December 1979

A Comparison of the Performance of Five Randomly Selected Groups of 1978-1979 Eighth Grade Students on Five Different Stanford Achievement Test Batteries Standardized in 1929, 1940, 1952, 1964, and 1973

Vaughn D. Chambers
East Tennessee State University

INFORMATION TO USERS

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

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

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

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

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

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

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

University Microfilms International

300 N. ZEEB ROAD, ANN ARBOR, MI 48106
18 BEDFORD ROW, LONDON WC1R 4EJ, ENGLAND

CHAMBERS, VAUGHN D.

A COMPARISON OF THE PERFORMANCE OF FIVE RANDOMLY SELECTED GROUPS OF 1978-1979 EIGHTH GRADE STUDENTS ON FIVE DIFFERENT STANFORD ACHIEVEMENT TEST BATTERIES STANDARDIZED IN 1929, 1940, 1952, 1964, AND 1973

East Tennessee State University ED.D. 1979

A COMPARISON OF THE PERFORMANCE OF FIVE RANDOMLY SELECTED GROUPS OF 1978-1979 EIGHTH GRADE STUDENTS ON FIVE DIFFERENT STANFORD ACHIEVEMENT TEST BATTERIES STANDARDIZED IN 1929, 1940, 1952, 1964, and 1973

A Dissertation

Presented to

the Faculty of the Department of Supervision and Administration

East Tennessee State University

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Education

---

by

Vaughn D. Chambers

December 1979

APPROVAL

This is to certify that the Advanced Graduate Committee of

VAUGHN D. CHAMBERS

met on the

29 th day of NOVEMBER, 1979.

The committee read and examined his dissertation, supervised his defense of it in an oral examination, and decided to recommend that his study be submitted to the Graduate Council and the Dean of the School of Graduate Studies in partial fulfillment of the requirements for the degree Doctor of Education.

Floyd H. Edwards
Chairman, Advanced Graduate Committee

Glen K. Greninger

Clinton P. Moody

Charles C. Seibert

Norma E. Hankins

Signed on behalf of
the Graduate Council

Elizabeth L. McMahon
Dean, School of Graduate Studies

A COMPARISON OF THE PERFORMANCE OF FIVE RANDOMLY SELECTED GROUPS OF 1978-1979 EIGHTH GRADE STUDENTS ON FIVE DIFFERENT STANFORD ACHIEVEMENT TEST BATTERIES STANDARDIZED IN 1929, 1940, 1952, 1964, and 1973

by

Vaughn D. Chambers

The purpose of this study was to examine the test performance of five randomly selected groups of 1978 students on five different versions of the Stanford Achievement Test. Three types of comparisons were made. First, the test scores of the five groups of 1978 students in grade 8.1 were compared with each other on the 1929, 1940, 1952, 1964, and 1973 Stanford Achievement Tests. Second, the test scores of each 1978 test group were compared with the test scores of the 8.1 norming group for each test. Last, the test scores of 1978 students were compared with the test scores of students of the same age in the norming groups for the five different tests.

A total of 236 subjects from one middle school in Upper East Tennessee was used. The 236 subjects were randomly assigned to five groups. The five groups were randomly paired with the five different Stanford Achievement Tests and were tested under the same testing conditions. A computer comparison of the past achievement of the five 1978 test groups proved the groups equal in ability at the time of testing.

In making the comparisons, it was found that students in the 1978 test groups were not achieving less than students in the past in all subjects. Reading and language achievement scores were as high or higher than in the past. Mathematics scores were lower than in the past except for 1973. Recommendations for future research were given.

Institutional Review Board

This is to certify that the following study has been filed and approved by the Institutional Review Board of East Tennessee State University.

Title of Grant or Project A Comparison of the Performance of Five Randomly Selected Groups of 1978-1979 Eighth Grade Students on Five Different Stanford Achievement Test Batteries Standardized in 1929, 1940, 1952, 1964, and 1973

Principal Investigator Vaughn D. Chambers

Department Supervision and Administration

Date Submitted November 22, 1979

Principal Investigator Vaughn D. Chambers

Institutional Review Board Approval, Chairman

ACKNOWLEDGEMENTS

My appreciation is extended to the members of my doctoral committee for their patience, assistance, and contributions to the development of this dissertation, Dr. Floyd Edwards (Chairman), Dr. Gem Kate Greninger, Dr. Clinton Moody, Dr. Charles Burkett, and Dr. Norman Hankins.

To Dr. William Acuff, I also express my appreciation for suggesting an area of study and for encouragement throughout the study. Also, to Dr. William Evernden for his friendship.

To my mother, father, and sister for their successful roles in the development of a family filled with sincere support, motivation, love, and trust. My deepest appreciation is expressed to them for always having had faith in my dignity and worth.

To my wife, Peggy, and daughter, Gina, for granting me the time to do the work. I can but acknowledge our fortuitous interactions which allowed not only educational growth but also complete dispersion while retaining the faith and love for one another that I need so much.

Finally, to Lynn for insisting that I continue in spite of the obstacles and Dr. Measel who got me started in the first place.

CONTENTS

Chapter Page

Chapter Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Chapter 1

INTRODUCTION

> The 1978 classroom is supposed to be relaxed, informal, and fun-fun-fun. Compared to the classroom of 1938 or 1948, it is. The trouble? Nobody learns much of anything measurable in it any more. And all tests show it.[1]

Max Rafferty's criticism of education was not alone in the literature surveyed. Gordon Cawelti reported that national polls showed that two out of three adults believed that the quality of education was declining.[2] The critics of education were convinced that the schools of the 1970's were not equal in quality to schools of earlier years. The critics were not, however, clearly supported in their thinking by all test results.

Robert Ebel reported a decline in the scores of applicants for college admission on the Scholastic Aptitude Test of the College Entrance Examination Board (CEEB). The mean score on the verbal portion of the test dropped from 478 in the 1962-63 academic year to a mean score of 437 in 1974-75. This 41 point score drop was accompanied, he added, by a drop of 29 score points in mathematics. He further reported that a similar decline was recorded by college bound students on the American College Testing Program (ACT). Even though the tests were not compared

---

[1]Max Rafferty, "Decline and Fall of Education--Part II," The Knoxville [Tennessee] Journal, May 23, 1978, p. 6.

[2]Gordon Cawelti, "National Competency Testing: A Bogus Solution," Phi Delta Kappan, XLIX (May, 1978), 619.

with each other, the decline in scores for these two prestigious tests tended to support Rafferty's claim.[3] Evidence to the contrary, however, also existed.

Cawelti continued, in the report cited earlier, by stating that the actual evidence presented a mixed picture. While the College Entrance Examination Board and the Scholastic Aptitude Tests showed declines in scores, some other test scores increased or did not change: American College Test (science), Iowa Test of Basic Skills (early grades), National Assessment of Educational Progress (reading achievement).[4]

Charles Silberman also took an opposing view to Rafferty's. He pointed out there is remarkably little information on how much students learn from school, or on how much they know, whatever the sources of their knowledge. He gave an example of some information available from The Educational Testing Service. Comparable tests were given to roughly representative national samples of students at two different times during the postwar period; in 186 instances the results suggested an average improvement on scores by the later group tested over the group tested earlier. Finally, he added a conclusion by the Department of Health, Education and Welfare that until further evidence was presented, the tentative judgment was that children in the sixties were learning more than their older brothers and sisters learned in the fifties.[5]

---

[3]Robert L. Ebel, "Declining Scores: A Conservative Explanation," Phi Delta Kappan, LVIII (December, 1976), 306.

[4]Cawelti, p. 619.

[5]Charles E. Silberman, Crisis in the Classroom (New York: Random House Incorporated, 1970), p. 18.

Discussing satisfactory student achievement, William Hedges quoted research findings that most students (perhaps more than 90 percent) can master what we have to teach them. He then added that from studying children we find evidence of this. We observe, for example, that children coming to us in school have mastered the structure of a language by the time they enter the first grade. Hedges stated that by demonstrating their ability to do the complex and difficult task of speaking a language, children demonstrate that they are not dumb.[6]

Hedges was convinced that children were often improperly labeled by being below the median, mean, or mode for their grade level. The same type of concern was expressed in an article on one state's testing program. An article in the New Jersey Journal of Education reminded readers that test results would tell them almost anything they wanted to read into them.[7]

Oscar Lenning warned also in an article on student achievement in junior colleges that institutions should not be judged on their outputs alone, but by their outputs relating to their inputs.[8] Eric Gardner supported the concern by stating that norms represent only an appropriate level of average achievement for a particular group of students. Even then, he added, by the very definition of a norm it is expected that

---

[6]William D. Hedges, "Are Forty Percent of Our Children Really Unsatisfactory?," The Clearing House, L (May, 1977), 418.

[7]"The Results Are In--The Controversy Continues," NJEA Review, L (May, 1977), 14.

[8]Oscar T. Lenning, "Assessing Student Progress in Academic Achievement," New Directions for Community Colleges, XVIII (Summer, 1977), 15.

students will exceed it and half will fall below.[9]

Of special importance to this research was the incongruity of the reported studies of test scores used to measure achievement--was there evidence of a decrease in student achievement, or had student achievement increased? The evidence was not clear.

The Problem

The problem, the delimitations, and assumptions of this study are stated below.

Statement of the Problem

The problem of this study was to compare the performance of five randomly selected groups of eighth-grade students on the 1929, 1940, 1952, 1964, and 1973, versions of the Stanford Achievement Test and to compare each group's performance with the norms established for the test administered to the group.

Delimitations of the Study

The study was limited in the following ways:

1. The study considered only the performance of five randomly selected groups of eighth-grade students in one Upper East Tennessee school.

2. By selecting students from only one school in Upper East Tennessee, the generalization of the results of the study was possibly limited.

3. The tests were presented in their original form. No attempt was made to restate questions made obsolete by history.

---

[9]Eric F. Gardner, "Interpreting Achievement Profiles: Uses and Warnings," Journal of Research and Development in Education, X (Spring, 1977), 53.

4. No attempt was made to match the groups in terms of race, sex, or economic status with students in 1929, 1940, 1952, 1964, or 1973.

Assumptions

It was assumed in this study that:

1. the Stanford Achievement Tests measured achievement.

2. age, previous test scores, and sex could affect performance on achievement tests and should be considered.

3. random selection of the groups permitted treating the five groups as equal or as the same group.

4. all groups tested in 1978-79 were equal or were not significantly different from the norming groups in 1929, 1940, 1952, 1964, and 1973. This permitted comparisons of the performance of 1978 students with the performance of the five norming groups.

5. by converting raw scores to content scales, performances by the 1978 groups on the 1929, 1940, 1952, 1964, and 1973 Stanford Achievement Test could be compared.

6. by comparing equal groups any differences in test performances would indicate differences in abilities of students in 1978 from students in the past.

7. achievement tests were ability measures and were better predictors of achievement than attitude measures.

8. significant results could be attained by permitting random sampling to handle factors such as number of free lunches, socioeconomic factors within the schools, make up of the feeder community, and average daily attendance considerations.

9. all tests over the years were measuring achievement.

Hypotheses

Given the statement of the problem and the incongruity of conclusions drawn from the review of related literature, the following hypotheses were formulated:

H1: The scores of the group taking the 1973 test will not differ significantly from the scores of the group taking the 1964 test.

H2: The scores of the group taking the 1973 test will not differ significantly from the scores of the group taking the 1952 test.

H3: The scores of the group taking the 1973 test will not differ significantly from the scores of the group taking the 1940 test.

H4: The scores of the group taking the 1973 test will not differ significantly from the scores of the group taking the 1929 test.

H5: The scores of the students taking the 1964 test will not differ significantly from the scores of the group taking the 1952 test.

H6: The scores of the group taking the 1964 test will not differ significantly from the scores of the group taking the 1940 test.

H7: The scores of the group taking the 1964 test will not differ significantly from the scores of the group taking the 1929 test.

H8: The scores of the group taking the 1952 test will not differ significantly from the scores of the group taking the 1940 test.

H9: The scores of the group taking the 1952 test will not differ significantly from the scores of the group taking the 1929 test.

H10: The scores of the group taking the 1940 test will not differ significantly from the scores of the group taking the 1929 test.

H11: The 1978 students who take the 1973 test will achieve a grade equivalent score equal to or greater than 8.1 on the 1973 norms.

H12: The 1978 students who take the 1964 test will achieve a grade equivalent score equal to or greater than 8.1 on the 1964 norms.

H13: The 1978 students who take the 1952 test will achieve a grade equivalent score of 8.1 or above on the 1952 norms.

H14: The 1978 students who take the 1940 test will achieve a grade equivalent score of 8.1 or above on the 1940 norms.

H15: The 1978 students who take the 1929 test will achieve a grade equivalent score of 8.1 or above on the 1929 norms.

H16: The scores of 1978 students who took the 1973 test will equal to the scores of students of the same age in 1973 norming group.

H17: The scores of 1978 students who took the 1964 test will equal to the scores of the students of the same age in the 1964 norming group.

H18: The scores of 1978 students who took the 1952 test will equal to the scores of the students of the same age in the 1952 norming group.

H19: The scores of 1978 students who took the 1940 test will equal to the scores of the students of the same age in the 1940 norming group.

H20: The scores of 1978 students who took the 1929 test will equal to the scores of the students of the same age in the 1929 norming group.

Significance of the Study

Since the evidence about test performance of the students of the 1970's in comparison to their predecessors was not conclusive, a comparison of the performance of 1978 students with students from earlier years was in order. If the experimental group differences were controlled, it was assumed that any difference in performance could be attributed to differences in 1978 students and students in the 1929, 1940, 1952, 1964,

and 1973 norming groups.

John Flanagan listed two requirements for equating test scores: the tests should be as similar as possible, and the test groups should be as similar as possible to the initial national samples on which the norms were obtained.[10] This study took a different approach. Five test groups that were similar enough to be considered equal were given tests assumed to be equal. The test scores were analyzed to determine differences in the performances of 1978 students on the different tests and to determine differences in 1978 students and the initial national norming groups for each test.

James Popham and others stated that though experts may not have agreed, many educators and most citizens felt standardized achievement tests were the only instruments one should consider when determining how well schools were working.[11] In educational settings it would be important, then, to determine what test scores reveal.

Vincent Rogers and Joan Baron reported an overall downward trend in test scores in the late 1960's and early 1970's.[12] Ebel wrote that, to educators and laymen alike, these reports of score declines were disturbing. The question in the mind of every concerned citizen, he added, was, why?[13]

---

[10]John C. Flanagan, "Obtaining Useful Comparable Scores for Non-Parallel Tests and Test Batteries," Journal of Educational Measurement, I (Spring, 1964), 1-4.

[11]W. James Popham and others, Of Measurement and Mistakes, Testimony before the General Subcommittee on Education, Committee on Education and Labor, U.S. House of Representatives, Washington, D.C. (March 29, 1973), p. 3.

[12]Vincent R. Rogers and Joan Baron, "Declining Scores: A Humanistic Explanation," Phi Delta Kappan, LXIII (December, 1976), 311.

[13]Ebel, "Declining Scores: A Conservative Explanation," p. 306.

Since some writers were reporting score declines and since many citizens were concerned, more information seemed necessary to determine whether knowledge levels were decreasing.

