

A CORRELATION OF THE
STANFORD ACHIEVEMENT TEST LEVEL TASK II
AND THE
COMPREHENSIVE TEST OF BASIC SKILLS LEVEL 19/20

EUGENIA LANE PARKER

A CORRELATION OF THE
STANFORD ACHIEVEMENT TEST LEVEL TASK II
AND THE
COMPREHENSIVE TEST OF BASIC SKILLS LEVEL 19/20

An Abstract
Presented to the
Graduate and Research Council of
Austin Peay State University

In Partial Fulfillment
of the Requirements for the Degree
Master of Arts in School Psychology

by
Eugenia Lane Parker

July 12, 1990

ABSTRACT

This study was made to determine whether or not there would be any score variance when two different achievement tests were administered to the same sample of students. The sample consisted of 173 Sophomores from Hunters Lane High School in Nashville, Tennessee. The sample was administered the Stanford Achievement Test in their freshman year and the Comprehensive Test of Basic Skills in their sophomore year.

The scores of the two tests were correlated and an analysis of the results indicated that the students performed similarly on both achievement tests with correlations of .70 and better, with the exception of the equivalent Spelling subtests. The sample means were above the expected 50 of the norm group on both tests. The standard deviations ranged from 14 to 21 on the equivalent subtests and were acceptable with the exception of the Stanford Spelling subtest which had a standard deviation of 36.

In conclusion, there was little score variance between the two tests; therefore, they seem to measure similar abilities.

A CORRELATION OF THE
STANFORD ACHIEVEMENT TEST LEVEL TASK II
AND THE
COMPREHENSIVE TEST OF BASIC SKILL LEVEL 19/20

A Thesis
Presented to the
Graduate and Research Council of
Austin Peay State University


In Partial Fulfillment
of the Requirements for the Degree
Master of Arts in School Psychology

by
Eugenia Lane Parker


July 12, 1990


To the Graduate and Research Council:

I am submitting herewith a Thesis written by Eugenia Lane Parker entitled "A Correlation of the Stanford Achievement Test Level Task II and the Comprehensive Test of Basic Skills Level 19/20." I have examined the final copy of this paper for form and content, and I recommend that it be accepted in partial fulfillment of the requirements for the degree Master of Arts in Psychology, with a major in School Psychology.

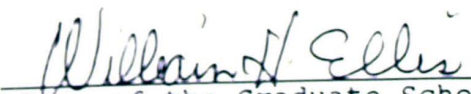

Major Professor

We have read this thesis
and recommend its
acceptance:


Minor Professor
or
Second Committee Member


Third Committee Member

Accepted for the Graduate and
Research Council:


Dean of the Graduate School

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to Dr. Susan Kupisch, Professor of Psychology, Austin Peay State University, for her aid, guidance and time given during the entire study.

Appreciation is extended to Mr. A.D. Hancock, Executive Principal of Hunters Lane High School, and his staff; Dr. Ed Binkley, Director of Research and Evaluation for Metro Nashville Schools, and his staff; and to the State Department of Education, Department of Research and Evaluation, for their valuable assistance in making this study possible.

Additionally, the author wishes to especially thank her mother, Dorothy C. Parker, family and friends for their help, encouragement, and understanding during this study.

TABLE OF CONTENTS

CHAPTER	PAGE
1. INTRODUCTION	1
2. REVIEW OF LITERATURE	2
The History of Educational Reform	2
Review of the Stanford Achievement Test.	6
Review of the CTBS/4	9
Item Response Theory	13
3. METHODOLOGY.	16
Subjects	16
Instruments.	16
Procedures	17
Analysis	18
4. RESULTS.	19
5. DISCUSSION AND SUMMARY	23
BIBLIOGRAPHY.	27

LIST OF TABLES

TABLE	PAGE
1. Stanford and CTBS/4 Equivalent Subtests.17
2. Correlation Coefficients of Equivalent Subtests.19
3. Equivalent Subtests and their Means.20
4. Equivalent Subtests and Standard Deviations. .	.21

CHAPTER 1

INTRODUCTION

A spotlight has been placed on the nation's schools. One cannot listen to a news report or read a newspaper without the state of our schools being mentioned. This is a time of educational reform. Schools have changed their programs so that basic proficiency skills are in the forefront because schools are being held accountable for the achievement level of their students and the success of their programs.

