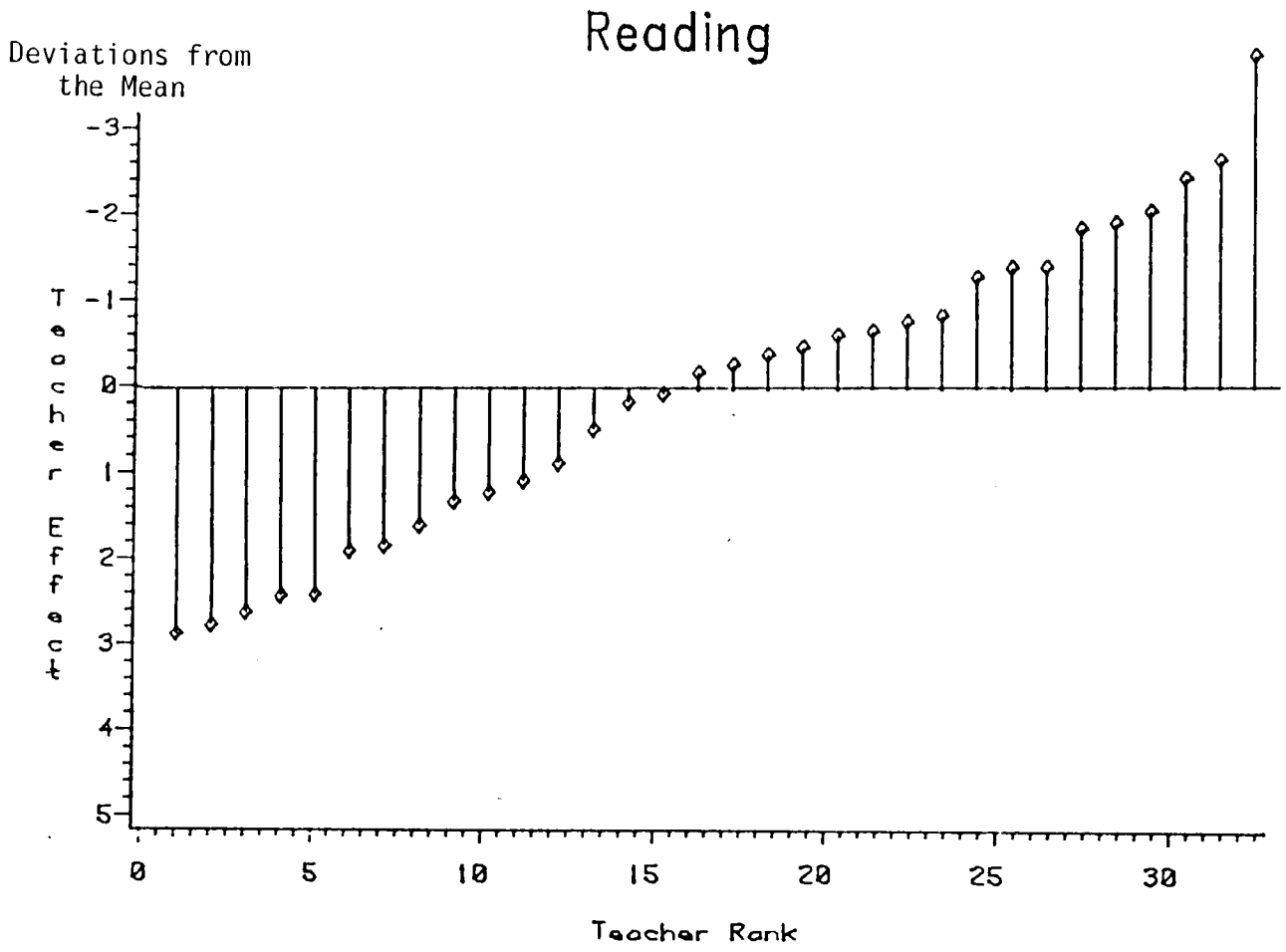


FIGURE 21

READING COMPOSITE RANKING GRAPHIC PRESENTATION BASED UPON TWO YEAR ACHIEVEMENT DATA, 1983-1984

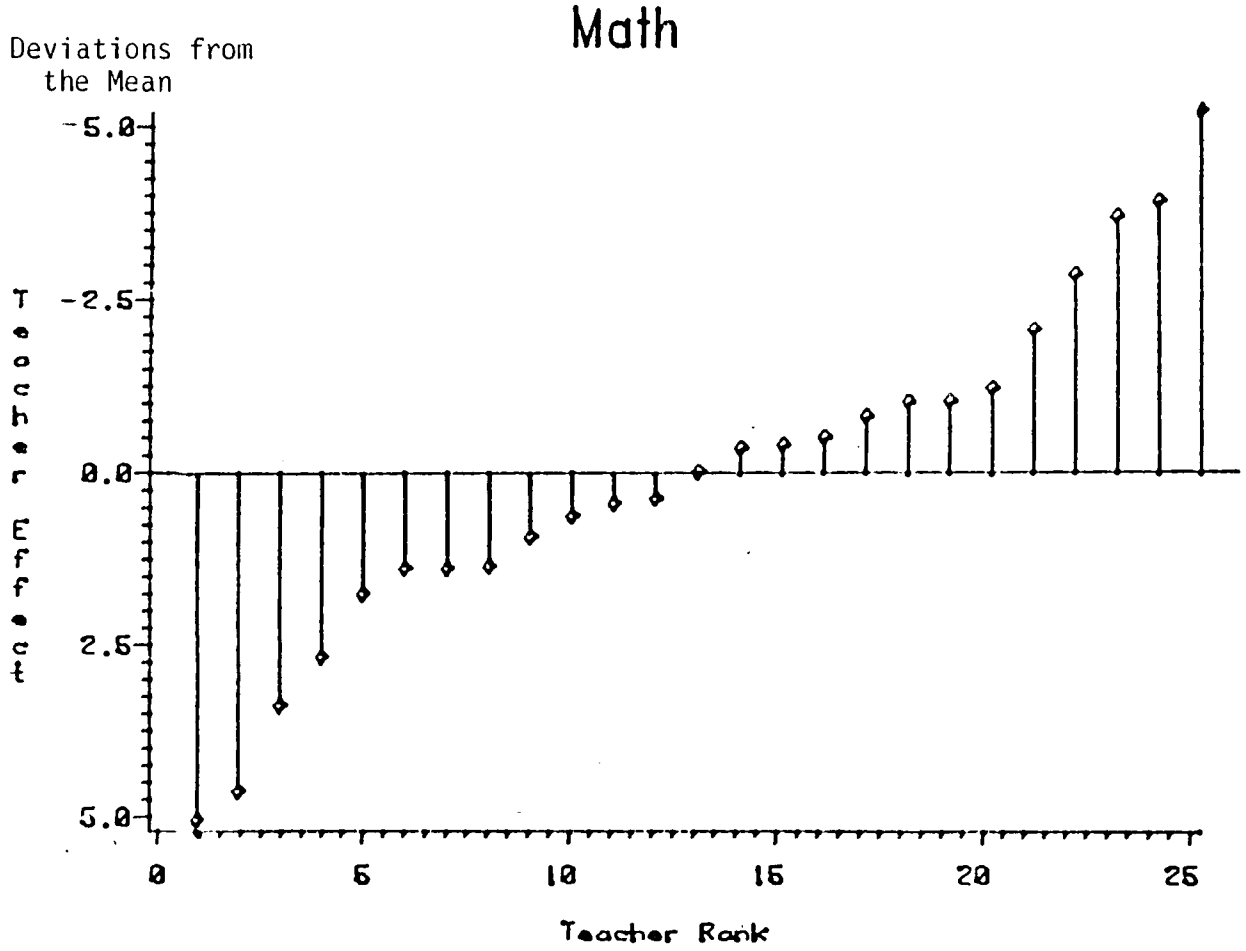


FIGURE 22

MATH COMPOSITE RANKING GRAPHIC PRESENTATION BASED UPON TWO YEAR ACHIEVEMENT DATA, 1983-1984

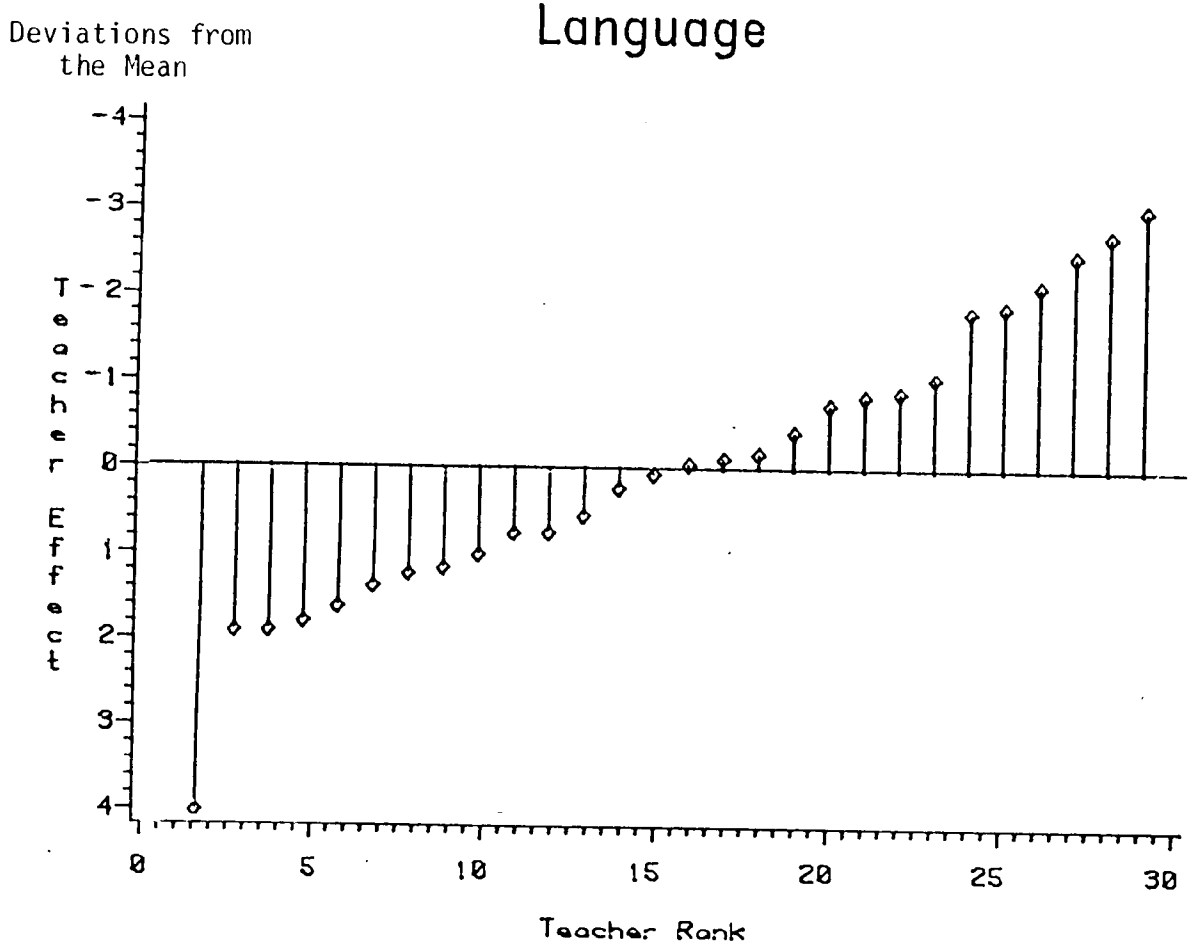


FIGURE 23

LANGUAGE COMPOSITE RANKING GRAPHIC PRESENTATION BASED UPON TWO YEAR ACHIEVEMENT DATA, 1983-1984

in the Appendix was created and used in assimilating the data for this aspect of the study.

The percentages of agreement between the principal subjective rankings and the actual ranking developed from standardized achievement test data are presented in Table XXXIV. Percentage of agreement are presented for total ranking and for the three levels: top third, middle third, and low third. The percentage of agreement between the total ranking provided by achievement data and principal opinion was 71 percent in reading, 61 percent in math, and 46 percent in language arts.

The ranking information was also used to create a 3 by 3 table to show more clearly the relationship between the principal's expected ranking and the actual achievement test ranking. Figure 24 shows this data for reading, Figure 25 for math, and Figure 26 for language arts. In reading there was a slight tendency for principals to underestimate. However, in both math and language arts the tendency was for the principals to overestimate their fifth grade teachers in terms of effectiveness as measured by standardized achievement test data.

Three elementary instructional supervisors were also asked to rank fifth grade Blount County teachers as to where they felt each fifth grade teacher would fall in the final ranking developed from achievement data in the three subject areas. The supervisory ranking form found in the Appendix was created to assimilate this data. The percentages of agreement between the supervisors' subjective rankings and the actual ranking developed from achievement test data is

TABLE XXXIV

RANKING COMPARISON: PRINCIPAL EVALUATION AND ACHIEVEMENT RANKING

Category	Total	Top Third	Mid Third	Low Third
Reading	71%	72%	100%	50%
Math	61%	75%	75%	55%
Language Arts	46%	33%	88%	30%

ACHIEVEMENT RANKING

PRINCIPAL EXPECTED RANKING	HIGH	MEDIUM	LOW
HIGH	5	---	3
MED.	6	10	2
LOW	---	---	4

FIGURE 24