One of the most appropriate instruments for evaluating any changes seemed to be the Stanford Achievement Test. The test first appeared in 1923 and was described in the Mental Measurements Yearbook in 1953 as the foremost test since 1923 and that with the 1952 revision, it was likely to retain its position as one of the finest achievement tests.[14] Later, the Stanford Achievement Test was described as the patriarch of the achievement test batteries in the Seventh Mental Measurements Yearbook and perhaps the most widely used test of its kind over the longest period of time.[15]

Though tests had many critics and admittedly made many mistakes in individual cases, there was a body of evidence that showed in a variety of situations, tests did a better job than other available evaluation methods.[16] This study attempted to help clarify whether students were performing as well as students in the past and attempted to explain possible changes using the Stanford Achievement Tests as evaluation instruments. The study was significant because educators and the public were concerned that test scores seemed to be declining and because evidence about possible declines was insufficient.

---

[14]Oscar K. Buros, ed., The Fourth Mental Measurements Yearbook (Highland Park: The Gryphon Press, 1953), p. 62.

[15]Oscar K. Buros, ed., The Seventh Mental Measurements Yearbook (Highland Park: The Gryphon Press, 1972), p. 46.

[16]Frederick G. Brown, Principles of Educational and Psychological Testing (Hinsdale: The Dryden Press, 1970), p. 2.

Definition of Terms

Raw data were converted to content scales for the study. The definitions in this section were included to explain the conversion.

A Standardized Norm Referenced Test

A standardized norm referenced test is a published test, accompanied by specific directions for administration and scoring, that has been given to a group of subjects representative of the group of students for whom the test was designed. The performance of any subsequent examinee can be compared with the performance of typical examinees through the use of derived scores and norms.[17]

Measurement

Measurement is the process of assigning numerals to objects, events, or people using a rule.[18]

Raw Scores

The number of items answered correctly is called the raw score. Sometimes raw scores are used in test analysis and interpretation. Raw scores usually, however, are transformed to another scale and thus become derived or transformed scores.[19]

Derived or Transformed Scores

Derived or transformed scores are any scores obtained by transforming raw scores to another, more useful scale.[20]

---

[17]C. Mauritz Lindvall and Anthony J. Nitko, Measuring Pupil Achievement and Aptitude (2d ed.; New York: Harcourt Brace Jovanovich, 1975), p. 135.

[18]Victor R. Martuza, Norm-Referenced and Criterion-Referenced Measurement in Education (Boston: Allyn and Bacon, 1977), p. 1.

[19]Brown, p. 14.

[20]Brown, p. 14.

Content Scales

Content scales compare an individual's performance to some ideal performance. They represent the closest approximation available in educational and psychological testing. They are used infrequently--only with achievement tests. The simplest type is percentage correct:

items correct ÷ total items X 100 = percentage correct.[21]

A Subtest

A subtest is one test in the set of subtests which make up an achievement test. Scores on the subtests may be combined to obtain a total score, or they may be treated separately.[22]

Abbreviations for Hypotheses

A capital H and an arabic number were combined as an abbreviation for each hypothesis. H1, for example, was the abbreviation for the first hypothesis of the study and H15 was the abbreviation for the fifteenth hypothesis.

Organization of the Study

The study was organized into five chapters. Chapter 1 contains an introduction to the study, a statement of the problem, delimitations of the study, assumptions of the study, research hypotheses, and significance of the study. Definitions of terms and organization of the study were also included.

A review of the related literature is presented in Chapter 2.

Procedures by which the study was conducted are described in Chapter 3.

---

[21]Brown, pp. 163-164.

[22]Brown, pp. 87-88.

An analysis of the findings of the study is presented in Chapter 4. The summary, conclusions, and recommendations of the study are included in Chapter 5.

Chapter 2

REVIEW OF RELATED LITERATURE

Introduction

The review of related literature revealed a very limited number of studies designed to compare student achievement over a period of years. Critics who translated the decline of ACT and SAT scores into a general decline in achievement were easy enough to find but so were supporters of the education establishment who argued that declines in SAT and ACT scores did not indicate a decline in overall achievement. A void was discovered in the study of achievement other than ACT and SAT over periods of time long enough to indicate trends.

The purpose of the review of literature was to summarize the literature related to achievement trends over the years and to summarize the most significant studies. The following sections present the summaries.

Comparing and Interpreting Test Scores

As early as 1930 Truman Kelley addressed the problem of interpreting test scores or results. Responding to an article a year earlier by Dr. Guy Wilson, he disagreed with Wilson's statement that the first fundamental criterion of a test should be to serve the curricular aim of the subject being tested and the second should be to reinforce good methods of teaching. He claimed the main value of the Stanford Spelling Test, for

example, was that of proper classification. To classify he argued tests must include questions of such varying degrees of difficulty that not all eighth graders commonly know them. The different points of view of these two men pointed to a general disagreement in 1930, as to the purpose of achievement tests.[1]

In 1935, T. C. Foran and Edmund Loyes found that the New Stanford, the Modern School, the Metropolitan, and the Unit Attainment Scale were the four general achievement tests for use in the elementary schools. Rather than question the purpose of tests, they reported that they compared the tests and found the tests to vary in difficulty and that difficulties varied not only between the tests but they also varied with the subject measured on each test. The conclusion was that identification of skills and deficiencies by means of any one of these tests risked contradiction by some other test.[2]

Writing in 1961, Warren Findley offered the thesis that the way tests are used and interpreted in the ongoing process of education in schools is another dimension of the validity of test results. He theorized that testing for achievement by standardized tests at annual intervals makes for a comparability not attained when testing is done within the school year. His concern for measuring the individual fairly at his level of confidence caused him to suggest the practice of measuring the achievement

---

[1]Truman L. Kelley, "A Communication Concerning Difficulty of Achievement Test Scores," Journal of Education Research, XXVI (November, 1930), 309-314.

[2]T. C. Foran and M. Edmund Loyes, "The Relative Difficulty of Three Achievement Examinations," Journal of Educational Psychology, XXVI (March, 1935), 218-222.

of all children in the same grade with the same test.[3] Findley's thesis was not the most popular approach to test use.

At the 1964 annual meeting, the National Council on Measurement in Education offered a symposium, The Equating of Non-Parallel Test Scores, which addressed the greatest concern in analyzing test scores. Flanagan, one of the speakers, recognized a demand for comparable scores for various tests and test batteries used for the same purpose and added his view that the difficulties, limitations, and likelihood for misinterpretation of comparable scores from non-parallel tests had produced a real problem.[4]

Two basic requirements for obtaining comparable scores, according to Flanagan, were that the content of the test or combination of tests should be as similar as possible and that the sample used for equating should be as similar as possible to the initial national sample on which the norms were obtained. If either condition was fully met, he stated, the other could be ignored.[5]

E. F. Lindquist followed, with what he considered a foregone conclusion, that we could in a certain sense establish comparable scales, but we could not use non-parallel tests interchangeably.[6]

William Angoff followed Lindquist and discussed the technical problems of conversion of scores obtained on one test to the score scale of the test

---

[3]Warren G. Findley, "Use and Interpretation of Achievement Tests in Relation to Validity," Yearbook: National Council on Measurement in Education, XVIII (Spring, 1961), 23-24.

[4]John C. Flanagan, "Obtaining Useful Comparable Scores for Non-Parallel Tests and Test Batteries," Journal of Educational Measurement, I (Spring, 1964), 1.

[5]Flanagan, p. 2.

[6]E. F. Lindquist, "Equating Scores on Non-Parallel Tests," Journal of Educational Measurement, I (Spring, 1964), 9.

of another publisher. He stressed that the enumeration of limitations in the use of comparable scores should not be considered a wholesale condemnation of their use.[7] Robert Lennon explained the development of comparison tables using an anchor test. The best anchor test would be an I.Q. test since a high correlation was common between I.Q. and achievement tests. He suggested further studies to determine the direct equivalence between various pairs or among combinations of tests to determine the goodness of the anchor-test approach.[8]

In spite of the fact that many authors pointed out the problems of test interpretation for educators, Popham testified at a congressional hearing that many educators and most citizens assume that standardized achievement tests are the only respectable instruments one should use when attempting to find out how well schools are working.[9] If Popham were correct, educators and the public would accept test results as an acceptable means of judging student performance in the schools. The literature search was then directed at whether test scores were low and whether scores were declining.

Low or Declining Test Scores--Achievement

Lester Paldy reported the fact that blacks, poor whites, Hispanos, and females performed consistently below national averages at the three levels

---

[7]William H. Angoff, "Technical Problems of Obtaining Equivalent Scores on Tests," Journal of Educational Measurement, I (1964), 12.

[8]Robert T. Lennon, "Equating Non-Parallel Tests," Journal of Educational Measurement, I (1964), 18.

[9]W. James Popham and others, Of Measurement and Mistakes, Testimony before the General Subcommittee on Education, Committee on Education and Labor, U.S. House of Representatives, Washington, D.C. (March 29, 1973), p. 3.

tested in the 1978 National Assessment Science Survey and that this was contrary to the American belief that equal opportunity is the well-spring of American democracy. He emphasized that those subjected to equal opportunities in public schools had not performed equally. The highest performances on the NAEP Science Survey were white, male, and from advantaged urban communities.[10] The fact that disadvantaged students performed below national averages on the NAEP Science Survey was an important aspect of the discussion about tests and changing test scores in America in the 1970's.

A review of the literature concerning testing and declining test scores revealed a very hot controversy as to whether scores were declining. Part of the cause for the controversy was reported by Cawelti, who reported that national polls showed two out of three adults shared the belief that the quality of education was declining.[11] Ebel reported that most of the evidence of declines in pupil achievement had come from tests. He added that educators feared that test scores might be misinterpreted and lead to unwarranted criticism. Also, he introduced the fear of the public that educators were not always doing a good job of educating their children. He warned that unless educators could develop more valid and dependable measures of pupil achievements than tests provide, the use of tests was not likely to diminish.[12]

---

[10]Lester G. Paldy, "Science Achievement Disparities 'Jarring,'" National Assessment of Educational Progress Newsletter, XII (February, 1979), 2.

[11]Gordon Cawelti, "National Competency Testing: A Bogus Solution," Phi Delta Kappan, XLIX (May, 1978), 619.

[12]Robert L. Ebel, "Declining Scores: A Conservative Explanation," Phi Delta Kappan, LVIII (December, 1976), 307.

Both the American College Testing Program (ACT) and the Scholastic Aptitude Test (SAT) score declines since 1962 were reported by Ebel. He rejected explanations that tests were more difficult, that a different kind of student was taking the tests, that fewer repeaters were taking the tests, or that there were lowered pressures for high test scores. The possibility that most people thought of first and still found most plausible and the possibility educators were most reluctant to consider was that students who took the tests in 1975 were actually less well educated than were students who took the tests in 1962.[13] Ebel emphasized the data to check the reasons were simply not available.[14]

Leo Munday, with three quotes, helped put the whole controversy in perspective. The first appeared in an article entitled "Reading Then and Now" by Mabel E. Boss.

> . . . [Many observers compare] schools today with those of "the fathers." Surveys have shown many ways in which schools may be improved. Often the work of particular schools or individual teachers has been shown to be ineffective. In an age of scientific study of how the human race educates its young, it is not uncommon for the public and the rank and file of teachers to possess the uneasy feeling that something is badly wrong. Survey after survey has revealed unsuspected inadequacy or inefficiency in American education. Both teachers and teaching have been exposed to severe public censure.[15]

The second was a statement by Harry J. Fuller, a university professor, in 1951:

> As one who is now embarking on his fifteenth year of university teaching, I am well acquainted with this decline in the quality of pre-university training, and, since I first

---

[13]Ebel, pp. 306-307.

[14]Ebel, p. 307.

[15]Leo A. Munday, "Changing Test Scores, Especially Since 1970," Phi Delta Kappan, LX (March, 1979), 496, citing Mabel E. Boss, "Reading, Then and Now," School and Society, LI (January, 1940), 62-64.

> took chalk to hand, I have sadly observed the shrinking knowledge of spelling, arithmetic, English grammar, geography, history, and science in our freshmen.[16]

The third was from pages 23 and 24 of a 1953 book: Quackery in the Public Schools.

> Shoals of comparative "proof" of achievement mean little to an employer who cannot find among recent high school graduates one girl in 20 who can write a letter or a report to a standard of literacy which was a minimum requirement for high school graduates before the . . . war.[17]

Historically, it seemed there have always been critics of schools and student performances. The belief that present day students were less well educated than students in the past was not unique to 1978-79, but what about data to support that belief?

Munday found the available information scanty at best. He reported that "then and now" studies shared several concerns: the comparability of their samples from one time period to another, the extent to which test exercises in one period are suitable for children in another, the grade-to-grade promotion practice changes, and the percentage of students who drop out over the years. In general the studies reported by Munday reflected achievement gains in the 1940's, 1950's, and early 1960's, with peaks in the mid-1960's, followed by declines until 1970, and then little change or a leveling off in the 1970's.[18]

A 1978 report on SAT scores seemed to support Munday's findings on

---

[16]Munday, p. 496, citing Harry J. Fuller, "The Emperor's New Clothes or Pruis Dementat," Scientific Monthly (January, 1951), 35.

[17]Munday, p. 496, citing Albert Lynd, Quackery in the Public Schools (Boston: Little, Brown, 1953), pp. 23, 24.

[18]Munday, pp. 498-499.

the scores in the 1970's. Education U.S.A. reported that test scores did not fall on the SAT in 1978 for the first time since 1967.[19]

The state of Tennessee published a report in 1979 which helped focus the issue concerning declining test scores in one Southern state. During Tennessee's three and one half years of assessment which started in 1975, there was no evidence, according to the report, that test scores were declining in Tennessee. While not rising, Tennessee's scores were also not declining. This was noted in the report as contrary to national media reports that test scores everywhere were declining.[20]

The literature search revealed several studies which were relevant to the topic of changing test scores. These studies could generally be described as "then and now" studies.

"Then and Now" Studies and Related Research

Introduction

The only review of "then and now" studies found in the literature was by Leo A. Munday in the March, 1979, Phi Delta Kappan. He reported thirteen "then and now" studies in his review of changing test scores.[21] The studies reported by Munday were considered relevant enough to be summarized here.

Mabel Boss did a comparison of reading in 1938, as compared to reading in 1916, in the St. Louis public schools. The study, complicated

---

[19]Education U.S.A., XXI (December 25, 1978), 129.

[20]Capsule Report: Tennessee Looks at Its Schools, 1977-78 State Education Assessment of Schools (Knoxville, Tennessee: Tennessee State Testing and Evaluation Center, 1979), p. 1.

[21]Munday, pp. 498-499.

by changes in promotion policies and age differences in groups, did not yield clear-cut findings. From grade 4 on, 1916 students were older and were better in reading than their 1938 counterparts. Age adjustments narrowed the gap, but students in 1916 were still stronger.[22]

Joseph Sligo's 1955 study compared ability and achievement in Iowa high schools over the twenty year span from 1934-1954. He found that mental ability test scores on the American Council on Education Mental Ability Test increased significantly over the twenty year period. Achievement scores on the Iowa Every-Pupil Test varied depending on the subject. In 1934, students were stronger in algebra and U.S. history, but the groups were about the same in science and English. Sligo concluded the late 1930's, 1940's, and early 1950's may have been periods of achievement gains.[23]

Two national studies were conducted by Arthur I. Gates as part of the national norming of his reading tests in 1937 and 1957. He found that children in 1957 were ahead of children in 1937 when he analyzed data by age rather than grade level. In 1937, students were stronger by grade level but were older.[24]

Benjamin Bloom's 1955 study was reported as one of the few unambiguous studies because it covered achievement between 1943 and 1955, a

---

[22]Mabel E. Boss, "Reading, Then and Now," School and Society, LI (January, 1940), 62-64.

[23]Joseph R. Sligo, "Comparisons of Achievement in Selected High School Subjects in 1934 and 1954" (PhD dissertation, University of Iowa, 1955), pp. 148-163.