READING RANKING COMPARISON (PRINCIPAL EXPECTED RANKING-ACHIEVEMENT RANKING)

ACHIEVEMENT RANKING

PRINCIPAL EXPECTED RANKING	HIGH	MEDIUM	LOW
HIGH	5	1	1
MED.	2	7	3
LOW	1	---	5

FIGURE 25

MATH RANKING COMPARISON (PRINCIPAL EXPECTED RANKING-ACHIEVEMENT RANKING)

ACHIEVEMENT RANKING

PRINCIPAL EXPECTED RANKING	HIGH	MEDIUM	LOW
HIGH	3	1	3
MED.	6	8	4
LOW	---	---	3

FIGURE 26

LANGUAGE ARTS RANKING COMPARISON (PRINCIPAL EXPECTED RANKING-ACHIEVEMENT RANKING)

presented in Table XXXV. Average individual agreement percentages were as follows: reading, 22 percent; math, 24 percent; and language arts, 44 percent.

A group consensus percentage was also calculated: reading 31 percent; math 40 percent; language arts 46 percent. These percentages were calculated as a result of a group meeting (using the same ranking form as the individuals completed) in which the collective group ranking decision was recorded.

Elementary school principals provided strong corroboration for the rankings developed from standardized achievement test data. Elementary central office supervisory staff's corroboration was weaker, however.

II. SUMMARY

This chapter has presented and analyzed the data of the study. First, an intensive study of the distribution frequency of the various test score groups indicated that there was no significant ceiling or floor effect within the data that might bias the final ranking. Second, student mean scores were adjusted to control for prior ability through utilizing the analysis of covariance (ANCOVA). Third grade achievement data and fourth grade school ability test scores were chosen as the covariables for this adjustment. Third, to control for the regression to the mean, the teachers' mean scores in reading, math, and language arts were adjusted again by employing the Best Linear Unbiased Prediction procedure. From these adjusted scores, deviations

TABLE XXXV

RANKING COMPARISON: SUPERVISOR EVALUATION AND ACHIEVEMENT RANKING

Category	Individual Supervisor 1	Individual Supervisor 2	Individual Supervisor 3	Average
Reading	25%	22%	19%	22%
Math	36%	32%	4%	24%
Language Arts	46%	43%	43%	44%

from the mean of all teachers were calculated in each subject area and used to rank teachers for both single year periods and for a two year period. Teachers were ranked according to their instructional effectiveness as calculated from the adjusted standardized achievement test scores.