[24]Arthur I. Gates, Reading Attainment in Elementary Schools: 1957 and 1937 (New York: Teachers College, Columbia University, 1961), pp. 22-23.

period generally conceded to be a time of solid achievement gains at the high school level. He found that seniors in 1955 performed higher in English, social studies, natural sciences, literary materials, and mathematics on the Tests of General Educational Development than seniors in 1943.[25]

The Iowa Test of Basic Skills was nationally normed in 1956, 1964, and 1971. These national data showed gains from 1956 to 1964 at all grades for all skills except vocabulary at grade 8. From 1964 to 1971 there was an average decline across grades 3-8 with a greater decline in the upper grades.[26]

From 1969 to 1976 California State Assessment data showed little change in reading, language, and arithmetic scores. Three Southern states not wishing to be identified by name reported assessments. Two found a leveling off or no change in scores while the third reported gains in all skill areas from 1972 to 1975. Between 1971 and 1977 declines were recorded at grades 6 and 8 on the Iowa Tests of Basic Skills. Ohio reported a general leveling off or even a slight decline in grades 8 and 10 on the Ohio Survey Test.[27]

The National Assessment of Educational Progress study results show that reading ability among young children and "functional literacy" among seventeen-year-olds increased from 1970 to 1979 but that writing mechanics scores decreased from 1969-70 to 1973-74 among nine-, thirteen-,

---

[25]Benjamin S. Bloom, "The 1955 Normative Study of the Tests of General Educational Development," The School Review, LXIV (January-December, 1956), 110-124.

[26]Munday, pp. 498-499.

[27]Munday, pp. 498-499.

and seventeen-year-olds.[28]

Roger Farr and Leo Fay completed a "then and now" reading skills study of the scores of sixth and tenth graders in 1944 and 1976 in Indiana. They found that in 1976 Indiana children in both grades did as well as their 1944 counterparts.[29] In his analysis of the Farr and Fay study, Munday concluded that the Indiana students in 1976 may have done better for two reasons. First, the drop out rate decreased from 14 percent to 5 percent which implied that more weak students may have taken the tests in 1976. Second, the 1976 students were ten months younger than the students tested in 1944. Munday added that the researchers in Indiana found students in Indiana in 1976 acquired reading skills at a younger age than those in 1944.[30]

Finally, Munday reported the results of the 1976-77 norming of the Gates-MacGinities Reading Tests. This study found small gains for fifth graders from 1964 to 1977, hardly any change in grade 6, and slight drops for grades 7 and above.[31]

Munday concluded from these studies that achievement levels of today's elementary children are probably above that of children in the historically highest periods of achievement, the 1960's, and that the only marked decline in achievement scores was at the high school level

---

[28]Roger Farr and Leo Fay, Then and Now: Reading Achievement in Indiana (1944-45 and 1976) (Bloomington: Indiana University School of Education, 1978), p. 16.

[29]Farr and Fay, pp. 101-138.

[30]Leo A. Munday, "Changing Test Scores, Especially Since 1970," Phi Delta Kappan, LX (March, 1979), pp. 498-499, citing Roger Farr and Leo Fay, Then and Now: Reading Achievement in Indiana (1944-45 and 1976) (Bloomington: Indiana University School of Education, 1978), pp. 106-107.

[31]Munday, pp. 498-499.

in the late 1960's. That decline, he stated, seemed to have ended in the 1970's.[32]

Other research reports discovered in a computer assisted search of the literature were analyzed for relevance to the study of eighth grade achievement in one eighth class in one Upper East Tennessee school in 1978-79. Studies related to this research were selected and reported.

Ability Measures vs. Attitude Measures as Predictors of Academic Achievement

A 1972 study conducted in Australia by Kevin Marjoribanks examined the differences between ability measures such as I.Q. tests and attitude measures as predictors of academic achievement. Data were collected on 396 twelve-year-old, high school students in an English provincial town. A battery of cognitive and attitude measures was administered along with two group intelligence tests commonly used in England.

Marjoribanks concluded that for each academic subject within each sex group, the ability measures were more powerful predictors of achievement than were the attitude scores. His conclusion was the basis for the assumption that previous achievement tests could be used to equate groups and that attitude surveys were not essential to equating treatment groups.[33]

---

[32]Munday, pp. 498-499.

[33]Kevin Marjoribanks, "School Attitudes, Cognitive Ability, and Academic Achievement Exhibited by Middle- and Lower-Class Black and White Elementary School Boys," Journal of Educational Psychology, LXVIII (December, 1976), 653-660.

Comparisons of Achievement and Attention of Middle and Lower Class Students

In 1977, the results of a study by Vernon Hall, John Huppertz, and Alan Levi were published. The purpose of the study was to determine whether there were differences in attending behavior between middle- and lower-class black and white boys. Race was considered because many authors had used disadvantaged and black interchangeably, and many teachers believed blacks did not behave as well as white children. Boys were used because most teachers believed boys had the most difficulty with attention in elementary school.[34]

Four groups of twenty students each were randomly selected from a larger group of 600. The four groups were given the Peabody Picture Vocabulary Test (PPVT), the Raven Coloured Progressive Matrices (RCPM), and the Test of Basic Experience (TOBE). The data were analyzed using a 2 X 2 (Race X Social Class) analysis of variance.[35]

The concluions were that lower-class children were not more disruptive or nonattentive than middle-class children. Also, the results did not show that the tests used were lower in predictive validity for disadvantaged students. The implications were that having blacks and disadvantaged students in school should not lower achievement expectations

---

[34]Vernon C. Hall, John W. Huppertz, and Alan Levi, "Attendance and Achievement Exhibited by Middle- and Lower-Class Black and White Elementary School Boys," Journal of Educational Psychology, LXIX (April, 1977), 115-120.

[35]Hall, Huppertz, and Levi, pp. 115-120.

for the school.[36] The implications, however, were somewhat contradicted by the Anita Summers and Barbara Wolfe study in Philadelphia.

Do Schools Make a Difference?

Summers and Wolfe conducted a microeconometric examination of the pupil files of 627 sixth-grade elementary school students in Philadelphia in 1970-71. They also analyzed an eighth grade sample and found the eighth grade results were similar to the sixth grade findings. Thus their findings generally applied to other groups in addition to sixth graders.[37]

Using a single equation multiple regression equation, they examined relationships of students for the three years previous to the sixth grade. Cautioning that little theory, economic or otherwise, was available to describe the determinants of educational achievement, they recorded some very interesting conclusions.[38]

They found that low-achievers, low-income, and black students do respond in terms of achievement to variations in school inputs. They further concluded that males performed at lower levels than females, that physical facilities did not have much effect on achievement, that training of teachers and principals past minimum levels had no impact on achievement, and that factors such as number of students on free lunch, student mobility, median income of feeder areas, average educational level of

---

[36] Hall, Huppertz, and Levi, pp. 115-120.

[37] Anita A. Summers and Barbara L. Wolfe, "Do Schools Make a Difference?," The American Economic Review, LXVII (September, 1977), 639-652.

[38] Summers and Wolfe, pp. 639-652.

adults in the feeder areas, and average daily attendance were not significant factors in relation to achievement.[39]

1916-1938 Reading Comparisons in St. Louis

In 1938, Boss conducted a study for the St. Louis, Missouri, Board of Education which could be classified as a "then and now" study. Boss duplicated a 1916 study by Judd and Gray which measured the achievement in reading in St. Louis. Judd and Gray selected 10,549 pupils at random and then reported on 8,928 of those selected. The 1916 students were given silent and oral reading tests in the second and fourth quarters of various grades. In 1916, in all grades after the first, girls were superior to boys in their performance on the oral reading tests.[40]

In 1938, the same tests were given to 1,156 students selected as a typical sampling of the students in St. Louis. In 1938 the performance of girls was even more superior to that of boys than it had been in 1916 on oral reading. The ranges of ages had decreased in each grade level and the students were younger. Even after age adjustments were made, children in 1916 scored higher in general than students in 1938. Boss concluded that the only significance of the differences was that they confirmed that educational practice in 1938 had departed from the procedures used in 1916.[41]

Changes in Mathematical Literacy 1950-1975

Milton Beckman's study was a comparison of scores on his mathematical literacy test in Nebraska in 1950, 1965, and 1975. Beckman constructed

---

[39]Summers and Wolfe, pp. 639-652.

[40]Boss, "Reading Then and Now," p. 63. [41]Boss, p. 64.

the literacy test in 1950 to determine how students had mastered twelve math competencies. He administered the test to 1,296 students in 1950. The same test was administered to 1,385 students in 1965-66, and 1,302 students in 1975.[42]

The findings were that considerable gain was recorded from 1950 to 1965. Students beginning the ninth grade in 1965 did as well as students finishing the ninth grade in 1950. Students in 1975 scored significantly lower than students in 1965 and slightly higher than students in 1950. The overall conclusions included one indicating students in 1975 were less literate in mathematics than students in 1965 and were about the same as students in 1950.[43]

Performance on the Tests of General Educational Development 1943-1955

An older but significant study on more than one area of achievement was conducted in 1955 to norm the Tests of General Educational Development, to check differences in scores from state to state, and to study variations in test performance in relation to other social data from state to state.[44]

At the time of the study most states used the Tests of General Educational Development (GED) for granting high school equivalency

---

[42]Milton W. Beckman, "Basic Competencies--Twenty-five Years Ago, Ten-Years Ago, and Now," Mathematics Teacher, LXXI (February, 1978), 102-106.

[43]Beckman, pp. 102-106.

[44]Bloom, "The 1955 Normative Study of the Tests of General Educational Development," pp. 110-124.

certificates to individuals with high enough scores. The first norming of the GED was in 1943. In 1955 the University of Chicago was contracted to carry out another normative study. The conditions for administering the tests in 1955 were identical to those of 1943. While the purpose of the 1955 testing was to provide new norms for the GED, the collection of data was also analyzed by Benjamin Bloom to determine answers to three questions: (1) What changes in test performance occurred from 1943 to 1955 for the country as a whole? (2) Had test performances varied from state to state? (3) How were variations in performance on the tests related to other social data about the states?[45]

Bloom recognized difficulties in insuring that parallel samples of students were involved and that conditions of student motivation were similar. Not being able to guarantee that these difficulties were relieved, Bloom did make sure similar sampling methods, the same test conditions, and the same test instructions were used.[46]

The results showed that 1955 seniors performed at higher levels on each GED subtest than did the 1943 sample. Bloom concluded that 1955 students were achieving to a greater extent the objectives measured by the GED. The greatest change was in mathematics and the least was in social studies. Bloom further concluded that high schools were doing a significantly better job in 1955 than they were doing in 1943. Finally Bloom summarized that the national level of competence as measured by The Tests of General Educational Development had risen significantly from 1943 to 1955.[47]

---

[45] Bloom, p. 112.

[46] Bloom, p. 111.

[47] Bloom, p. 124.

Then and Now Reading
Achievement in Indiana

The last study selected was the most publicized study found in the review of related literature. The researchers first stated that they recognized that comparisons of achievement of today's children to that of children of former years is difficult. Then they stated that studies are difficult to effect, school settings change, and instruments that measure development change. Comparisons, they said, must be made with caution and must consider as many variables as possible. They defended their study by saying charges made by critics of education that students do not perform as well as in the past and that schools are to blame demand that responsible comparisons be made. Farr and Fay stated that in spite of criticisms that 1978 students did not perform as well as students used to, for example, no solid proof of that existed. Their study investigated whether Indiana student performance had changed in reading from 1944 to 1976.[48]

The introduction to the Indiana study included a very good discussion of the debate over whether test scores were declining. They concluded, based upon their review of the literature, that scores of children generally rose until the mid 1960's and declined after that. The point was not made that the decline was checked in the late 1970's.[49]

The Indiana study replicated the 1944-45 statewide assessment of reading achievement conducted in Indiana for grades 6 and 10. The evaluation instrument, the Iowa Silent Reading Tests (ISRT) BM edition

---

[48]Farr and Fay, Then and Now: Reading Achievement in Indiana (1944-45 and 1976), pp. 1-141.

[49]Farr and Fay, pp. 3-20.

(1943) was given to a stratified sample of students in 1976 just as it was administered in 1944-45.[50]

The comparisons of performance in 1944-45 with those in 1976 showed little overall difference. With an adjustment for a ten month age difference, however, the younger 1976 students outscored their earlier counterparts on every subtest and on the total median score. Unadjusted scores showed a slight advantage for the 1944-45 group. Since the 1976 sophomores were fourteen months younger, however, the researchers adjusted scores to consider age. When they did, the 1976 sophomores performed significantly higher than children in 1944-45.[51]

The conclusions of the study included a conclusion that the results of the study contradicted the national alarm that students did not read as well as students in the past. They recommended further study to consider factors responsible for the change between 1944 and 1976.[52]

Historical Development of the Stanford and Other Achievement Tests

To determine how a comparison of achievement could be best made between students in 1978 with students in the past, a review of the literature on achievement testing was performed. Also, because of its availability the Stanford Achievement Test was also studied.

The first objective educational or achievement test in the United States was developed by Rice in 1895. Rice's test was a spelling test which he followed by lesser known tests in arithmetic and language.

---

[50] Farr and Fay, pp. 25-26.

[51] Farr and Fay, pp. 106-107.

[52] Farr and Fay, pp. 125-126.

Though best known for his spelling test, Rice's greatest contribution to standardized achievement testing was his objective and scientific approach to the assessment of pupil knowledge.[53]

William Mehrens and Irvin Lehmann reported that the Stone Arithmetic Reasoning Test was published in 1908. Then, they stated, Thorndike published his Scale for Handwriting of Children in 1909. Thorndike also taught many students who were later to make their contributions to achievement testing.[54] Following Thorndike's scale and beginning in 1910, a number of studies were published which indicated the unreliability of teachers' grading and led to the search for, and development of, more objective procedures for testing and grading students.[55]

In the early 1920's and 1930's an important development was the publication of test batteries. In 1923, the first standardized survey battery, the Stanford Achievement Test was published. Since that time, hundreds of achievement tests have been developed.[56]

The 1923 Stanford Achievement Test was published in Forms A and B. It reflected the chaotic state of achievement testing in that the norms for the different subtests were established at different times on different groups with different procedures. Not until the 1929 revision was it possible to use the Stanford Achievement Tests for constructing profiles of relative achievement in different subjects and to make growth studies. The care with which the 1929 Stanford Achievement Tests were

---

[53]William A. Mehrens and Irvin J. Lehmann, Standardized Tests in Education (New York: Holt, Rinehart and Winston, 1973), p. 164.

[54]Mehrens and Lehmann, p. 164.

[55]Mehrens and Lehmann, p. 164.

[56]Mehrens and Lehmann, p. 164.

constructed placed them among the very best of comparable tests.[57]

The norms of the ten tests on the 1929 Stanford were equated so that the score norms for a given age or grade were the same for all tests. For example, a score of 40 (or any other) was an equally good score on all the tests. This made the interpretation of pupils' scores much easier.[58]

After almost ten years of use, new developments in achievement testing called for a new test battery. The authors of the Stanford Achievement Test embarked upon a program for the construction of five entirely new forms of the Stanford. The 1940 revision, then, was really a new test, not a revision.[59]

The pioneer in the achievement test field was revised again in 1952 and was expected to retain its position as one of the finest available achievement tests.[60] It was reviewed in The Fifth Buros Mental Measurements Yearbook as a plodding, useful, dependable workhorse that could serve the middle-of-the-road school system well.[61] The most significant review, however, was to come in the Sixth Mental Measurements Yearbook.

---

[57]Oscar K. Buros, ed., The Third Mental Measurements Yearbook (Highland Park: The Gryphon Press, 1947), pp. 32-33.

[58]Truman L. Kelley, Giles M. Ruch, and Lewis M. Terman, New Stanford Achievement Test: Directions for Administering (New York: World Book Company, 1929), p. 2.

[59]Truman L. Kelley, Giles M. Ruch, and Lewis M. Terman, Stanford Achievement Test: Directions for Administering (New York: World Book Company, 1940), p. 2.

[60]Oscar K. Buros, ed., The Fourth Mental Measurements Yearbook (Highland Park: The Gryphon Press, 1953), p. 62.

[61]Oscar K. Buros, ed., The Fifth Mental Measurements Yearbook (Highland Park: The Gryphon Press, 1959), p. 80.

In 1965, Oscar Buros included a review of the 1953 revision of the Stanford by Miriam M. Bryan. She reported that scores on the 1953 revision, as with the various previous editions, were directly comparable to scores on earlier forms.[62] In discussing the 1964 version Bryan recommended the test for use in the analysis of group differences among school subjects and also differences in abilities of individual pupils in the various subjects for purposes of planning individualized instruction, grouping pupils for instructional purposes, determining and evaluating rate of progress, and evaluating achievement.[63] In addition to having comparable scores and good reviews, the Stanford Achievement Tests were described as the patriarch of the standardized achievement test batteries and perhaps the most widely used tests of this kind over the longest period of time.[64]

Summary

All of the related literature seemed to state that "then and now" studies were very difficult to perform. School philosophies have changed considerably over the years as have drop out rates, average daily attendance, grading procedures, promotion procedures, physical facilities, racial balance in student bodies, student motivation, and others. Comparability of groups was a concern of all researchers included in the review

---

[62]Oscar K. Buros, ed., The Sixth Mental Measurements Yearbook (Highland Park: The Gryphon Press, 1965), p. 114.

[63]Buros, The Sixth Mental Measurements Yearbook, p. 121.

[64]Oscar K. Buros, ed., The Seventh Mental Measurements Yearbook (Highland Park: The Gryphon Press, 1972), p. 46.

of literature. There was, however, an implication that studies of this type should be done.

The most used procedures seemed to be to select groups to take the same tests taken in the past while assuming that student motivation, promotion policies, grading differences, differences in racial balance, and socioeconomic variables were adequately controlled by sampling procedures or could not be controlled. For most studies scores were comparable because they were obtained by having the "now" students take the same tests as the "then" students.

Even though comparisons were considered difficult and many complex variables were almost impossible to control, some studies were made by responsible researchers. While every study indicated the limitations to making conclusions were numerous, the consensus was that achievement scores rose until the mid-1960's, declined through the early 1970's, and probably started to level off or possibly increase in the mid-1970's.

Regardless of the trends it seemed critics of education have always existed, and it has always been easy to find statements in the literature for any given time which state present day students are less well educated than students used to be. More studies were needed to clarify the situation.

Chapter 3

DESIGN OF THE STUDY

The design of the study involved the following steps: (1) selection of the subjects; (2) assignment to treatment groups; (3) selection of the tests; (4) obtaining permission to use the tests; (5) testing; and (6) analysis and interpretation of the data.

Selection of Subjects

Eighth grade students from one Upper East Tennessee school comprised the population for the study. With the approval of the school system's superintendent, the principal who was also the researcher estimated the 1978-79 eighth grade class to total almost 300 students. Using a table of random numbers and three digit student numbers from 000 to 300, five test groups were selected at random. Names of students who started the school year as eighth graders were alphabetized and assigned consecutive numbers starting at 000 and ending at 263, the number of students who reported the first day of school. All new students for the first four weeks of school were assigned numbers as they registered starting at 264 for the first to register. Students who dropped were crossed off and their numbers were dropped from the test groups. It was predetermined that all students would be given an achievement test battery, but that the names of all students who did not complete all subtests and the names of those students not having completed a seventh grade achievement test battery would be deleted from the five test groups. This procedure

yielded the 236 students who comprised the five groups listed below:

Group (1) - 48 students

Group (2) - 46 students

Group (3) - 49 students

Group (4) - 45 students

Group (5) - 48 students

Assignment to Treatment Groups

By predicting 300 eighth-grade students and by assigning 300 numbers at random to five groups, a potential experimental set of five groups of 60 students each was planned. Deleting students who dropped or had no seventh grade scores reduced the numbers as did the fact that the 263 students who actually reported when school opened was fewer than expected. The five randomly formed groups were then matched at random with five Stanford Achievement Test batteries. The groups were relabeled as groups 1 through group 5 according to the test they were paired with in the random matching. The resulting groups were as follows:

Group 1 - 49 students - 1929 Stanford Achievement Test

Group 2 - 48 students - 1940 Stanford Achievement Test

Group 3 - 45 students - 1952 Stanford Achievement Test

Group 4 - 46 students - 1964 Stanford Achievement Test

Group 5 - 48 students - 1973 Stanford Achievement Test

Figure 1 is a graphic presentation of the random selection assignment of the five student groups to the five achievement test batteries. The treatment groups are also described by the table in Appendix A.

Random Selection of Groups	Test	Renamed Test Group
(3) → → → →	1929 → → →	1
(1) → → → →	1940 → → →	2
(4) → → → →	1952 → → →	3
(2) → → → →	1964 → → →	4
(5) → → → →	1973 → → →	5

Figure 1

Graphic Representation of Random Selection of Groups and Random Assignment of Groups to Tests

Selection of the Tests

The eighth graders in the school selected for the study were scheduled to take the 1973 Stanford Achievement Test as their state achievement test for 1978-79. A review of the literature on the Stanford Achievement Test revealed that it first appeared in 1923, was in use in 1978, and was a respected achievement test. The 1923 test was revised and published in five forms in 1929. Revisions were published in 1940, 1952, 1964, and 1973. The Pyschological Corporation provided reference copies of the 1929 through the 1964 tests. The school, as mentioned above, expected to give the 1973 tests as part of their regular state testing program. By selecting the 1929 through 1973 tests for study, one test for each test group was possible should the Psychological Corporation have sufficient copies for use and should they grant permission for their use.

Obtaining Permission to Use the Stanford Achievement Tests

The Psychological Corporation provided one copy of each of the test batteries along with directions for scoring and interpreting the completed tests. They gave permission to make copies of the tests since they did not have additional loan copies. The Psychological Corporation attached only one condition to the permission to use the tests: each copy had to have "reproduced with permission" printed on the front. The condition was met and the tests were given with the knowledge and permission of the publishers.

Testing Procedures

The Stanford Achievement Tests were designed to be administered by a classroom teacher following the specific instructions printed in the directions for administering each subtest. The testing situation was designed to be identical in appearance to all students. They were told to report to the testing rooms--large group instruction rooms--where a teacher would administer the tests with assistance from three monitor teachers. The test was given in successive sittings as usual for testing at John Sevier Middle School except for two differences in routine: (1) The teachers administering the test were told that the tests were different and were asked to coordinate the giving of subtests so that, for example, all teachers gave the spelling subtests at the same time; and (2) Students in groups 1, 2, 3, and 4 were told to place their answers in the test booklets, an option provided by the directions for administering. Group 5 used computer answer sheets.

All teachers gave the tests at the same time, collected them, and returned them to the school guidance office. Tests for groups 1, 2, 3, and 4 were graded by clerical workers hired for test grading. The computer answer sheets filled in by group 5 were scored by The State of Tennessee Testing Service located on the University of Tennessee campus in Knoxville, Tennessee.

The scores for all subtests and for the total batteries were recorded along with the age, sex, and 1977-78 achievement test battery score and sent to the East Tennessee State University Computer Services Center.

Analysis and Interpretation of the Data

The statistical technique of analysis of variance was utilized in the analysis and interpretation of the raw scores after they were transformed to content scores. The analysis of variance was performed by the East Tennessee State University Computer Center. The purpose of the analysis of variance was to determine whether there was a difference in the performance of equal groups of 1978 students on different versions of Stanford Achievement Test. Consideration for age, sex, and score on the previous year's achievement test was included in the analysis. This portion of the study was to determine whether 1978 students could perform equally with each other on the 1929, 1940, 1952, 1964, and 1973 achievement tests.

The five experimental groups' scores were also converted to grade equivalent scores using the norms originally published for each test. These grade equivalent scores were graphically compared to the 8.1 grade equivalent for each subtest and for the battery to determine how 1978 students compared to students in the past. The scores were then graphed with reference to age instead of grade level to see if age made any difference.

The five experimental groups and the five Stanford Achievement Test comprised the design for the study:

O_8 R O_1 O_6 O_7

O_8 R O_2 O_6 O_7

O_8 R O_3 O_6 O_7

O_8 R O_4 O_6 O_7

O_8 R O_5 O_6 O_7

The treatment groups were observed as follows:

1. O_1 - The first experimental group (49 students) was given the 1929 Stanford Achievement Test.

2. O_2 - The second experimental group (48 students) was given the 1940 Stanford Achievement Test.

3. O_3 - The third experimental group (45 students) was given the 1952 Stanford Achievement Test.

4. O_4 - The fourth experimental group (46 students) was given the 1964 Stanford Achievement Test.

5. O_5 - The fifth experimental group (48 students) was given the 1973 Stanford Achievement Test.

6. O_6 - The age for each student was recorded.

7. O_7 - The sex was recorded for each student.

8. O_8 - The score of each student on the seventh grade achievement test was recorded.

Differences between the mean content scales earned by the groups on the subtests and the total battery were tested for statistical significance in a single classification analysis of variance (ANOVA). The Student-Newman-Keuls Multiple Range Procedure was then used to determine the specifics of any differences found. Tables 1 and 2 are examples of the way the ANOVA and Student-Newman-Keuls procedure can be exhibited together.

Appendix A is a table showing treatment groups, the tests taken, and the covariants considered. Appendix B shows the average content scale of each group for the variables included in the study. Appendix C is a sample of the type graph used to compare each group with groups in the past. The data generated by the ANOVA and the Student-Newman-Keuls

Procedure, along with information from the tables in the appendix were used to test the twenty hypotheses. The findings are presented in Chapter 4.

Table 1

Analysis of Variance for Spelling Variable

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	F Probability
Between Groups	4	34861.05	8715.26	17.27	0.0000
Within Groups	231	116567.25	504.62		
Total	235	151428.25			

Table 2

Multiple Range Test for Spelling Variable
Homogeneous Subsets for Student-Newman-Keuls Procedure
Ranges for the 0.050 Level

Subset 1			
Group	Group 1		
Mean	30.04		
Subset 2			
Group	Group 4		Group 2
Mean	42.56		51.62
Subset 3			
Group	Group 2		Group 5
Mean	51.62		59.29
Subset 4			
Group	Group 5		Group 3
Mean	59.29		63.87
Subset 5			
Group			
Mean			

Chapter 4

AN ANALYSIS OF THE FINDINGS OF THE STUDY

Introduction

The study was designed to determine whether students in 1978 were achieving less than students in the past. Three types of comparisons were used. First, the performances of five equal groups of 1978 students on the 1929, 1940, 1952, 1964, and 1973 Stanford Achievement Tests were compared. Second, the performances of the 1978 students were compared with the performances of students in the same grade (8.1) for each test battery. Last, the performances of students in 1978 were compared with the performances of students of the same age for each subtest and battery. Twenty specific hypotheses were tested in the three types of comparison. Details of the findings are included in the following sections.

The Comparison of the Performances of 1978 Students on the 1929, 1940, 1952, 1964, and 1973 Stanford Achievement Tests

Ten combinations of the five 1978 groups compared two at a time were possible. The ten combinations were transformed into the first ten hypotheses of the study:

H1: The scores of the group taking the 1973 test will not differ significantly from the scores of the group taking the 1964 test.

H2: The scores of the group taking the 1973 test will not differ significantly from the scores of the group taking the 1952 test.

H3: The scores of the group taking the 1973 test will not differ significantly from the scores of the group taking the 1940 test.

H4: The scores of the group taking the 1973 test will not differ significantly from the scores of the group taking the 1929 test.

H5: The scores of the students taking the 1964 test will not differ significantly from the scores of the group taking the 1952 test.

H6: The scores of the group taking the 1964 test will not differ significantly from the scores of the group taking the 1940 test.

H7: The scores of the group taking the 1964 test will not differ significantly from the scores of the group taking the 1929 test.

H8: The scores of the group taking the 1952 test will not differ significantly from the scores of the group taking the 1940 test.

H9: The scores of the group taking the 1952 test will not differ significantly from the scores of the group taking the 1929 test.

H10: The scores of the group taking the 1940 test will not differ significantly from the scores of the group taking the 1929 test.

The statistical technique of single classification analysis of variance was utilized to determine the relationships between the achievement of the five groups on the five achievement test batteries. The Student-Newman Keuls Multiple Range Test was used to make priori comparisons related to the ANOVA to determine specifically how the groups differed. The results of the ANOVA and the Student-Newman Keuls Multiple Range Test on total test battery performance are shown in Tables 3 and 4.

Students taking the 1929, 1973, and 1952 test batteries scored higher than students taking the 1940 and 1964 test batteries. The group taking the 1929 test scored significantly higher than the students taking

Table 3

Analysis of Variance for Total Battery Variable

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	F Probability
Between Groups	4	5504.86	1376.22	5.96	0.0001
Within Groups	231	53317.02	230.81		
Total	235	58821.88			

Table 4

Multiple Range Test for Total Battery Variable
Homogeneous Subsets for Student-Newman-Keuls Procedure
Ranges for 0.050 Level

Subset 1			
Group	Group 4	Group 2	
Mean	48.20	50.10	
Subset 2			
Group	Group 2	Group 1	
Mean	50.10	55.63	
Subset 3			
Group	Group 1	Group 5	Group 3
Mean	55.63	59.65	60.13

the 1964 test but not significantly higher than those who took the 1940 test. While not having significantly higher total battery scores than the 1940 group, the group which took the 1929 test had battery scores which were not significantly lower than battery scores of the students who took the 1973 test or the students who took the 1964 test. The ANOVA and the Student-Newman-Keuls Procedure indicated that for the total battery variable the following hypotheses should not be rejected:

H2: The scores of the group taking the 1973 test will not differ

significantly from the scores of the group taking the 1952 test.

H4: The scores of the group taking the 1973 test will not differ significantly from the scores of the group taking the 1929 test.

H6: The scores of the group taking the 1964 test will not differ significantly from the scores of the group taking the 1940 test.

H9: The scores of the group taking the 1952 test will not differ significantly from the scores of the group taking the 1929 test.

H10: The scores of the group taking the 1940 test will not differ significantly from the scores of the group taking the 1929 test.

The data generated by ANOVA and the Student-Newman Keuls Procedure indicated that the hypotheses listed below should be rejected:

H1: The scores of the group taking the 1973 test will not differ significantly from the scores of the group taking the 1964 test.

H3: The scores of the group taking the 1973 test will not differ significantly from the scores of the group taking the 1940 test.

H5: The scores of the students taking the 1964 test will not differ significantly from the scores of the group taking the 1952 test.

H7: The scores of the group taking the 1964 test will not differ significantly from the scores of the group taking the 1929 test.

H8: The scores of the group taking the 1952 test will not differ significantly from the scores of the group taking the 1940 test.

Further Findings from the ANOVA and the Student-Newman-Keuls Multiple Range Procedure

The ANOVA and the Student-Newman-Keuls Procedure were applied to the average content scales for the subtests common to the test batteries.

These comparisons are shown in Tables 5 through 28 and are discussed in the following sections.

The Vocabulary Subtest Comparisons

The 1964 Stanford Test did not include a vocabulary subtest. On the vocabulary subtests of the other four batteries, the students who took the 1940 test scored significantly lower than the students who took the 1929, the 1952, and the 1973 tests. The group scores for the 1929, 1952, and 1973 tests were not significantly different.