Independent corroboration of this teacher ranking was sought from elementary principals and supervisory personnel. The building principal group provided moderate to strong corroboration for this ranking system. However, the supervisory group provided weaker corroboration.

CHAPTER V

SUMMARY, CONCLUSIONS, DISCUSSION, AND RECOMMENDATIONS

I. SUMMARY

The problem investigated in this study dealt with the need for a more objective measure of teacher effectiveness. The purpose of the study was to determine if standardized achievement tests could be used to objectively evaluate classroom teachers. This study also sought to determine if an alternate year achievement testing schedule could provide adequate data for an objective teacher evaluation plan based upon standardized achievement measures. The study attempted to provide answers to the following questions:

1. Are there significant differences in teacher effectiveness?

2. Can standardized achievement tests be used appropriately and accurately to rank classroom teachers in terms of classroom (teaching) effectiveness?

3. What are the appropriate procedures or designs for using student achievement data in teacher evaluation?

4. How do student achievement test rankings correlate with principal and supervisory ratings?

5. Is yearly achievement testing a requirement if one wishes to evaluate teaching effectiveness through the use of achievement test data?

The population was composed of fifth grade students and teachers within the Blount County School System (Tennessee) for the

1982-83 and 1983-84 school years. Computer tapes of Blount County achievement test data were obtained from the State Testing Center. Individual records were edited to insure consistent enrollment of students in adjacent years and to encode teacher identification.

An Analysis of Covariance was performed with third grade achievement and fourth grade school ability test information as independent variables (to control for prior ability) and fifth grade achievement test scores in reading, math, and language arts as the dependent variable. This analysis also determined if there were significant differences in teachers' effectiveness (F test). Henderson's Best Linear Unbiased Prediction (BLUP) procedure was then used with the data to correct for regression to the mean. The scores were then used to rank teachers in terms of their effectiveness in teaching reading, math, and language arts.

Correlations were calculated to determine if a teacher's effectiveness was consistent across years. Finally, the rankings developed from standardized achievement tests were compared to principal and supervisory expected rankings as a measure of independent corroboration.

The following findings were obtained:

1. There were significant differences in teachers' effectiveness within this study. The analysis of covariance with its accompanying F test found statistically significant differences in teachers' effectiveness as measured by adjusted mean student achievement scores at the .01 level of significance.

2. Mean standardized achievement scores were successfully used to rank teachers in terms of their instructional effectiveness. Teachers were ranked in the areas of reading, math and language arts for both a single year and for a two year period.

Achievement test data used for such purposes should be free of any ceiling or floor effect problems. A frequency distribution study was performed on the data initially to detect any ceiling or floor effect. No significant ceiling or floor effect problem was discovered. After this preliminary diagnostic procedure, it was necessary to adjust student scores to control for initial student differences by using the analysis of covariance. Henderson's Best Linear Unbiased Prediction Procedure was then used to control for the regression to the mean.

3. Consistency of teacher effectiveness as measured by achievement test data was significantly correlated across the two years of this study in reading and math. The total correlation in reading (.45) was significant at the 0.06 level. Math (.58) was significant at the 0.02 level. However, language arts (.20) was significant at the 0.45 level, a much lower correlation. The correlation calculated for total teacher effect (reading, math, and language arts) between the two years was 0.36 which was significant at the 0.01 level.

4. The principals' ratings and achievement test rankings had a moderate to strong level of agreement for both reading and math (reading, 71 percent and math, 61 percent). A lower agreement percentage was found in language arts (46 percent).

The agreement factor between the achievement test rankings and supervisory expected ratings was lower (reading, 22 percent;

math, 24 percent; and language arts, 44 percent.) The group collective consensus ranking percentages were reading 31 percent, math 40 percent, and language arts 46 percent.

5. An alternate year achievement testing schedule provided sufficient data for developing a ranking of teachers based upon achievement test data for those grades with baseline achievement scores available.

II. CONCLUSIONS AND DISCUSSION

1. This study has identified and documented significant differences in teachers' effectiveness. When there are significant differences in teachers' effectiveness, it is feasible to develop an objective component of teacher evaluation based upon achievement test data.

For too long, the principal's subjective appraisal of teachers provided the only means of distinguishing the performance levels among various teachers. The subjectivity of this method has been heavily criticized. The accountability issue over the past several years has forced upon the educational community the realization that its manpower appraisal and reward system must become more comparable to that of the business sector in our society. The identification of an objective measure of evaluating a teacher's performance has become a necessity. A related means or method of establishing a connection or link between this evaluation or performance appraisal and compensation has also become a pressing need in the education field. In reference to the

first need, this research has identified an objective teacher evaluation system based upon standardized achievement test data. The development of this objective component to be used in teacher evaluation has helped to bring the educational system's performance appraisal methodology into alignment with that of the business sector.

This research is supported by and extends other major studies in arriving at the conclusion that teachers differ in their teaching effectiveness and that this effectiveness can be determined by analysis of student achievement test data. The design or procedure identified built upon the results of previous research, but it also improved upon the previous statistics employed. As discussed in Chapter II, various statistics have been employed to produce an adjusted achievement score for evaluational purposes. Regression analysis was employed by Veldman and Brophy (1973)[114] and Evertson and Brophy (1973).[115] This method of adjustment was advocated by Klein and Alkin[116] and Cronbach.[117] McGowan employed a teacher by level of analysis of variance to evaluate teachers.[118] The statistical method of analysis of covariance has been advocated by such noted authorities as Jason

---

[114]Veldman, op. cit., p.2.

[115]Evertson and Brophy, op. cit., p. 1.

[116]Alkin and Klein, op. cit., p. 8.

[117]Cronbach, op. cit., p. 80.

[118]McGowan, op. cit., pp. 49-50.

Millman[119] and L. R. Gay.[120] Last, Sanders and McLean (1983) used a gain score concept adjusted through the use of the BLUP (Best Linear Unbiased Prediction) procedure to correct for the regression to the mean.[121] Each of these studies, employing different statistical methods, have concluded that teachers differ in their effectiveness and that student achievement test data can be used to objectively evaluate classroom teachers.

This research has produced a unique design or set of recommended procedures for utilizing student achievement scores in teacher evaluation. The study utilized data from a school system employing an alternate year achievement testing schedule that eliminated the use of a gain score concept. First, all student data was subjected to a frequency distribution study to determine if any significant ceiling or floor effect was present. Such effects would bias final results. No significant ceiling or floor effect was found with the test data of this study. Evidence of a slight ceiling effect was identified in both third grade language arts data sets (language arts, 1982 and language arts, 1981). However, the researcher determined that the ceiling effect problem was not significant because the magnitude of the effect (4 percent or less) was low and because both effects were found only in data used as an adjustment factor and

---

[119]Millman, op. cit., p. 161.

[120]Gay, op. cit., p. 322.