The Reading Comprehension Subtest Comparisons

All five test batteries included reading comprehension subtests. The groups who took the 1940 and the 1964 tests scored lower than the groups who took the 1929, 1952, and 1973 tests. Students in 1978 found the 1940 and the 1964 tests difficult.

The Average Reading Subtest Comparisons

Average reading was not recorded for the 1940 and 1964 test batteries. The 1929, 1952, and 1973 test scores were in the high group for both the vocabulary and reading comprehension subtest discussed above. With only the 1929, 1952, and 1973 tests represented on average reading no difference in scores was expected and none was found.

The Language Usage Subtest Comparisons

The Student-Newman-Keuls Procedure was applied to the language usage subtests for all five test batteries. Two subsets were found. The scores for the 1964, 1940, and 1973 groups composed the low subset. The

scores for the 1940, 1973, 1929, and 1952 groups composed the high group. The conclusion was that the group taking the 1964 language usage subtest scored significantly lower than the groups who took the 1929 and the 1952 language usage subtests.

Table 5

Analysis of Variance for Vocabulary Variable

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	F Probability
Between Groups	3	7239.84	2413.28	7.77	0.0001
Within Groups	186	57794.64	310.72		
Total	189	65034.48			

Table 6

Multiple Range Test for Vocabulary Variable
Homogeneous Subsets for Student-Newman-Keuls Procedure
Ranges for the 0.050 Level

Subset 1			
Group	Group 2		
Mean	49.04		
Subset 2			
Group	Group 5	Group 3	Group 1
Mean	57.58	60.69	65.92

Table 7

Analysis of Variance for Reading Comprehension Variable

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	F Probability
Between Groups	4	20072.82	5018.20	13.691	0.0000
Within Groups	231	84668.12	366.53		
Total	235	104740.88			

Table 8

Multiple Range Test for Reading Comprehension Variable
Homogeneous Subsets for Student-Newman-Keuls Procedure
Ranges for the 0.050 Level

Subset 1			
Group	Group 2	Group 4	
Mean	42.62	48.37	
Subset 2			
Group	Group 5	Group 3	Group 1
Mean	59.04	65.36	65.51

Table 9

Analysis of Variance for Average Reading Variable

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	F Probability
Between Groups	2	1232.66	616.33	2.13	0.1223
Within Groups	139	40153.91	288.88		
Total	141	41386.57			

Table 10

Multiple Range Test for Average Reading Variable
Homogeneous Subsets for Student-Newman-Keuls Procedure
Ranges for the 0.050 Level

Subset 1			
Group	Group 5	Group 3	Group 1
Mean	58.46	62.96	65.51

Table 11

Analysis of Variance for Language Usage Variable

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	F Probability
Between Groups	4	3323.68	830.92	4.19	0.0027
Within Groups	231	45805.09	198.29		
Total	235	49128.76			

Table 12

Multiple Range Test for Language Usage Variable
Homogeneous Subsets for Student-Newman-Keuls Procedure
Ranges for the 0.050 Level

Subset 1				
Group	Group 4	Group 2	Group 5	
Mean	57.35	61.52	62.08	
Subset 2				
Group	Group 2	Group 5	Group 1	Group 3
Mean	61.52	62.08	65.51	68.58

The Math Concepts
Subtest Comparisons

The scores on all math subtests were lower than the scores on the reading and language subtests. One of the math subtests, math concepts, was included in all batteries except the 1929 Stanford. The lowest score was recorded by the group taking the 1940 test. The groups taking the 1952 and the 1973 tests scored highest. By successfully answering only 19 percent of the questions on the 1940 math concepts subtest, students in 1978 demonstrated that the test was most difficult.

The Math Application Subtest Comparisons

Only two batteries had math application subtests. The mean score of the group which took the 1964 test was significantly lower than the mean score of the group which took the 1973 test. Math applications were not tested separately on the Stanford prior to 1964.

The Math Computation Subtest Comparisons

The math computation subtest was added to all Stanford Achievement Test batteries after the 1929 battery. A significant difference was found in the performances of each of the four groups who took the 1940, 1952, 1964, and 1973 tests. The 1940, 1964, 1952, and 1973 test groups scored 28.90, 37.54, 49.44, and 61.50 respectively. The highest score for students in 1978 was on the 1973 test. The math computation scores on all tests prior to the 1973 test were low and differed significantly from each other.

Table 13

Analysis of Variance for Math Concepts Variable

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	F Probability
Between Groups	3	39104.30	13034.77	38.37	0.0000
Within Groups	183	62170.67	339.73		
Total	186	101274.94			

Table 14

Multiple Range Test for Math Concepts Variable
Homogeneous Subsets for Student-Newman-Keuls Procedure
Ranges for the 0.050 Level

Subset 1		
Group	Group 2	
Mean	19.29	
Subset 2		
Group	Group 4	
Mean	43.54	
Subset 3		
Group	Group 3	Group 5
Mean	53.22	55.44

Table 15

Analysis of Variance for Math Application Variable

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	F Probability
Between Groups	1	17585.25	17585.25	48.86	0.0000
Within Groups	92	33109.79	359.89		
Total	93	50695.04			

Table 16

Multiple Range Test for Math Application Variable
Homogeneous Subsets for Student-Newman-Keuls Procedure
Ranges for the 0.050 Level

Subset 1	
Group	Group 4
Mean	34.83
Subset 2	
Group	Group 5
Mean	62.19

Table 17

Analysis of Variance for Math Computation Variable

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	F Probability
Between Groups	3	28880.87	9626.96	28.57	0.0000
Within Groups	183	61672.74	337.01		
Total	186	90553.56			

Table 18

Multiple Range Test for Math Computation Variable
Homogeneous Subsets for Student-Newman-Keuls Procedure
Ranges for 0.050 Level

Subset 1	
Group	Group 2
Mean	28.90
Subset 2	
Group	Group 4
Mean	37.54
Subset 3	
Group	Group 3
Mean	49.44
Subset 4	
Group	Group 5
Mean	61.50

The Total Mathematics Subtest Comparisons

Total math scores were included on only three test batteries. The 1929, 1952, and 1973 Stanford Achievement Tests had provisions for combining all math subtest scores to produce one total mathematics score. Students in 1978 scored highest on the 1973 test. They found the 1929 and 1952 tests equally difficult, and they scored significantly lower on these tests.

Table 19

Analysis of Variance for Total Mathematics Variable

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	F Probability
Between Groups	2	5756.51	2878.26	9.83	0.0001
Within Groups	139	40714.16	292.91		
Total	131	46470.66			

Table 20

Multiple Range Test for Total Mathematics Variable
Homogeneous Subsets for Student-Newman-Keuls Procedure
Ranges for 0.050 Level

Subset 1		
Group	Group 1	Group 3
Mean	44.82	51.02
Subset 2		
Group	Group 5	
Mean	60.15	

The Spelling Subtest Comparisons

All test batteries had spelling subtests. The 1929 test had an oral spelling test which was officially listed as a dictation test. Students in 1978 scored lowest on the dictation test and highest on the 1952 spelling test. Other spelling test scores were between the 1929 and the 1952 test scores.

The Social Science Subtest Comparisons

The content of the social science subtests differed more from test to test than did the content from test to test for other subjects. Social

science scores for the 1978 test groups were placed into two groups by the Student-Newman-Keuls Procedure. The low group contained the 1964, 1952, and 1940 groups. The high group contained the 1952, 1940, 1929, and 1973 groups. The comparisons indicated that the 1929 and the 1973 scores were significantly higher than the 1964 scores for social science.

The Science Subtest Comparisons

The science subtest scores except for the 1940 group scores were not significantly different. The 1940 science subtest scores were significantly higher than other science scores. Students in 1978 found the 1940 science subtest questions easier than the science questions on all the other science subtests.

The Literature Subtest Comparisons

The literature subtest comparisons revealed very little. Only the 1929, 1940, and 1952 tests had literature subtests. Students in 1978 scored lower on the 1929 literature subtest than they did on the 1940 or 1952 literature tests. The fact that the 1964 and the 1973 batteries did not have literature subtests prevented detailed then and now comparisons for literature.

Table 21

Analysis of Variance for Spelling Variable

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio	F Probability
Between Groups	4	34861.05	8715.26	17.27	0.0000
Within Groups	231	116567.25	504.62		
Total	235	151428.25			

Table 22

Multiple Range Test for Spelling Variable
Homogeneous Subsets for Student-Newman-Keuls Procedure
Ranges for 0.050 Level

Subset 1			
Group	Group 1		
Mean	30.04		
Subset 2			
Group	Group 4		Group 2
Mean	42.57		51.62
Subset 3			
Group	Group 2		Group 5
Mean	51.62		59.29
Subset 4			
Group	Group 5		Group 3
Mean	59.29		63.87

Table 23

Analysis of Variance for Social Science Variable

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	F Probability
Between Groups	4	3368.12	842.03	3.18	0.0145
Within Groups	231	61228.51	265.63		
Total	235	64596.63			

Table 24

Multiple Range Test for Social Science Variable
Homogeneous Subsets for Student-Newman-Keuls Procedure
Ranges for 0.050 Level

Subset 1				
Group	Group 4	Group 3	Group 2	
Mean	48.33	53.82	55.69	
Subset 2				
Group	Group 3	Group 2	Group 1	Group 5
Mean	53.82	55.69	57.43	59.44

Table 25

Analysis of Variance for Science Variable

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	F Probability
Between Groups	4	10598.84	2649.71	8.13	0.0000
Within Groups	231	75257.56	325.79		
Total	235	85856.38			

Table 26

Multiple Range Test for Science Variable
Homogeneous Subsets for Student-Newman-Keuls Procedure
Ranges for 0.050 Level

Subset 1				
Group	Group 4	Group 1	Group 5	Group 3
Mean	50.76	58.22	59.21	60.22
Subset 2				
Group	Group 2			
Mean	71.62			

Table 27

Analysis of Variance for Literature Variable

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	F Probability
Between Groups	2	2405.53	1202.76	5.10	0.0073
Within Groups	139	32806.17	236.02		
Total	141	35211.69			

Table 28

Multiple Range Test for Literature Variable
Homogeneous Subsets for Student-Newman-Keuls Procedure
Ranges for 0.050 Level

Subset 1		
Group	Group 1	
Mean	51.10	
Subset 2		
Group	Group 2	Group 3
Mean	58.62	60.62

A Summary of All Comparisons

Except for average reading, math application, spelling, science, and literature students taking the 1940 test had scores which placed them in the lowest subsets. Only three groups had average reading scores, and these scores indicated no significant difference in average reading scores on the 1973, 1952, and 1929 tests.

Further examination of the data revealed that students who took the 1940 test had higher science scores than students taking all other subtests except for the language usage scores earned by the students who took the 1952 test.

Except for the average reading scores, all applications of the Student-Newman-Keuls Procedure produced subsets which varied significantly. The inference was that achievement varied for all variables except average reading.

Comparisons of 1978 Students with Students in the Same Grade in 1973, 1964, 1952, 1940, and 1929

Graphs and tables were prepared for each of the five test groups to demonstrate how 1978 students performed in comparison with students in the past who were in the same grade. Each test battery had a norm group for

grade 8.1 which was the grade level of the 1978 students involved in this study.

The Comparison of the Performance of 1978 Students with 1973 Students

Table 29 shows the score, age equivalent, and grade equivalent 1978 students achieved on each subtest of the 1973 battery. The battery median and age comparisons were also included. Figure 2 was prepared to pictorially present the data in Table 29. Table 29 and Figure 2 were then used to test hypothesis H11: The 1978 students who take the 1973 test will achieve a grade equivalent score equal to or greater than 8.1 on the 1973 norms. H11 was accepted because 1978 students had grade equivalents above 8.1 on all subtests and on the total battery.

Table 29

The Performance of 1978 Students on the 1973 Test

Test	1978 Score	1973 Age Equivalent	1973 Grade Equivalent
1. Vocabulary	57.58	15.7	10.1
2. Reading Comprehension	61.50	15.2	9.8
3. Math Concepts	55.44	13.1	8.4
4. Math Applications	62.19	14.8	9.2
5. Spelling	59.29	14.1	8.7
6. Language	62.08	15.3	9.9
7. Social Science	59.44	15.1	9.7
8. Science	59.21	15.3	9.9
Total Reading	58.46	15.2	9.8
Total Math	60.15	14.2	8.8
Total Battery	59.65	14.9	9.3

Average age of 1978 students = 13.6

Average age of 1973 students = 13.7

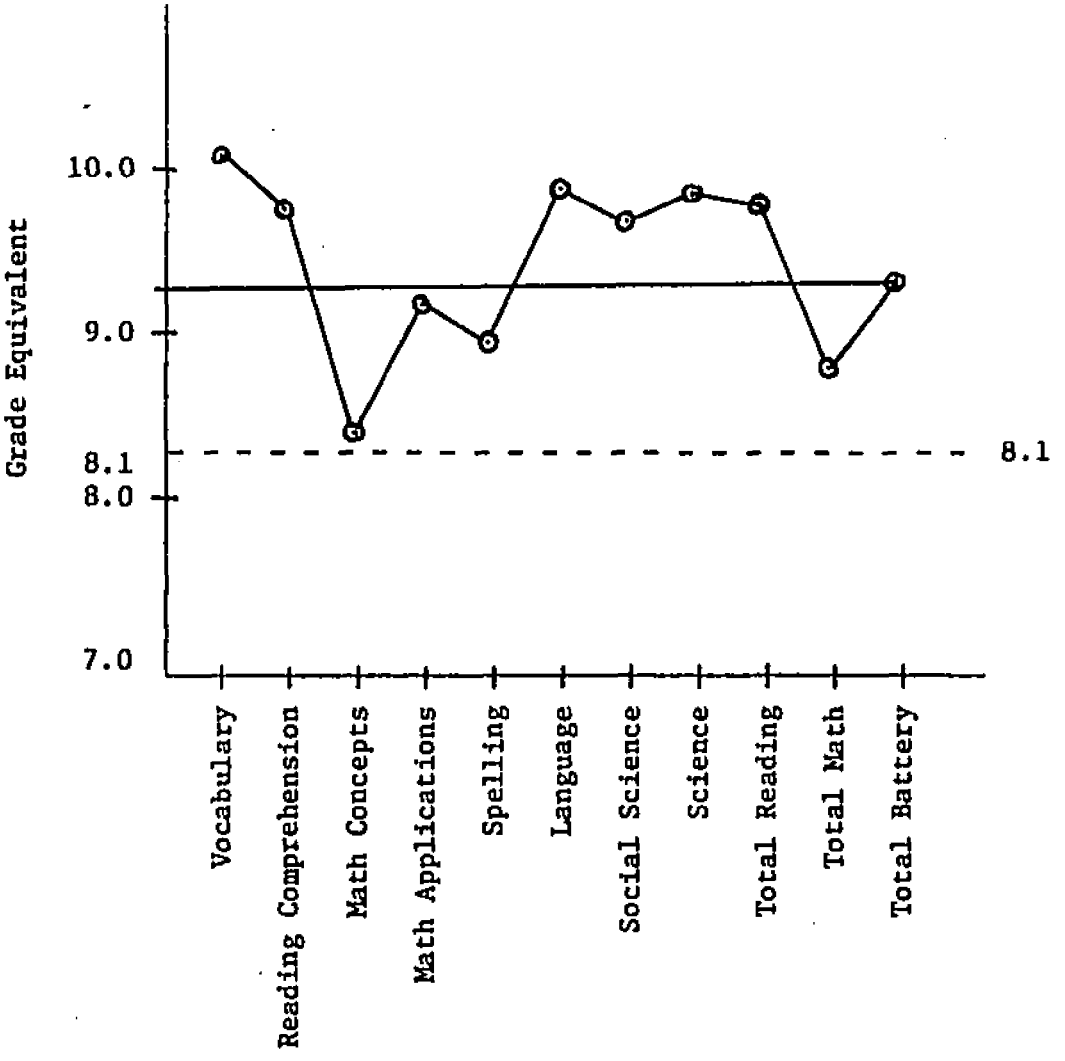

Figure 2

The Performance of the 1978 Students on the 1973 Test

The Comparison of the Performance of 1978 Students with 1964 Students

Table 30 shows the performance of 1978 students on the 1964 test. Figure 3 pictorially presents the data in Table 30. Table 30 and Figure 3 were used to test H12: The 1978 students who take the 1964 test will achieve a grade equivalent score equal to or greater than 8.1 on the 1964 norms. H12 was rejected because 1978 students failed to achieve a grade equivalent above 7.4 on any subtest and, their total battery grade equivalent was only 7.0.