[121]McLean and Sanders, op. cit., pp. 4-5.

not in the primary data set (fifth grade achievement). Second, student scale scores were adjusted to control for prior ability through the use of the analysis of covariance. Third grade achievement and fourth grade school ability test scores were used as adjustment factors. Significant differences, (.01 level of significance) were found in teachers' effectiveness even after this adjustment. This significant difference had to be documented before any ranking system could be considered from achievement based data. Most research has stopped after this initial adjustment for prior knowledge and teachers have been ranked or evaluated on the basis of the adjusted scores. Critics have been quick to denounce such methodology because of the researchers neglect of the regression to the mean problem (Soar and Soar, 1975;[122] and Borich, 1977[123]). This study, building upon the work of Sanders and McLean, went one step further in adjustment and utilized the Best Linear Unbiased Prediction (BLUP) procedure to control for regression to the mean.

This research has also shown that these teacher effects were consistent across the years of the study (1982-83 and 1983-84) The total correlation in reading was .45 which was significant at the 0.06 level. Math (.58) was significant at the 0.02 level. Both levels of significance were considered adequate for the purpose of this study,

---

[122]Robert S. Soar and Ruth Soar, "Problems in Using Pupil Outcomes for Teacher Evaluation," (Washington, D.C.: National Education Publication, April 1975), p. 12.

[123]Borich, op. cit., p. 51.

ranking or evaluating teachers on the basis of achievement test data. Language arts (.20) was significant at the 0.45 level, a lower correlation. The correlations calculated for total teacher effect (reading, math, and language arts) between the two years was .36, which was significant at the 0.01 level. Consistency of teacher effects was considered an important factor in the development of an evaluation design based upon student achievement test data. It is hoped that further refinements in the evaluation procedures will help to improve this consistency measure.

Similar findings in regard to the consistency of teacher effect were provided by Fortune (1966),[124] Veldman and Brophy (1973),[125] McGowan (1982)[126] and Sanders.[127] McLean and Sanders[128] utilized student achievement data over a three year period to establish this consistency. Fortune (1966) used an experimental design with a group of Head Start teachers to document the stability of teacher effects.[129] The consistency of teacher effectiveness is well documented.

---

[124]Rosenshine, op. cit., p. 657.

[125]Veldman, op. cit., p. 3.

[126]McGowan, op. cit., p. 64.

[127]McLean and Sanders, op. cit., p. 9.

[128]Ibid.

[129]Rosenshine, op. cit., p. 657.

2. The design or procedures for implementing this teacher evaluation system based upon achievement test data is feasible and applicable to most school systems under the following conditions:

a. The system has an agreed upon curriculum.

b. The achievement tests have been carefully selected to match goals and objectives of the curriculum.

c. The integrity of the testing process has been protected.

d. Appropriate computer programs and services are available both for the scoring of student achievement tests and for the calculation of teacher effectiveness measures and ranking.

Discussion

If teachers are to be evaluated upon achievement test data, it is imperative that a central curriculum guide the instructional program. Teachers should be evaluated upon similar goals and objectives at each grade level. Uniformity of curriculum goals and objectives could be system-wide or state-wide. After an agreed-upon curriculum has been established, safeguards should be maintained to monitor the instructional program to make sure the curriculum is being implemented.

The selection of a standardized achievement test is the second major decision to be made in the implementation of a teacher evaluation system based upon achievement test data. The test or

tests should definitely test the goals and objectives that are guiding the instructional program. This might appear to the reader to be generally understood. However, in many settings, test selection is taken very lightly and the result is a lack of congruence between test objectives and instructional objectives.

Educational institutions have essentially two choices in test selection. They can select from the numerous standardized achievement tests available on the market or they can elect to construct their own tests. The latter option is most practical for very large educational systems or at the state level due to the time and expense involved. At the state level, it would be feasible to develop test instruments that would be based upon a state-wide adopted curriculum. As a safeguard of the testing process, a test bank of items containing equivalent test questions could be developed to allow for nonredundant tests over years. Such procedures would be very time consuming and expensive.

The integrity of the testing process should be protected if teachers are to be evaluated upon their results. This matter is quite broad and complex. Among others, the following items should be considered:

A. Adequate test monitoring should be done in order to assure that students do not cheat among themselves.

B. Outside monitors should perhaps be used during the testing, to eliminate any possible biased teacher aid to students during the testing.

C. There should be adequate safeguards to assure that the actual test instruments are secure to prevent general public distribution.

D. If possible, actual test questions should not be given to the classroom teacher to eliminate possible "teaching of the test." Instead, lists of objectives, etc., could be provided to each instructor for fairness. Teachers do need to know upon what their instruction is to be evaluated.

E. Distribution frequency patterns of the actual test data should be reviewed to identify any gross "teacher manipulation" of test data (e.g., teacher changing of test answers, etc.).

The protection of the integrity of the testing process could be greatly facilitated by the creation of a central coordination office to control the annual testing and scoring procedure. Such an office would in effect allow a third party to coordinate the scoring and data reporting, a third party that in no way would be biased or have any personal interests in the test data results. Also, general educators do not need to be burdened with any additional time-consuming record-keeping. Teachers and most teacher organizations are very critical of the existing paperwork and record-keeping that burden teachers now. Teachers in effect need time to teach. A central coordinating office could be an efficient and fair mechanism for coordinating the testing program.

The use of computers and computer programs is essential in the development of a teacher evaluation system based upon student achievement data. This study has shown that computer software is

currently available for both the required editing and for performing actual required mathematical calculations. The CMS XEDIT program produced by IBM provides the required editing services, and the SAS (Statistical Analysis System) software system performs the statistical calculations required. Adequate computer software is available to implement such an evaluation system. It is maintained, however, that an individual system or educational unit would need to improve and individualize existing computing techniques to meet its individual needs and to increase speed and efficiency of calculation.

If the following four assumptions are met, it is highly likely that classroom teachers' effectiveness can be evaluated from standardized achievement test data. A system would have to have (1) an agreed upon curriculum, (2) achievement tests devised or selected to match the curriculum, (3) safeguards established to protect the integrity of the testing process, and (4) adequate computer programs and services available for editing of records and for performing statistical analysis.

3. As the instructional leader of a school, the building principal has the greatest knowledge and daily experience of any administrator in the educational organization to most accurately subjectively rank classroom teachers.

Discussion

As a source of corroboration for the results obtained in this evaluation design, a comparison was developed between the achievement based ranking and the expected rankings provided by building

principals. The principals' expected rankings and the achievement based ranking had a moderate to strong positive relationship especially in reading and math. The percentage of congruence between the two rankings was 71 percent in reading, 61 percent in math, and 46 percent in language arts. The lower percentage in language arts may be the result of less knowledge and awareness of the language arts curriculum on the part of the school principal. Also, in reading and math the system followed the state's structured Basic Skills First Program. This program with its lists of skill specific objectives, suggested activities, and criterion tests could have provided the principal with definite guidelines to direct his/her observations and eventual ranking expectations. There was no comparable program for language arts. This lack of a finite, structured program may not only be a factor in the lower percentage of agreement between the two rankings, but it may also be a partial cause of the lower consistency correlation in language arts previously discussed.

The 3 x 3 grid produced from the principal data set (Chapter IV, pages 155-157) offers some very interesting observations. Principals in the study appear to have a better idea of who are the bad or ineffective teachers (bottom 1/3) than they do the best teachers (top 1/3). They, also, do not have a good idea of the difference between the good and average teacher. These observations help to highlight the need for an objective component in teacher evaluation upon which to base or accompany the principal's subjective evaluation.

A comparison was also developed between achievement based teacher ranking and central office supervisory expected ranking. The congruency percentages of agreement for all subjects (reading, math, and language arts) were lower. The reading agreement (individual) percentage was 22 percent, math 24 percent, and language arts 44 percent. The collective consensus percentages, somewhat higher, were reading 31 percent, math 40 percent and language arts 46 percent. Individuals were allowed, in a group setting, to share knowledge of the teachers with one another and arrive at a consensus ranking. Due to the extremely high supervisor/teacher ratio, this consensus ranking may be a fairer agreement factor.