Table 30

The Performance of 1978 Students on the 1964 Test

Test	1978 Score	1964 Age Equivalent	1964 Grade Equivalent
1. Reading Comprehension	48.37	12.6	7.0
2. Language Usage	57.35	12.1	6.3
3. Math Concepts	43.54	12.10	7.4
4. Math Applications	34.83	12.10	7.4
5. Math Computation	37.54	12.2	6.4
6. Spelling	42.57	12.6	7.0
7. Social Science	48.33	12.8	7.2
8. Science	50.76	12.5	6.9
Total Battery	48.20	12.6	7.0

Average age of 1978 students = 13.8

Average age of 1964 students = 13.7

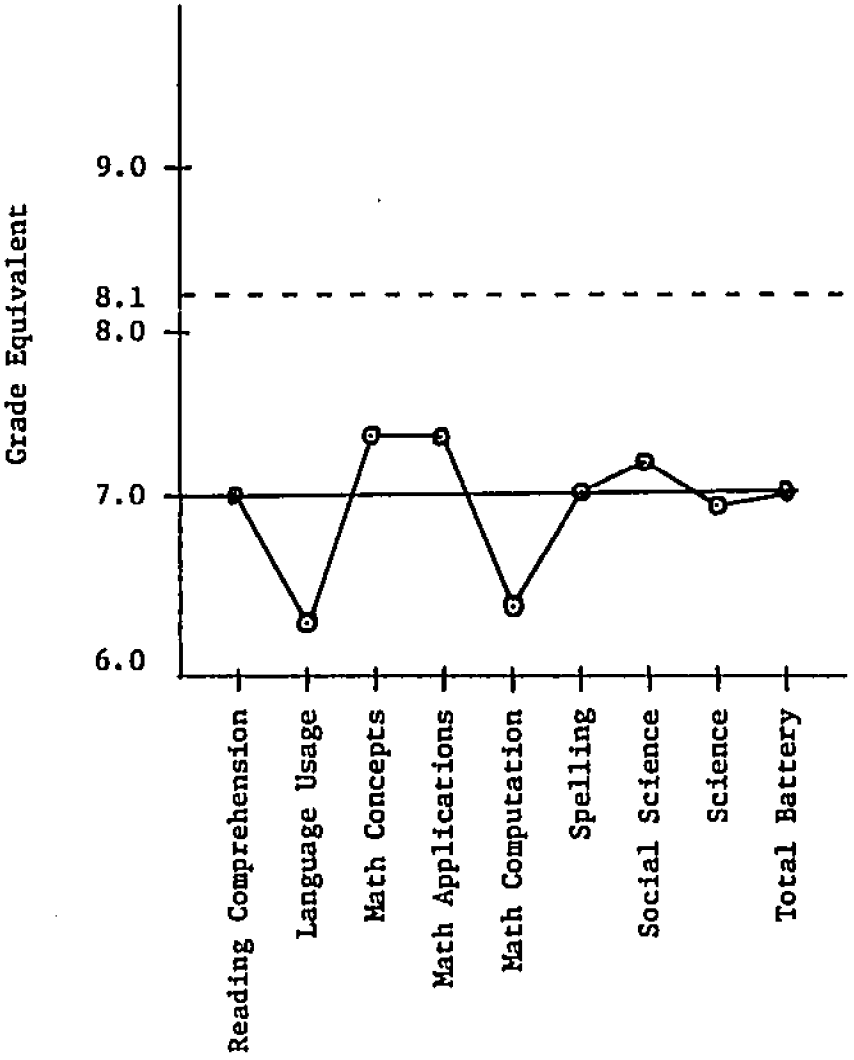

Figure 3

The Performance of 1978 Students on the 1964 Test

The Comparison of 1978 Students with 1952 Students

Table 31 and Figure 4 display data related to H13: The 1978 students who take the 1952 test will achieve a grade equivalent score of 8.1 or above on the 1952 norms. H13 was rejected for the total battery; however, subtests related to reading, spelling, and language revealed grade equivalent scores above 8.1.

Table 31

The Performance of 1978 Students on the 1952 Test

Test	1978 Score	1952 Age Equivalent	1952 Grade Equivalent
1. Paragraph Meaning	76	13 - 1	8.1
2. Word Meaning	79	13 - 4	8.4
Average Reading	78	13 - 2	8.2
3. Spelling	79	13 - 3	8.3
4. Language	111	16+	11.5
5. Arithmetic Reasoning	71	12 - 4	7.4
6. Arithmetic Computation	69	11 - 9	6.9
Average Arithmetic	70	12 - 2	7.2
7. Social Science	69	12 - 4	7.4
8. Science	68	12 - 3	7.3
9. Study Skills	60	11 - 3	6.3
Battery Median	54	12 - 11	7.9

Average age of 1978 students = 13.7

Average age of 1952 students = 13.1

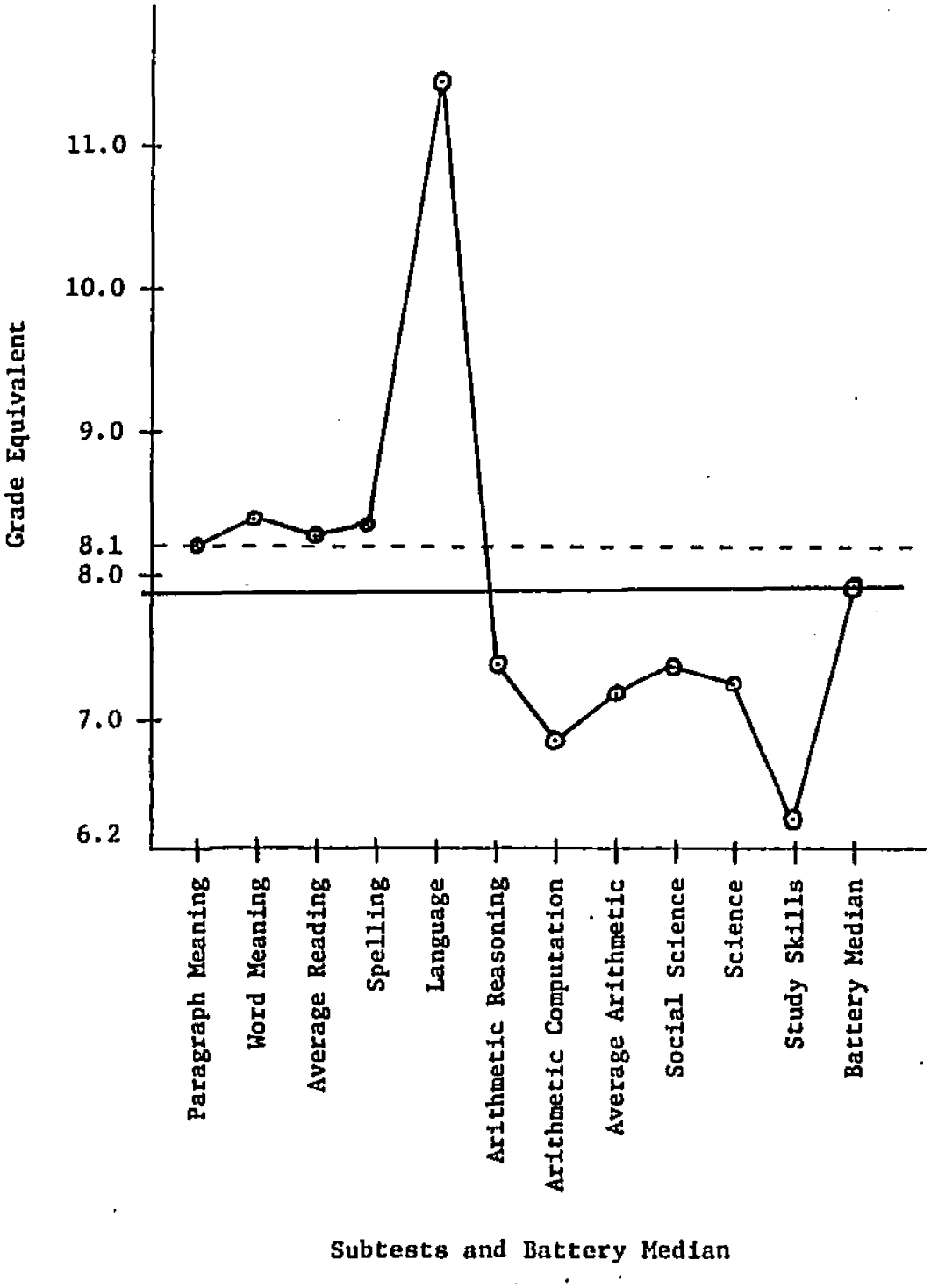

Figure 4

The Performance of 1978 Students on the 1952 Test

The Comparison of 1978 Students with 1940 Students

Table 32 and Figure 5 were used to consider H14: The 1978 students who take the 1940 test will achieve a grade equivalent score of 8.1 or above on the 1940 norms. H14 was accepted because the battery grade equivalent score was above 8.1. Arithmetic grade equivalent scores, however, were well below 8.1 for all three arithmetic subtests.

Table 32

The Performance of 1978 Students on the 1940 Test

Test	Score	Age Equivalent	Grade Equivalent
1. Paragraph Meaning	73	14 - 9	9.8
2. Word Meaning	74	15 - 0	10.0
Average Reading	73	14 - 10	9.9
3. Language Usage	101	16 - 0	11+
4. Arithmetic Reasoning	62	12 - 2	7.2
5. Arithmetic Computation	59	11 - 7	6.6
Average Arithmetic	60	11 - 10	6.9
6. Literature	74	15 - 0	10.0
7. Social Science	73	14 - 9	9.8
8. Elementary Science	78	16 - 0	11.0
9. Spelling	58	11 - 5	6.4
Total Battery	72	14 - 6	9.5

Average age of 1978 students = 13.6

Average age of 1940 students = 13.1

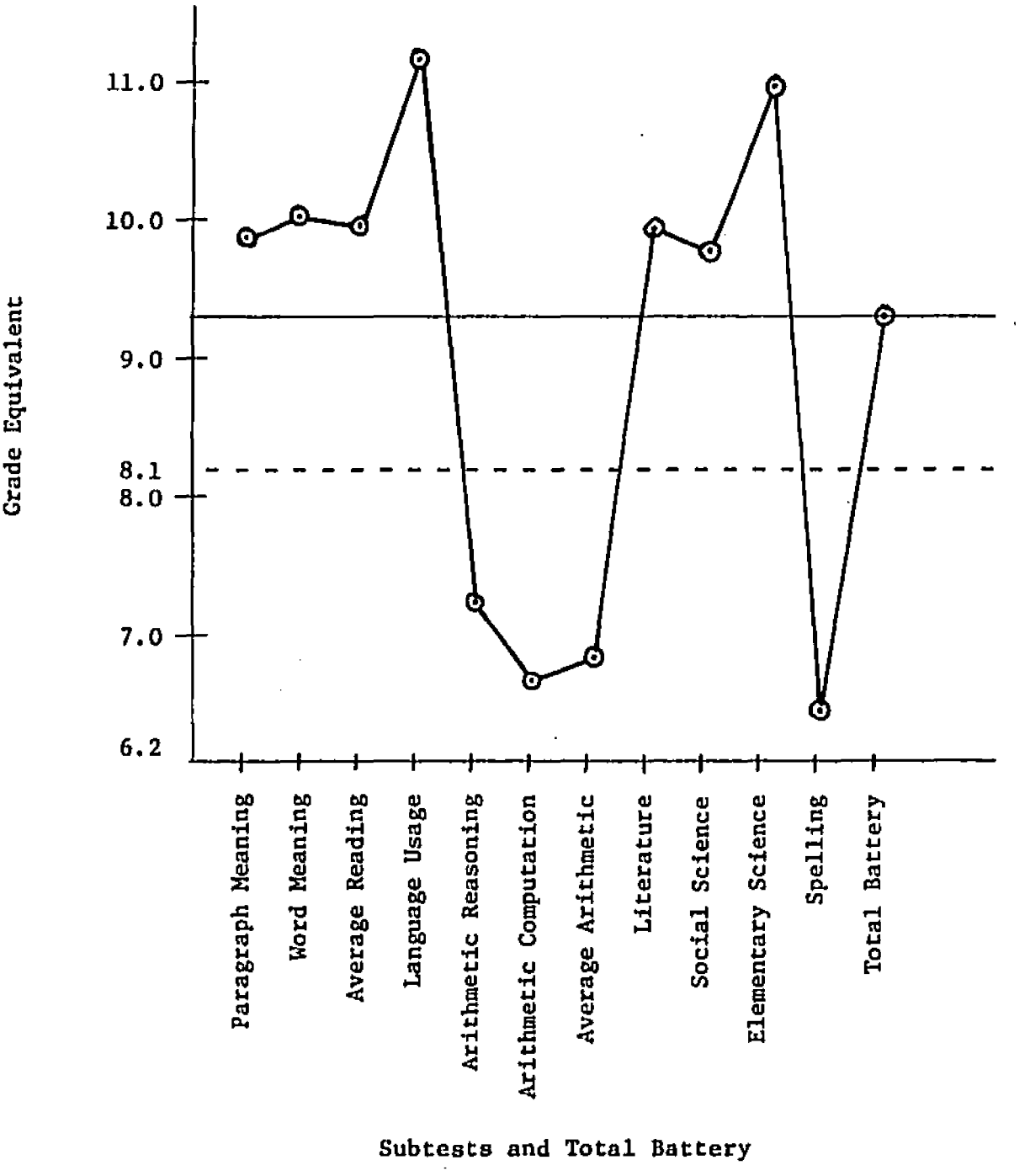

Figure 5

Performance of 1978 Students on the 1940 Test

The Comparison of 1978 Students with 1929 Students

Table 33 and Figure 6 were used to test H15: The 1978 students who take the 1929 test will achieve a grade equivalent score of 8.1 or above on the 1929 norms. H15 was rejected because the battery median was 6.7 which was below 8.1. Paragraph meaning was above 8.1 and total reading was 8.1, but all other scores were below 8.1.

Table 33

The Performance of 1978 Students on the 1929 Test

	Test	Score	Age Equivalent	Grade Equivalent
1.	Paragraph Meaning	92	14 - 4	8.4
2.	Word Meaning	89	13 - 9	7.9
	Total (Average) Reading	90	14 - 0	8.1
3.	Dictation (Spelling)	48	9 - 10	4.0
4.	Language Usage	76	12 - 0	6.2
5.	Literature	84	12 - 11	7.2
6.	Social Science	87	12 - 3	6.4
	History			
	Civics			
	Geography			
7.	Science (Physiology and Hygiene)	80	12 - 6	6.7
8.	Arithmetic Reasoning	79	12 - 4	6.6
9.	Arithmetic Computation	80	12 - 6	6.7
	Total (Average) Arithmetic	79	12 - 4	6.6
	Battery Median	80	12 - 6	6.7

Average age for 1978 students = 13.7

Average age for 1929 students = 13.11

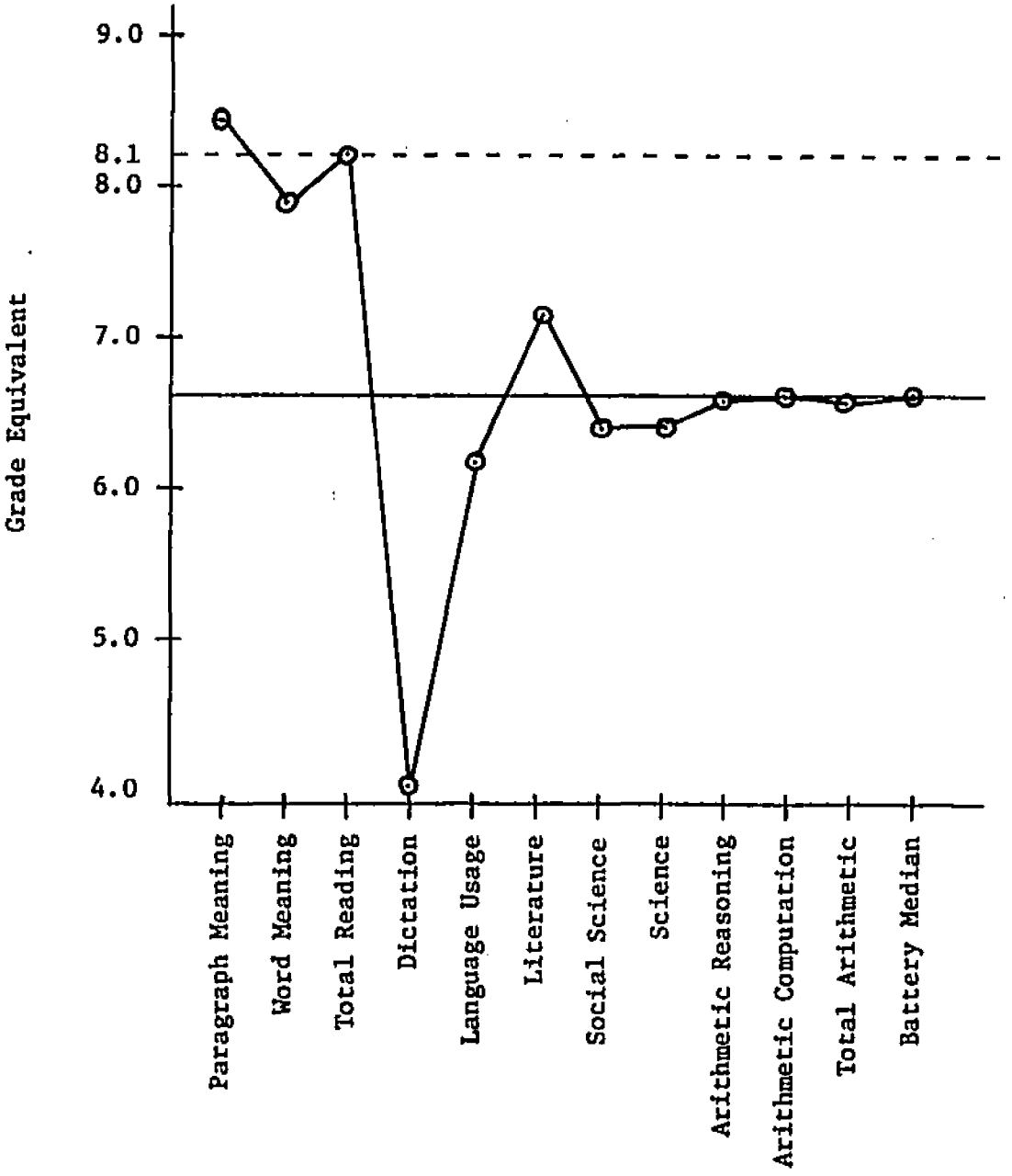

Figure 6

The Performance of 1978 Students on the 1929 Test

The Comparison of 1978 Students with 1973, 1964, 1952, 1940, and 1929 Students of the Same Age

After comparing the five 1978 groups with each other and with students in the same grade in the past, the five groups were compared in a third way. The average age of each group was computed and that age was used to compare 1978 students with students of the same age in 1929, 1940, 1952, 1964, and 1973. Tables 29 through 33 were used to compare students of the same age for the following hypotheses:

H16: The scores of 1978 students who took the 1973 test will equal to the scores of students of the same age in 1973 norming group.

H17: The scores of 1978 students who took the 1964 test will equal to the scores of the students of the same age in the 1964 norming group.

H18: The scores of 1978 students who took the 1952 test will equal to the scores of the students of the same age in the 1952 norming group.

H19: The scores of 1978 students who took the 1940 test will equal to the scores of the students of the same age in the 1940 norming group.

H20: The scores of 1978 students who took the 1929 test will equal to the scores of the students of the same age in the 1929 norming group.

The Comparison of 1978 Students with 1973 Students

Table 29 shows the 1978 group averages 13.7 years of age as compared to an average age of 13.6 for students in grade 8.1 in 1973. Since the average score for eighth graders in 1973 was 8.1 and since the 1978 group averaged the same scores as students 14.9 years old in 1973, H16 was rejected. Students in 1978 scored higher than students of the same age in 1973.

The Comparison of 1978 Students with 1964 Students

Table 30 was studied to test hypotheses H17. Eighth grade students in 1978 averaged 13.8 years of age as compared to 13.7 in 1964. Since the 1978 students scored an age equivalent of only 12.6, H17 was rejected. Students in 1978 scored lower than students of the same age in 1964.

The Comparison of 1978 Students with 1952 Students

Table 31 was used to test H18. Students in 1978 averaged 13.7 years of age compared to an average age of 13.1 for 1952 students. Since the 1978 students scored an average 12.11 on the test battery, H18 was rejected. Students in 1978 scored lower than students of the same age in 1952.

The Comparison of 1978 Students with 1940 Students

Table 32 was used to test H19. Students in 1978 were 13.6 years of age compared to 13.1 years of age as the average for eighth graders in 1940. Since the 1978 students scored 14.6 on the 1940 test, H19 was rejected. Students in 1978 scored higher than their 1940 peers.

The Comparison of 1978 Students with 1929 Students

Table 33 was used to test hypotheses H20. Students in 1978 were 13.7 years of age compared to 13.11 for eighth graders in 1929. Since the 1978 students scored 12.6 on the 1929 test, H20 was rejected. Students in 1978 scored lower than students of the same age in 1929.

Further Findings

The ANOVA and Student-Newman-Keuls Test was sent to the computer twice. The first time the ANOVA and Student-Newman-Keuls ran without regard for sex. The second time sex was entered as a covariate to see if sex made a difference. Table 34 shows the F ratio and the significance of F when sex was a covariate. Four subtests revealed that girls scored higher than boys. These were Average Reading, Language Usage, Total Math, and Spelling. On all other subtests no difference was found in the performances of boys and girls.

Table 34

A Comparison of the Performance of 1978 Boys and Girls on the Subtests of All Achievement Tests

Test	F Ratio	Significance of F
Vocabulary	1.766	0.186
Reading Comprehension	1.469	0.227
Average Reading	6.871	0.010*
Language Usage	4.802	0.029*
Math Concepts	0.048	0.826
Math Applications	0.044	0.835
Math Comprehension	1.136	0.288
Total Math	6.280	0.013*
Spelling	6.011	0.015*
Social Science	0.049	0.825
Science	1.503	0.221
Total Battery	2.646	0.105
Literature	3.186	0.077

*Sex made a difference, girls scored higher than boys.

Four other tables were prepared for the study. Tables 35 and 36 were prepared to show the ANOVA and Student-Newman-Keuls Procedure on age of the 1978 groups. Tables 37 and 38 were prepared for the ANOVA and Student-Newman-Keuls Procedure for scores the 1978 students earned on the Stanford Achievement Test given to them in 1977 for grade 7.1. These four tables revealed no significant difference in the ages of the five 1978 test groups and no significant difference in their ability as measured by the previous year's achievement test. Thus, the assumption that the groups were of equal age and ability was valid.

Table 35

Analysis of Variance for Age Variable

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	F Probability
Between Groups	4	264.28	66.07	0.934	0.4451
Within Groups	231	16344.92	70.76		
Total	235	16609.20			

Table 36

Multiple Range Test for Age Variable
Homogeneous Subsets for Student-Newman-Keuls Procedure
Ranges for 0.050 Level

Subset 1					
Group	Group 2	Group 5	Group 3	Group 1	Group 4
Mean	162.60	163.67	164.67	164.73	165.72

Table 37

Analysis of Variance for 1977 Achievement Variable

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio	F Probability
Between Groups	4	6353.35	1588.34	0.202	0.9368
Within Groups	231	1812214.56	7845.08		
Total	235	1818567.00			

Table 38

Multiple Range Test for 1977 Achievement Variable
Homogeneous Subsets for the Student-Newman-Keuls Procedure
Ranges for 0.050 Level

Subset 1					
Group	Group 3	Group 4	Group 1	Group 2	Group 5
Mean	254.07	254.96	261.51	263.06	268.02

Summary

The statistical and graphical findings were presented in this chapter. The students in 1978 performed at a higher level than students in the past years of 1973 and 1940. They did not equal 8.1 graders in 1964, 1952, and 1929. Age was not found to be a factor affecting scores, but indications were that sex should be considered on reading, language usage, total math, and spelling. In general, students in 1978 performed at a level equal to or above past students in paragraph meaning, word meaning, and reading. They generally performed below past students in mathematics and spelling. The discussion, conclusions, and recommendations were reserved for Chapter 5.

Chapter 5

DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS

Introduction

The intention of this study was to examine the test performances of five equal groups of 1978 students on five different versions of the Stanford Achievement Test. Specifically, the study compared the performances of the five 1978 groups in three ways. First, the performances of the five groups of 1978 students in grade 8.1 were compared with each other on the 1929, 1940, 1952, 1964 and 1973 Stanford Achievement Tests. Single classification ANOVA and the Student-Newman-Keuls Procedure were used to analyze the data. Second, the performance of each 1978 test group was compared with the performance of students in the grade 8.1 norming group, when the test the 1978 group took was standardized. For example, the 1978 students who took the 1929 Stanford Achievement Test were compared with the 1929 students who acted as the 8.1 norming group for that test. Last, the performance of 1978 students was compared with norming students of the same age for the five different tests. The comparison of age groups, then, was not necessarily comparisons of students in grade 8.1. For example, the 1978 group which took the 1929 Stanford Achievement Test averaged 13.7 years of age while students of that age in 1929 were in grade 7.8. The second and third ways of comparing performance involved the use of tables and figures rather than the ANOVA and the Student-Newman-Keuls Procedure.

The remaining sections of this chapter were designed to present the discussion, conclusions, and recommendations of this study.

Discussion

Table 39 was prepared from the figures and tables in Chapter 4 to aid in the discussion of the results. The scores listed in Table 39 are content scores and were used in the single classification analysis of variance. Interpretation of the data beyond accepting or rejecting the hypotheses was difficult, but it was obvious that the 1978 students found the 1940 and 1964 tests more difficult than the 1929, 1952, and 1973 tests. A study of only the 1929, 1952, and 1973 test performances would have indicated that students' performance in 1978 was not significantly different from 1929 to 1973. By including the 1940 and 1964 tests, however, it became obvious that 1978 student achievement compares differently with past student groups depending upon the particular test selected. The 1978 students involved in this study found the tests unequal in difficulty.

The 1978 students scored low in mathematics subtests for all test batteries except for the 1973 test. On the 1973 math subtests the 1978 students scored well above their grade level.

A surprise was found among the science scores. The 1978 students found the 1940 test battery difficult except for science. The 1978 students' score on the 1940 science subtests was higher than scores on all other subtests on other batteries except for the 1952 language usage subtest.

In general the 1978 students scored below the 8.1 level on the 1929,

Table 39

Scores and Ages of 1978 Test Groups and Ages of the 8.1 Groups for 1929, 1940, 1952, 1964, and 1973

Test	Group				
	1	2	3	4	5
	1929	1940	1952	1964	1973
Vocabulary (Word Meaning)	65.92 (7.9)	49.04 (10.0)	60.69 (8.4)	X	57.58 (10.1)
Reading Comprehension (Paragraph Meaning)	65.61 (8.4)	46.62 (9.8)	65.36 (8.1)	48.37 (7.0)	59.04 (9.8)
Average Reading (Total Reading)	65.51 (8.1)	X	62.96 (8.2)	X	58.46 (9.8)
Language Usage	65.51 (6.2)	61.52 (11+)	68.58 (11.5)	57.35 (6.3)	62.08 (9.9)
Math Concepts	X	19.29 (7.2)	53.22 (7.4)	43.54 (7.4)	55.44 (8.4)
Math Applications	X	X	X	34.83 (7.4)	62.19 (9.2)
Math Computation	X	28.90 (6.6)	49.44 (6.9)	37.54 (6.4)	61.50 (8.8)
Total Math (Average Math)	44.82 (6.6)	X	51.02 (7.2)	X	60.15 (8.8)
Spelling	30.04 (4.0)	51.62 (6.4)	63.87 (8.3)	42.57 (7.0)	59.22 (8.7)
Social Science	57.43 (6.4)	55.69 (9.8)	53.82 (7.4)	48.33 (7.2)	59.44 (9.7)
Science	58.22 (6.7)	71.62 (11.0)	60.22 (7.3)	50.76 (6.9)	59.21 (9.9)
Total Battery (Battery Median)	55.63 (6.7)	50.10 (9.5)	60.13 (7.9)	48.20 (7.0)	59.65 (9.3)
Literature	51.10 (7.2)	58.62 (10.0)	60.62 (7.4)	X	X
1978 Group's Age in Months Grade 8.1	164.73	162.60	164.67	165.72	163.67
1978 Group's Age in Years and Months - Grade 8.1	13.7	13.6	13.7	13.8	13.6
Past (Norming) Group's Age in Months for grade 8.1	167	168	168	163	162
Past Group's Age in Years and Months for grade 8.1	13.11	13.1	13.1	13.7	13.6

( ) = grade level
X = indicates no subtest included for this subject area in the battery

1952, and 1964 tests. They scored above the 8.1 level on the 1940 and 1973 tests. The analyses of data did not support critics' statements that achievement test scores were declining. The evidence did suggest increases and decreases in test scores. This may be more related to changes in the tests or curriculum than to decreased student ability.

One particular exception to the overall picture was in the areas of word meaning, reading comprehension, and language. Except for the 1964 test the 1978 students scored equal to or higher than grade level 8.1 on the reading and language related subtests. This indicates that except for the early 1960's students in 1978 read and use their language as well as, and probably better than, students in the past.

The conclusions of the study were based upon a thorough study of the data in relation to the twenty hypotheses.

Conclusions

The study was designed to test twenty hypotheses. The results of the study indicated the following hypotheses should be accepted:

H2: The scores of the group taking the 1973 test will not differ significantly from the scores of the group taking the 1952 test.

H4: The scores of the group taking the 1973 test will not differ significantly from the scores of the group taking the 1929 test.

H6: The scores of the group taking the 1964 test will not differ significantly from the scores of the group taking the 1940 test.