This comparison of the principal and central office supervisory agreement factors should not result in a negative connotation in regard to the central office supervisor's role. One must remember that the three supervisors are, in effect, supervising some three hundred plus teachers in this particular setting. The principal, however, has daily contact with the teacher he was asked to rank. As in many similar settings, the central office supervisor's work load and responsibilities are too broad and his awareness of the individual classroom teacher's abilities is restricted. The results of this aspect of this research study could readily be used to highlight the need to hire additional supervisory personnel to allow for a reduction in work load to facilitate an improved instructional relationship between supervisor and teacher.

These lower percentages of agreement, however, helped to pinpoint the instructional leader of the schools included within this study. The building principal appeared to be the key individual with the greatest knowledge and daily experience with the instructional program within the educational organization studied. This individual had the expertise required to most accurately subjectively rank the classroom teachers as seen by the higher percentages of agreement found between rankings. This finding was contrary to the findings of another major study. In the Sanders and McLean study of 1983 the researchers state:

> . . . Ten teachers were selected from the top and bottom of each of the three grade levels; all supervisors agreed that the teachers at the top of the list were excellent teachers based upon their observations. This agreement was not as good for the bottom of the list. They agreed with about 90 percent of the low selections.[129]

However, differences in the procedures used in making the comparisons must be considered. Also, one would need to study the organizational structure of the two separate educational systems (goals, job descriptions of both principal and supervisory staff and number of supervisory personnel) in order to arrive at a valid comparison between the two studies. For the purposes of this study, strong corroboration was found with the principals' expected rankings. The building principals within this school system had the greater knowledge upon which to base their subjective rankings.

4. A school system does not have to move to a consecutive year achievement testing schedule and rely upon a gain score concept

---

[129]McLean and Sanders, op. cit., p. 12.

in order to implement an objective teacher evaluation system based upon achievement test data. An alternate testing achievement schedule can provide sufficient data.

Discussion

Consecutive (yearly) achievement testing is not a requirement for the implementation of a teacher evaluation system based upon student achievement data. As long as baseline achievement data are available, a system can use the analysis of covariance to calculate an adjusted mean score, independent of initial student ability, and use the Henderson BLUP procedure to control for the regression to the mean. This system could successfully be employed in school systems that practice an alternating achievement testing schedule. However, teachers at every grade level could not be evaluated yearly. Teachers at grade levels not administering an achievement battery could not be evaluated based upon their students' achievement test data. If, however, the testing schedule was planned in advance to provide baseline achievement data at every grade level on an alternating schedule, teachers at every level could be evaluated or ranked on an alternating schedule based upon achievement data. Such a rotating system could be built around a three year (or more) cycle with consecutive grade testing only on the end cycle year to insure adequate accumulation of student summative and diagnostic information. This type of plan would allow for all teachers to be evaluated or ranked based upon student achievement data every three

years (or whatever cycle period selected). The system would eliminate the additional expense and increase in testing time (loss of direct instruction) that would be required if yearly achievement tests had to be administered to implement a gain score concept plan.

5. A broader based teacher evaluation system is available. An objective component of teacher evaluation developed from achievement test data can be combined with many current evaluation procedures (subjective principal evaluation, peer evaluation, student evaluation, self evaluation) to provide a more detailed, accurate portrait of teaching behavior, useful in both formative and summative evaluation.

6. An objective component of teacher evaluation developed from achievement test data provides a mechanism of relating process and product educational research and evaluation.

Discussion

The development and refinement of this objective component of teacher evaluation makes possible a broader based teacher evaluation system. For too long teacher evaluation has relied heavily upon subjective measures such as principal checklists, peer evaluation and student evaluation. An objective component of teacher evaluation developed from achievement test data could be combined with many of these current evaluation procedures to provide a more detailed, accurate portrait of teaching behavior, useful in both formative and summative evaluation.

With the availability of this objective evaluation component to teacher evaluation, the educational system would need to examine its entire teacher evaluation system. Basic policy changes would probably result. Evaluation of teaching effectiveness based upon standardized achievement test data as this study has explored, should never be relied upon exclusively. Any evaluation policy or plan, that does so, should definitely be questioned. Evaluation of teachers based upon achievement test score data should be a component of a broader base of teacher evaluation. There are many facets of effective teaching that are not considered within an achievement based ranking of teaching effectiveness. This objective component of teacher evaluation, however, can provide objective/concrete data to substantiate or clarify other forms of teacher evaluation. Achievement test rankings could be an asset in implementing a diagnostic/prescriptive approach to teacher evaluation. For example, if a teacher ranked consistently low in the area of reading effectiveness, the building principal could examine past subjective evaluations of the teacher's instructional behavior or other data bases and hopefully identify weak areas that could possibly be resulting in poor student academic performance. These weak areas could be shared with the classroom teacher and a plan of remediation developed (formative evaluation). Eventual success or failure in obtaining these skills would become a part of the summative evaluation. The data from an achievement based teacher evaluation program would provide concrete/objective evidence to substantiate and/or clarify other evaluational data. This study

does not advocate the exclusive use of standardized student achievement data for evaluation of classroom teaching, but as a very valuable component to many current evaluational methods.

As a result of the development of a more comprehensive teacher evaluation system which incorporates the achievement based evaluation design, the development of existing merit pay plans and the growth of new merit pay plans could be fostered. One of the major criticisms and problems encountered in many merit pay plans has been the development of an objective base of summative evaluation. Teachers and other critics of merit pay have cited a lack of objective measure of evaluation in education as a major problem. Past subjective measures of teacher evaluation have been considered inadequate for decision making under a merit pay plan system. An objective component of teacher evaluation enhances the quality and reliability of summative evaluation, so often missing in many merit pay plans.

More importantly, however, a system of objectively evaluating teachers from standardized achievement data provides opportunities for improving the quality of a teacher's instructional skills. For example, if effective teachers can be identified, these individuals could be studied intensely and lists of characteristics developed on the qualities required to increase teaching effectiveness in particular subject or grade level areas. Such research results would be valuable in retraining of existing teachers (formative evaluation) through well planned in-service

growth activities and also in the initial training of new classroom teachers. Process-based research has helped guide these areas in the past. However, often certain process-related research, such as a teacher's questioning technique, has not been related to any actual objective product gain. Process educational research and evaluation focuses on aspects of teaching behavior considered to indicate teaching competence, such as teaching methods, classroom management, and student-teacher interaction. Product research and evaluation focuses on the results of teaching. For too long, process research and evaluation has been conducted in isolation. Now with an objective component of teacher evaluation possible as validated and documented by this study, it becomes possible to relate process and product research and evaluation. This improved relationship would result in improved instructional effectiveness of existing educational personnel and of future personnel through the identification and refinement of classroom teaching methods, classroom management techniques, and so forth.

This research has concluded that an objective measure of teacher performance based upon student achievement test scores is possible and quite feasible. The introduction of this objective system of evaluating teachers could prove to be a valuable addition to current educational evaluational methods and procedures. For too long, teacher evaluation has been void of any objective measures. A broader, more accurate evaluation of teaching becomes possible with the successful development of this student achievement test data

based teacher evaluation. This improved evaluation of teaching should result in improved classroom instruction.

III. RECOMMENDATIONS

Further Research

The following are the recommendations from this study.

Recommendation 1. This study should be replicated to further the collection of data on the use of achievement test data for evaluation of teaching. This study should be replicated across other grade levels and within other school systems. Hopefully, conclusions made from the current study would be validated by those additional studies and a further refined evaluation tool developed. This study has shown that a potential definitely exists for a more objective evaluation of teaching.