H9: The scores of the group taking the 1952 test will not differ significantly from the scores of the group taking the 1929 test.

H10: The scores of the group taking the 1940 test will not differ

significantly from the scores of the group taking the 1929 test.

H11: The 1978 students who take the 1973 test will achieve a grade equivalent score equal to or greater than 8.1 on the 1973 norms.

H14: The 1978 students who take the 1940 test will achieve a grade equivalent score of 8.1 or above on the 1940 norms.

The following hypotheses were rejected:

H1: The scores of the group taking the 1973 test will not differ significantly from the scores of the group taking the 1964 test.

H3: The scores of the group taking the 1973 test will not differ significantly from the scores of the group taking the 1940 test.

H5: The scores of the students taking the 1964 test will not differ significantly from the scores of the group taking the 1952 test.

H7: The scores of the group taking the 1964 test will not differ significantly from the scores of the group taking the 1929 test.

H8: The scores of the group taking the 1952 test will not differ significantly from the scores of the group taking the 1940 test.

H12: The 1978 students who take the 1964 test will achieve a grade equivalent score equal to or greater than 8.1 on the 1964 norms.

H13: The 1978 students who take the 1952 test will achieve a grade equivalent score of 8.1 or above on the 1952 norms.

H15: The 1978 students who take the 1929 test will achieve a grade equivalent score of 8.1 or above on the 1929 norms.

H16: The scores of 1978 students who took the 1973 test will equal to the scores of students of the same age in 1973 norming group.

H17: The scores of 1978 students who took the 1964 test will equal to the scores of the students of the same age in the 1964 norming group.

H18: The scores of 1978 students who took the 1952 test will equal to the scores of the students of the same age in the 1952 norming group.

H19: The scores of 1978 students who took the 1940 test will equal to the scores of the students of the same age in the 1940 norming group.

H20: The scores of 1978 students who took the 1929 test will equal to the scores of the students of the same age in the 1929 norming group.

The conclusions on the twenty hypotheses suggested that 1978 students were not achieving less than students in the past. Reading and language achievement scores were as high or higher than in the past. The scores of the 1978 students were higher than those of the 1973 students, possibly indicating a rise in achievement test scores in the 1970's.

Recommendations

This study established that the performances of 1978 students in one school were different from the performances of students in the past. It also proved five groups of 1978 students taking different tests from the past performed unequally. Areas which needed further investigation in reference to inferential conclusions are presented below:

1. The samples consisted of eighth graders from one upper East Tennessee school. The study should be replicated over a wider geographic area to provide greater external validity.

2. Achievement scores on the Stanford Achievement Tests were analyzed alone. Scores on other achievement tests need to be investigated.

3. The achievement tests studied were indicative of cognitive domain development. In addition to achievement tests, other types of tests

which measure not only the cognitive domain but also the affective and psychomotor domains should be investigated.

4. Some differences in the performances of males and females were found in this study. Further research should address the details of sex differences in student achievement.

5. The design of this study did not include consideration for difficulty of items or tests. Experimental designs which consider item and test difficulties should be included on other studies and the results compared to this study.

These recommendations were included not only to provide closure to this study but also to indicate the complexity of the problem. Closure was provided with respect to confining the research inferences for this study. The expanse of additional considerations suggested by the list needs investigation in order to advance information pertaining to a universal set of factors bearing upon the problem.

BIBLIOGRAPHY

BIBLIOGRAPHY

A. Books

Brown, Frederick G. Principles of Educational and Psychological Testing. Hinsdale: The Dryden Press, 1970.

Buros, Oscar K., ed. The Fifth Mental Measurements Yearbook. Highland Park, New Jersey: The Gryphon Press, 1959.

_____. The Fourth Mental Measurements Yearbook. Highland Park, New Jersey: The Gryphon Press, 1953.

_____. The Seventh Mental Measurements Yearbook. Highland Park, New Jersey: The Gryphon Press, 1972.

_____. The Sixth Mental Measurements Yearbook. Highland Park, New Jersey: The Gryphon Press, 1965.

_____. The Third Mental Measurements Yearbook. Highland Park, New Jersey: The Gryphon Press, 1947.

Gates, Arthur I. Reading Attainment in Elementary Schools: 1957 and 1937. New York: Teachers College, Columbia University, 1961.

Farr, Roger, and Leo Fay. Then and Now: Reading Achievement in Indiana (1944-45 and 1976). Bloomington: Indiana University School of Education, 1978.

Lindvall, C. Mauritz, and Anthony J. Nitko. Measuring Pupil Achievement and Aptitude. 2d ed. New York: Harcourt Brace Jovanovich, 1975.

Martuza, Victor R. Norm-Referenced and Criterion-Referenced Measurement in Education. Boston: Allyn and Bacon, 1977.

Mehrens, William A., and Irvin J. Lehmann. Standardized Tests in Education. New York: Holt, Rinehart and Winston, 1973.

Silberman, Charles E. Crisis in the Classroom. New York: Random House Incorporated, 1970.

B. Periodicals

Angoff, William H. "Technical Problems of Obtaining Equivalent Scores on Tests." Journal of Educational Measurement, I (1964), 11-13.

Beckman, Milton W. "Basic Competencies--Twenty-five Years Ago, Ten-Years Ago, and Now." Mathematics Teacher, LXXI (February, 1978), 102-106.

Bloom, Benjamin S. "The 1955 Normative Study of Tests of General Educational Development." The School Review, LXIV (January-December, 1956), 110-124.

Boss, Mabel E. "Reading Then and Now." School and Society, LI (January, 1940), 62-64.

Cawelti, Gordon. "National Competency Testing: A Bogus Solution," Phi Delta Kappan, XLIX (May, 1978), 619-621.

Ebel, Robert L. "Declining Scores: A Conservative Explanation." Phi Delta Kappan, LVIII (December, 1976), 306-310.

Education U.S.A., XXI (December 25, 1978), 129.

Findley, Warren G. "Use and Interpretation of Achievement Tests in Relation to Validity." Yearbook: National Council on Measurement in Education, XVIII (Spring, 1961), 23-24.

Flanagan, John C. "Obtaining Useful Comparable Scores for Non-Parallel Tests and Test Batteries." Journal of Educational Measurement, 1 (Spring, 1964), 1-4.

Foran, T. G., and M. Edmund Loyes. "The Relative Difficulty of Three Achievement Examinations." Journal of Educational Psychology, XXVI (March, 1935), 218-222.

Gardner, Eric F. "Interpreting Achievement Profiles: Uses and Warnings." Journal of Research and Development in Education, X (Spring, 1977), 51-63.

Hall, Vernon C., John W. Huppertz, and Alan Levi. "Attention and Achievement Exhibited by Middle- and Lower-Class Black and White Elementary School Boys." Journal of Educational Psychology, LXIX (April, 1977), 115-120.

Hedges, William D. "Are Forty Percent of Our Children Really Unsatisfactory?" The Clearing House, L (May, 1977), 417-422.

Kelley, Truman L. "A Communication Concerning Difficulty of Achievement Test Scores." Journal of Education Research, XXVI (November, 1930), 309-314.

Lenning, Oscar T. "Assessing Student Progress in Academic Achievement." New Directions for Community Colleges, XVIII (Summer, 1977), 1-20.

Lennon, Robert T. "Equating Non-Parallel Tests." Journal of Educational Measurement, I (1964), 15-18.

Lindquist, E. F. "Equating Scores on Non-Parallel Tests." Journal of Educational Measurement, I (Spring, 1964), 5-9.

Marjoribanks, Kevin. "School Attitudes, Cognitive Ability, and Academic Achievement Exhibited by Middle- and Lower-Class Black and White Elementary School Boys." Journal of Educational Psychology, LXVIII (December, 1976), 653-660.

Munday, Leo A. "Changing Test Scores, Especially Since 1970." Phi Delta Kappan, LX (March, 1979), 496-499.

Paldy, Lester G. "Science Achievement Disparities 'Jarring.'" National Assessment of Educational Progress Newsletter, XII (February, 1979), 2.

Rafferty, Max. "Decline and Fall of Education--Part II." The Knoxville [Tennessee] Journal, May 23, 1978.

Rogers, Vincent R., and Joan Baron. "Declining Scores: A Humanistic Explanation." Phi Delta Kappan, LXIII (December, 1976), 311-313.

Summers, Anita A., and Barbara L. Wolfe. "Do Schools Make a Difference?" The American Economic Review, LXVII (September, 1977), 639-652.

"The Results Are In--The Controversy Continues." NJEA Review, L (May, 1977), 14-16.

C. Other Works

Capsule Report: Tennessee Looks at Its Schools, 1977-78 State Education Assessment of Schools. Knoxville, Tennessee: Tennessee State Testing and Evaluation Center, 1979.

Kelley, Truman L., Giles M. Ruch, and Lewis M. Terman. New Stanford Achievement Test: Directions for Administering. New York: World Book Company, 1929.

_____. Stanford Achievement Test: Directions for Administering. New York: World Book Company, 1940.

Popham, W. James, and others. Of Measurement and Mistakes. Testimony before the General Subcommittee on Education, Committee on Education and Labor, U.S. House of Representatives, Washington, D.C., March 29, 1973.

Sligo, Joseph R. "Comparisons of Achievement in Selected High School Subjects in 1934 and 1954." PhD dissertation, University of Iowa, 1955.

APPENDIXES

APPENDIX A

THE TREATMENT GROUPS--TESTS TAKEN AND THE COVARIANTS

Table 40

Treatment Groups and the Covariants

Group	Number of Students	Stanford Achievement Test Taken	Mean 7th Grade Battery Score (Covariant)	Sex Male/ Female (Covariant)	Average Age
1	49	1929	261.5100	28/21	164.7347
2	48	1940	263.0625	21/27	162.6042
3	45	1952	254.0667	25/20	164.6667
4	46	1964	254.9565	19/27	165.7174
5	48	1973	268.0208	29/19	163.6667

APPENDIX B

TABLE OF SCORES ON ALL SUBTESTS, THE TOTAL BATTERY AVERAGE AGE, MEAN 1977-78 SCORE, AND RATIO BOYS TO GIRLS

Table 41

Means Scores for the Groups on Subtests and the Total Battery, Average Age for Each Group, Mean Score on the 1977-78 Test, Ratio Boys to Girls

		Word Mean. Vocab.	Para. Mean. Read. Comp.	Total Read Avg. Read	Language Usage	Math Concepts	Math Applications	Math Comp.	Total Math	Spelling	Social Science	Science	Total Battery	Literature	Age in Months	77-78 Score	Boys/Girls
$n_1 = 49$	Group 1 1929 Test	65.92	65.51	65.51	65.51	X	X	X	44.82	30.04	57.43	58.22	55.63	51.10	164.7	261.5	28/21
$n_2 = 48$	Group 2 1940 Test	49.04	42.62	X	61.52	19.29	X	28.90	X	51.62	55.69	71.62	50.10	58.62	162.6	263.1	21/27
$n_3 = 45$	Group 3 1952 Test	60.69	65.36	62.96	68.58	53.22	X	49.44	51.02	63.87	53.82	60.22	60.13	60.62	164.7	254.1	25/20
$n_4 = 46$	Group 4 1964 Test	X	48.37	X	57.35	43.54	34.83	37.54	X	42.57	48.33	50.76	48.20	X	165.7	254.9	19/27
$n_5 = 48$	Group 5 1973 Test	57.58	59.04	58.46	62.08	55.44	62.19	61.50	60.15	59.29	59.44	59.21	59.65	X	163.7	268.0	29/19

n = 236

APPENDIX C

GRAPH OF GRADE EQUIVALENTS BY EACH GROUP ON EACH SUBTEST AND ON THE TOTAL BATTERY

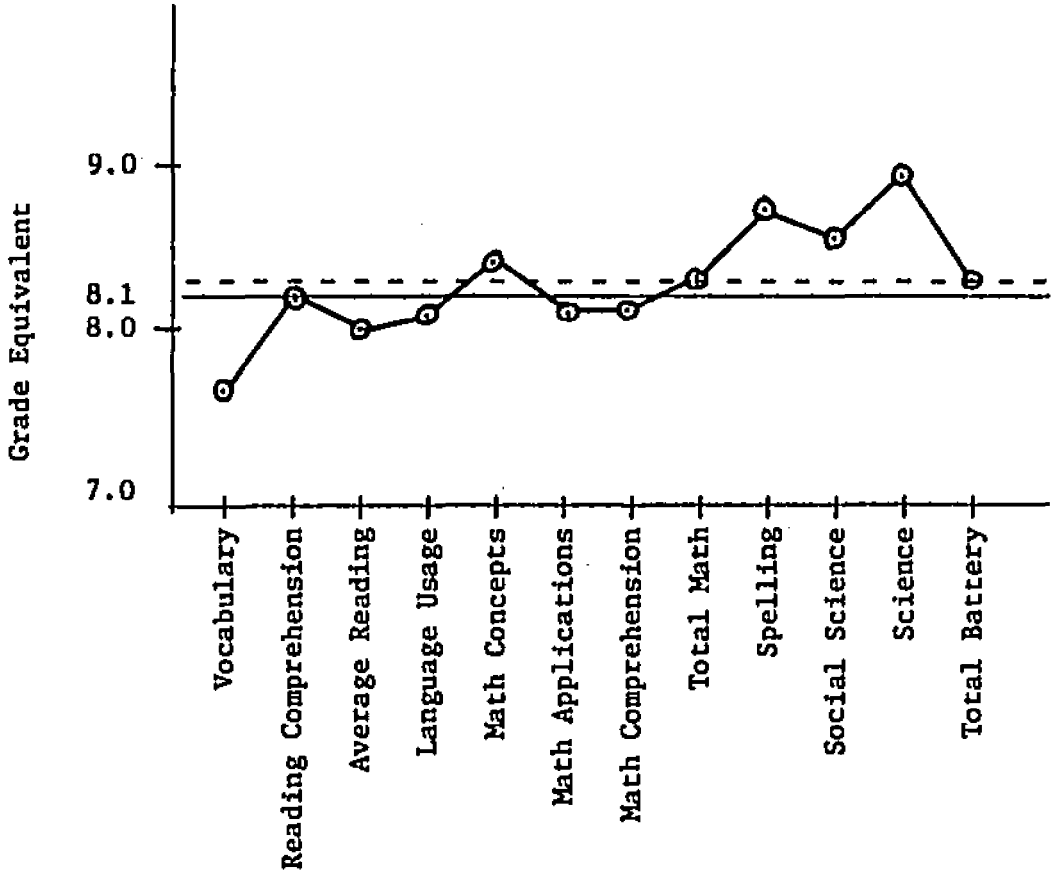

Figure 7

Graph of Grade Equivalents by Each Group on
Each Subtest and on the Total Battery

VITA

VAUGHN D. CHAMBERS

Personal Data:

Date of Birth: October 23, 1937
Place of Birth: Scott County, Tennessee
Marital Status: Married

Education:

Public Schools, Blount County, Tennessee.
University of Tennessee, Knoxville, Tennessee.
East Tennessee State University, Johnson City, Tennessee; mathematics, history, B.S.,
East Tennessee State University, Johnson City, Tennessee; guidance and counseling, M.A.,
East Tennessee State University, Johnson City, Tennessee; supervision and administration, Ed.D., 1979.

Professional Experience:

Workship Teacher, East Tennessee State University Mathematics Department, 1963-64.
Teacher, Kingsport City Schools; Kingsport, Tennessee, 1964-66.
Guidance Counselor, John Sevier Junior High School; Kingsport, Tennessee, 1966-67.
Principal, John Sevier Junior High School and John Sevier Middle School; Kingsport, Tennessee, 1967-79.