Recommendation 2. Further research be completed in order to refine an adjustment for non-independent samples. This study assumed that the samples in each class (students) were independent. In the real world of education this usually does not exist. For example, one teacher may be teaching reading, math, and language arts to the same group of 30 students while another teacher may be teaching the same subjects to 3 separate groups of students (90 students). There already exists some measure of correlation between the reading, math, and language arts for the first teacher due to the fact she is teaching the same student. The situation where the teacher is teaching the same students (reading, math and language arts) is different than the one teaching different students. There

is less information available for the first teacher than the second due to the already present correlation. The effectiveness measure of the teacher teaching three separate groups (90 students) will have a higher reliability measure. It was recommended for fairness and validity that further research be conducted on methods of adjusting for non-independent samples or students.

Recommendation 3. A weighting scheme should be developed to more heavily weight a teacher's most current years of teaching. In the composite scores of this study, each year was weighed equally. This study proposed that several years of data would be required to validly evaluate a teacher using achievement data. The most recent years should be most heavily weighed. It would be unfair to allow an individual who had worked hard in the past and yet in their later career years had become less instructionally effective to have these early years to count equally. Under such a system, some individuals would be able to "ride" on their past performance. Likewise, it would be unfair for the individual who had improved drastically from their early years to have them weighted equally. For fairness, a weighting scheme that provides major emphasis upon recent teaching years was advocated. Further research and development in this weighting scheme was recommended.

Recommendation 4. Further research was recommended in the area of the development of some form of confidence levels. If such a system of staff evaluation was to be implemented on a large scale, it would be important to be able to interpret and evaluate

results in terms of their levels of confidence. Without such a measure a certain amount of precision would be lost.

Recommendation 5. Micro-computer software should be developed to implement such an evaluation plan (achievement test data based). It has previously been stated within this study that computer statistical analysis was essential in implementing the evaluation plan. Further research into the use of the microcomputer for this analysis could prove to be most fruitful. The successful use of the microcomputer for the analysis required in this evaluation system would allow the implementation of the achievement based evaluation system at the individual building level. If such a system could operate at the building level, many further refinements of the actual evaluation program would probably be forthcoming. The development of microcomputer software to implement this evaluation system was recommended to allow further expansion in the use of the system.

Recommendation 6. Research should be conducted into what constitutes the difference between the effective and noneffective classroom teacher. Through using the teacher evaluation system as outlined by this study, teachers could be ranked in terms of their effectiveness. From this base, further research needs to be performed to further clarify the difference between the extremes (effective/ non-effective). Such research could be very valuable for both initial teacher training and for teacher inservice programs.

Recommendation 7. Evaluating teacher effectiveness based upon achievement data should not be relied upon exclusively. It should be viewed only as a component of a broader evaluation program. Further research into how this objective form of teacher evaluation would complement other forms of evaluation should be profitable and could possibly result in a weighting system for the various components of a multi-facet teacher evaluation program.

System Implementation

The following list of recommendations is provided for school systems to facilitate the implementation of the teacher evaluation system outlined within this study:

1. The school system should examine its complete teacher evaluation system and determine how it will use the standardized achievement test data rankings to complement its current system.

2. Students and teachers should be assigned a permanent identification number.

3. Achievement test answer forms should be revised to allow for the encoding of student and teacher identification numbers.

4. Provisions should be made for precoding of both student and teacher identification before test administration or, at least, closer teacher monitoring of this encoding of information.

5. In addition to obtaining computer print-out sheets on student achievement test results, the system should make provision to obtain student data on computer tapes.

6. Periodic review of test objectives and curriculum goals need to be conducted to assure a high level of congruence.

7. Data files should be maintained on all test forms and editions to facilitate the implementation of any standardization procedures that might be required.

BIBLIOGRAPHY

BIBLIOGRAPHY

A. PRIMARY SOURCES

Alkin, Marvin C. and Klein, Stephen P. "Accountability Defined Evaluating Teachers for Outcome Accountability." UCEA Evaluation Comment, U3, N3 (May 1972), 1-10.

Anastasi, Anne. Psychological Testing--Third Edition. London: The Macmillan Company, 1968.

Anderson, G. J. "Effects of Classroom Social Climate on Individual Learning." American Educational Research, 7 (March 1970), 135-153.

Averch, Harvey, et al. How Effective Is Schooling? Santa Monica, Calif.: Rand Corporation, 1972.

Banfield, Edward C. The Unheavenly City. 2nd ed., Boston: Little Brown, 1974.

Borich, Gary D. The Appraisal of Teaching--Concepts and Process. Reading, Massachusetts: Addison-Wesley Publications Co., 1977.

Baro, Stephen M. "An Approach to Developing Accountability Measures for the Public Schools." Phi Delta Kappan, 52, No. 4 (1970), 196-205.

Berliner, David C. "Impediments to the Study of Teacher Effectiveness." Washington, D.C.: National Institute of Education, Department of Health, Education and Welfare, 1975.

Brophy, Jere E. "Stability of Teacher Effectiveness." American Educational Research Journal, Vol. 10, No. 3 (Summer 1973), 247-250.

Coleman, James S., et al. Equity of Educational Opportunity. Washington, D.C.: U. S. Government Printing Office, 1966.

Cronbach, Lee J. and Furby, Lita. "How We Should Measure 'Change' or Should We?" Psychological Bulletin, Vol. 74, No. 1 (1970), 68-80.

Dunkin, Michael J. and Biddle, Bruce J. The Study of Teaching. New York: Holt Rinehart and Winston, Inc., 1974.

Dyer, Henry S. "Toward Objective Criteria of Professional Accountability in the Schools of New York City." Phi Delta Kappan, 51, No. 4 (1970), 206-211.

Educational Research Service. Merit Pay Plans for Teachers: Status and Descriptions. Arlington, Virginia: Educational Research Services, 1983.

Evertson, Carolyn M. and Brophy, Gere E. "High Inference Behavioral Ratings as Correlates of Teaching Effectiveness." Texas University, Texas: Research and Development Center for Teacher Education, March 1973.

Eynsenck, Hans J. The IQ Argument: Race, Intelligence and Education. New York: Library Press, 1971.

Gage, N. L., ed. Mandated Evaluation of Educators. Washington, D.C.: Capitol Publications, 1973, 111-132.

Gall, M. D. "The Problems of Student Achievement in Research on Teacher Effects." San Francisco, California: Far West Laboratory for Educational Research and Development, (Report A73-2), 1973.

Garber, M. and Ware, W. B. "The Home Environment as a Predictor of School Achievement." Theory Into Practice, 11 (June 1972), 190-195.

Gay, L. R. Educational Research Competencies for Analysis and Application. Columbus, Ohio: Charles F. Merrill Publishing Co., 1981.

Hawkins, Edward E. and Stoops, Emery. "Objective and Subjective Identification of Outstanding Elementary Teachers." The Journal of Educational Research, LIX (April 1966), 344-346.

Hays, William L. Statistics for the Social Science. New York: Rinehart and Winston, Inc., 1973.

Henderson, Charles R. and Henderson, C. R. "Analysis of Covariance in Mixed Models with Unequal Subclass Numbers." Communication in Statistics' Theory and Methods. 1979, 751-787.

Henderson, C. R. "Best Linear Unbiased Estimation and Predictions Under a Selection Model." Biometrics, 31 (1975), 423-477.

Herrnstein, Richard J. I.Q. in the Meritocracy. Boston: Little Brown, 1973.

Jencks, Christopher, et al. Inequality: A Reassessment of the Effect of Family and Schooling in America. New York: Basic Books, 1972.

Jensen, Arthur R. "How Much Can We Boost IQ and Scholastic Achievement?" Harvard Educational Review, (Winter 1969), 1-123.

Justiz, T. B. "A Reliable Measure of Teacher Effectiveness." Educational Leadership Research Supplement 3, No. 1 (1969), 49-55.

Klatt, Lawrence A., Murdick, Robert G., and Schuster, Fred E. Human Resources Management: A Behavioral Systems Approach. Homewood, Illinois: Richard D. Irwin, Inc., 1978.

Lord, F. M. "Elementary Models for Measuring Change." Problems in Measuring Change. (Edited by C. W. Harris) Madison: University of Wisconsin Press, 1963, Chapter 2, 21-38.

McGowan, William E. Classroom Effect on Fifth Grade Achievement. Dekalb, Illinois: Northern Illinois University, December 1982.

McLean, Robert A. and Sanders, William L. Objective Component of Teacher Evaluation--A Feasibility Study. December 1983.

McNeil, John D. and Popham, W. James. "The Assessment of Teacher Competence." Second Handbook of Research on Teaching. New York: Rand McNally, 1973, Chapter 7, 231-233.

McNeil, John D. and Popham, W. James. Second Handbook of Research on Teaching. New York: Rand McNally, 1973.

Millman, Jason. "Student Achievement as a Measure of Teacher Competence." Handbook of Teacher Evaluation. Beverly Hills: Sage Publication, Chapter 9, 1981, 146-167.

Morsh, J. E. and Wilder, E. W. Identifying the Effective Instructor: A Review of the Quantitative Studies 1900-1952. Chanute AFB, Ill.: Air Force Personnel and Training Center, 1954.

Mosteller, Frederick and Moyinham, David P. On Equality of Educational Opportunity. New York: Random House, 1972.

National Commission on Excellence in Education. A Nation at Risk: The Imperative for Educational Reform. (Washington, D.C.: U. S. Department of Education, 1983).

Ovard, G. F. "The Practitioner's Guide to Research Teacher Effectiveness and Accountability." NASSP Bulletin, Vol. 59 (January 1975), 87-95.

Pembroke, Eileen and Groedert, Edmund R. "What is the Key to Developing an Effective Teacher Evaluation System?" NASSP Bulletin, Vol. 66, No. 458 (December 1982), 29-37.

Popham, W. James. Alternative Teacher Assessment Strategies. University of California, Los Angeles and Instructional Exchange, 1973.

Popham, W. James. "Catch-22 for Teachers--The Penalty for Doing Well." Principal, Vol. 61, No. 5 (May 1982), 35-36.

Popham, W. James. "Teacher Evaluation--The Wrong Tests for the Right Job." Adapted from a symposium presented at a jointly sponsored session during the annual meetings of the National Council of Measurement in Education and the American Educational Research Association Los Angeles, California, April 1981. Research in Education (July 1981), 1-9.

Rosenshine, Barak. "The Stability of Teacher Effects Upon Student Achievement." Review of Educational Research, 40 (1970), 647-661.

Soar, R. S. "Optimum Teacher--Pupil Interaction for Pupil Growth." Educational Leadership Research Supplement 2 (December 1968), 275-280.

Soar, R. S. and Soar, R. M. Classroom Behavior, Pupil Characteristics and Pupil Growth for the School Year and for the Summer. University of Florida, Gainesville: National Institute of Mental Health, Education, and Welfare, 1973.

Soar, Robert S. and Soar, Ruth M. "Problems in Using Pupil Outcome for Teacher Evaluation." Washington, D.C.: National Education Association Publications, April 1975.

Solomon, D.; Bezdek, W. E.; and Rosenborg, L. Teaching Styles and Learning. Chicago: Center for the Study of Liberal Education of Adults, 1963.

Texas Education Agency. Interpretive Analysis of 1978-79 Research and Evaluation Reports Volume II. Austin, Texas, 1978, 627-650.

Veldman, Douale J. and Brophy, Jere E. "Measuring Teacher Effects on Pupil Achievement." Paper presented at a symposium of annual meeting of American Educational Research Association (New Orleans, Louisiana, February 25-March 1, 1973).

Wilson, Edward O. Socialbiology. Cambridge, Mass.: University Press, 1975.

B. SECONDARY SOURCES

Bolton, Dale L. "Implications of Pupil Growth Measures for Administrative Action." Paper presented at the annual meeting of the American Educational Research Association (Chicago, Illinois, April, 1974).

Brophy, Jere E. "Stability of Teacher Effectiveness." American Educational Research Journal. Vol. 10, No. 3 (Summer 1973), 245-252.

Fortune, Jimmie C. "Use of Classroom Distribution of Student Achievement Test Scores to Evaluate the Instructional Effectiveness of Teachers." Paper presented at the annual meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975).

Hawthorne, Phyllis. "Characteristics of and Proposed Models for State Accountability Legislation. A Report." Cooperative Accountability Project, Colorado State Department of Education (April 1973).

Klein, Stephen. "Evaluating Tests in Terms of the Information They Provide." Evaluation Comment, Vol. 2, No. 2 (June 1970), 1-6.

Lambert, Robert F. "Teacher Attitudes on Testing." The College Board Review, No. 118 (Winter 1980-81), 27-30.

McKenna, Bernard. "Task Force and Other Reports, Interim Report of the Task Force on Testing." Paper presented at the annual representative assembly of the National Education Association (Portland, Oregon, July 3-6, 1973).

Ogle, Donna and Fritts, James B. "Criterion-Referenced Reading Assessment Valuable for Process as Well as for Data." Phi Delta Kappan (May 1981), 640-641.

Ornstein, Allan C. "How Good Are Teachers in Effecting Student Outcomes?" NASSP Bulletin, Vol. 66, No. 458 (December 1982), 61-70.

Peery, Criss. The Prediction of Teacher Effectiveness: A Linear Regression Analysis of Mean Residual Student Achievement and Teacher Personality Characteristics. West Virginia: West Virginia University, 1979.

Popham, James W. "California's Precedent--Setting Teacher Evaluation Law." Educational Researcher, Vol. 1, No. 7 (July 1972), 13-14.

Popham, James W. and Baker, Eva L. "A Performance Test of Teaching Effectiveness." Paper presented at the 1966 American Educational Research Association Meeting (Chicago, Illinois, February 17-19, 1966).

Taneman, Ira Leonard. A Teaching Power Test: A Method of Using Pupil Gain as a Criterion in Teacher Evaluation. Los Angeles, Calif.: University of California, 1970.

Tarver, Linda K. and Carr, Dan B. "Teacher Proficiency and Performance in Louisiana." Paper presented at the annual meeting of the American Educational Research Association (Montreal, Quebec, April 1983).

Valette, Rebecca M. "Teaching for Mastery: The Role of Classroom Testing." Washington Association of Foreign Language Teachers (May 1970), 17-22.

Walton, Susan. "Mississippi Plan Links Schools' Achievement, Accreditation." Education Week, Vol. III, No. 33 (May 9, 1984), 1.

APPENDIX

1706 Bob White Drive
Maryville, TN 37801
May 1, 1984

Blount County School Board
Blount County Schools
Courthouse Annex
Maryville, TN 37801

Blount County School Board:

I would like to request permission to have access to achievement test records for the past five years, 1979-1984, for Blount County Schools. These records will be used in a proposed study on developing an objective evaluation of fifth grade teachers in our system. All information will be held confidential, as both teachers and students will be assigned identification numbers. A copy of the official proposal is attached to this letter.

This study will be used to satisfy dissertation requirements for my Ed.D. degree at The University of Tennessee, Knoxville. The accountability move in education is placing increasing pressure on educational administration to develop and implement a more objective measure of teacher evaluation. The topic of this study is, therefore, both relevant and timely.

Thank you for your consideration of this matter.

Sincerely,

David L. Cook

DLC/mlw

Attachment

May 21, 1984

Mr. Jim Ulseth
Tennessee State Testing Evaluation Center
The University of Tennessee
25 Health, P.E., and Recreation Building
1914 Andy Holt Avenue
Knoxville, TN 37916

Dear Mr. Ulseth:

The Blount County Board of Education granted a request by David Cook to conduct a study involving the use of achievement test data from Blount County schools at its May 10, 1984 meeting.

Please let Mr. Cook have access to any Blount County test scores.

Sincerely,

Fred Morton, Jr.
Superintendent of Schools

FM:MB

cc: David Cook

May 28, 1984

Memo: Elementary Principals

Re: Blount County Objective Teacher Evaluation Study

With permission from the Blount County school Board, a study involving the use of achievement test data for teacher evaluation is being conducted. To provide teacher/student identification, we must collect roll books for the past three years for all fifth grade teachers (1983-84, 1982-83 and 1981-82). Please collect these at your school and turn in to Mr. Miller by June 8th.

Also, please complete the enclosed rating sheet on each fifth grade teacher. Your time and help in this endeavor is appreciated.

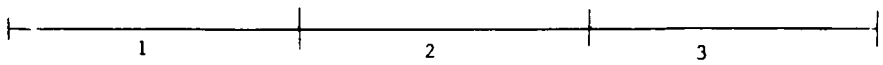
Thank you,

David L. Cook

Principal Rating Sheet

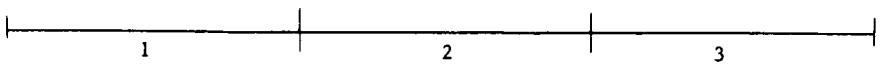
Please mark on the scale below where you believe each of your fifth grade teachers would belong in terms of their students academic gain as measured by the Stanford Achievement test. Consider the scale as a base line for all fifth grade Blount County teachers in terms of student gain in achievement. If a rank list of fifth grade teachers was developed from gains in achievement tests and divided into three levels where would you expect this teacher to fall. (One would be the highest gain, etc.)

Reading



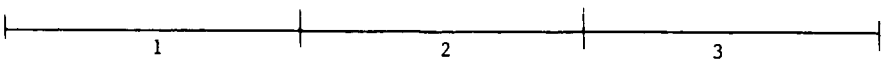
5th Grade Teacher _____ School _____

Math



5th Grade Teacher _____ School _____

Language Arts



5th Grade Teacher _____ School _____

*Teacher and students will be assigned identification numbers. All information will be confidential.

Thank you,

David L. Cook

MEMO: Elementary Supervisors
Blount County School System

RE: Independent collaboration of achievement based evaluation of teaching personnel.

. . . . Below you will find three lists of elementary (fifth grade) teachers (Reading/Math/Language Arts). These teachers have taught this subject during the past two consecutive school years (1982-83 and 1983-84) and a composite mean adjusted achievement score has been calculated for each. Using the continuium provided below, please indicate where you feel a particular teacher would fall relative to the entire population of fifth grade teachers within the Blount County School System having taught the particular subject.

Reading

	Top Third	Middle Third	Low Third
	1	2	3
1.	John Doe	1	
2.			
3.			

. . . etc.

Math

	Top Third	Middle Third	Low Third
	1	2	3
1.	John Doe	1	
2.			
3.			

. . . etc.

Language Arts

	Top Third	Middle Third	Low Third
	1	2	3
1.	John Doe	1	
2.			
3.			

. . . etc.

Language Arts Composites

1.	_____	(Lanier)	_____
2.	_____	(Friendsville)	_____
3.	_____	(Bungalow)	_____
4.	_____	(Friendsville)	_____
5.	_____	(Binfield)	_____
6.	_____	(Townsend)	_____
7.	_____	(Union)	_____
8.	_____	(Middlesettlements)	_____
9.	_____	(Hubbard)	_____
10.	_____	(Middlesettlements)	_____
11.	_____	(Lanier)	_____
12.	_____	(Hubbard)	_____
13.	_____	(Bungalow)	_____
14.	_____	(Porter)	_____
15.	_____	(Rockford)	_____
16.	_____	(Union)	_____
17.	_____	(Binfield)	_____
18.	_____	(Eagleton)	_____
19.	_____	(Eagleton)	_____
20.	_____	(Bungalow)	_____
21.	_____	(Porter)	_____
22.	_____	(Rocky Branch)	_____
23.	_____	(Forest Hill)	_____
24.	_____	(Fairview)	_____
25.	_____	(Fairview)	_____
26.	_____	(Fairview)	_____
27.	_____	(Lanier)	_____
28.	_____	(Eagleton)	_____

Reading Composites

1.	_____	(Lanier)	_____
2.	_____	(Friendsville)	_____
3.	_____	(Bungalow)	_____
4.	_____	(Friendsville)	_____
5.	_____	(Binfield)	_____
6.	_____	(Townsend)	_____
7.	_____	(Union)	_____
8.	_____	(Middlesettlements)	_____
9.	_____	(Hubbard)	_____
10.	_____	(Middlesettlements)	_____
11.	_____	(Lanier)	_____
12.	_____	(Hubbard)	_____
13.	_____	(Bungalow)	_____
14.	_____	(Bungalow)	_____
15.	_____	(Rockford)	_____
16	_____	(Union)	_____
17.	_____	(Binfield)	_____
18.	_____	(Eagleton)	_____
19.	_____	(Eagleton)	_____
20.	_____	(Porter)	_____
21.	_____	(Porter)	_____
22.	_____	(Porter)	_____
23.	_____	(Forest Hill)	_____
24.	_____	(Fairview)	_____
25.	_____	(Fairview)	_____
26.	_____	(Fairview)	_____
27.	_____	(Bungalow)	_____
28.	_____	(Lanier)	_____
29.	_____	(Bungalow)	_____
30.	_____	(Rocky Branch)	_____
31.	_____	(Bungalow)	_____
32.	_____	(Eagleton)	_____

Math Composites

1.	_____	(Lanier)	_____
2.	_____	(Friendsville)	_____
3.	_____	(Rockford)	_____
4.	_____	(Friendsville)	_____
5.	_____	(Binfield)	_____
6.	_____	(Townsend)	_____
7.	_____	(Union)	_____
8.	_____	(Middlesettlements)	_____
9.	_____	(Hubbard)	_____
10.	_____	(Middlesettlements)	_____
11.	_____	(Lanier)	_____
12.	_____	(Hubbard)	_____
13.	_____	(Bungalow)	_____
14.	_____	(Porter)	_____
15.	_____	(Forest Hill)	_____
16.	_____	(Union)	_____
17.	_____	(Binfield)	_____
18.	_____	(Rocky Branch)	_____
19.	_____	(Porter)	_____
20.	_____	(Porter)	_____
21.	_____	(Porter)	_____
22.	_____	(Fairview)	_____
23.	_____	(Lanier)	_____
24.	_____	(Bungalow)	_____
25.	_____	(Eagleton)	_____

VITA

David L. Cook was born in Maryville, Tennessee on January 24, 1951. He attended school in Maryville, Tennessee and was graduated from Everett High School in 1970. He entered The Uiversity of Tennessee, Knoxville immediately upon graduation.

In March, 1984, he received a Bachelor of Science degree from The University of Tennessee, Knoxville. He began his teaching career in the Blount County School System as a Comprehensive Developmental Classroom teacher in an elementary school in 1974. He taught for three years before becoming an elementary school principal within the Blount County School System.

He received the Master of Science degree with a major in Special Education from The University of Tennessee, Knoxville in August of 1975. In March, 1985, he received the Doctor of Education degree with a major in Educational Administration and Supervision from The University of Tennessee, Knoxville.

He is a member of the National Association of Elementary Principals, Tennessee Association of Elementary Principals, and the Tennessee Association of School Curriculum Development.