Due to the need to be accountable to the public and demonstrate pupil achievement, competency testing programs have been implemented. Metropolitan Nashville Public Schools are no exception. In the 1986-87 school year the Metropolitan Nashville Public School System began administering the 1982 Edition of the Stanford Achievement Test. In 1990 the school system is adopting the 1989 Edition of the Comprehensive Test of Basic Skills.

In the Fall of 1988, all ninth grade students at Hunters Lane High School were administered the Stanford Achievement Test (Level TASK II). In the Spring of 1990, these same students will be evaluated by the Comprehensive Test of Basic Skills Survey Test (Level 19/20). My hypothesis is that there will be score variance due to the fact that the scores are from two differently designed tests.

CHAPTER 2

REVIEW OF LITERATURE

Educational reform has been in existence since World War II. Colleges and industries have complained since the 50's that students are graduating without the abilities needed to go into higher education or fill positions in industry. As a result, the National Defense Education Act of the 1950's was mandated requiring schools to look at the outcomes of our educational system (Brandt, 1989).

In the 1970's, accountability came into effect. The feeling was that much money was being put into the schools, yet the SAT scores were going down. There was a demand to know what the country was getting for its money (Brandt, 1989; Brown, 1988).

Educational reforms have been in the form of improving the quality of teachers preparation through career ladderling, incentive pay programs, and certification tests. To improve instruction to students, we have extended the school day, added required courses for graduation, instated no pass - no play policy in extracurricular activity programs, developed enrichment and remedial programs, and required a passing score on a test of minimum competency skills to graduate (Airasian, 1987). The main goal of all these reforms is to improve the competence of people coming out of our educational system.

3

The use of testing systems has become the most visible and critical aspect of state government efforts to improve educational standards and gain increased control over the process of education in local school districts (Airasian, 1987). One criticism of the testing movement, according to Jane L. David (1988), is that a shift in test scores of a given school could be attributed to quality of instruction as well as a change in student population, curriculum, available resources, school leadership, or even the physical facilities in a school. Yet it is these test scores that we look at to judge the quality of our schools.

Another criticism of the testing movement is that the government is looking at the outcome of the schools as shown by test results, rather than the input into education. Test results change the behavior of administrators and instructors in order to improve the test scores. Schools, as reflected in their curriculums, tend to narrow their focus to instruction on isolated skills that can be easily measured by multiple choice items on standardized tests (David, 1988; Brandt, 1989).

In 1984, there were 29 states that required pupils to take competency tests at selected points in the educational ladder (Airasian, 1987). The 1989 state legislative sessions opened with approximately 30 states having accountability issues scheduled for discussion

(Pipho, 1989). In 1990, the number of states requiring testing has increased as more laws mandating accountability of our local schools have been passed.

As a result of the state and federal legislation to improve the plight of our schools, there has been a great number of task forces and research programs developed to study the quality of testing programs for use in the schools. One of these research programs is a five-year study at the Center for Research on Evaluation, Standards, and Student Testing, sponsored by the U.S. Office of Educational Research and Improvement. Eva L. Baker (1988), who is a member of the research team, stated that many policy-makers regard standardized tests as credible, objective, and the bottom line for the assessment of the schools. Another program, The National Assessment of Educational Progress (NAEP), has been periodically monitoring U.S. students in reading comprehension, writing, and mathematics. The federal government is considering changing the NAEP measurement practices to make state-by-state comparisons. Also, the National Association of State Boards of Education is developing a set of recommendations which would strengthen state capacity for education policy making (Cohen, 1988).

State leadership in educational policy has increased over the years as state support for local schools has increased. States pay about 50% of

education costs, and high-quality education is now viewed as a key to economic development (Cohen, 1988). The state of Tennessee, like other states, is interested in improving its economy and level of educational preparedness. Tennessee has expanded its interest in economic development, initiating programs to attract large businesses into the state. Local media has indicated businesses are reluctant to locate in Tennessee due to its standing in the national comparison of school systems. As a result, the state has become more concerned with the achievement standing of our schools. This concern has brought about the development of a state-wide testing program that will be used to evaluate school achievement.

This paper focuses on the impact of the testing movement in its narrowest sense by looking at only one local school district and one high school. The Metropolitan Nashville Public School System has been administering the 1982 Edition of the Stanford Achievement Test Series to its pupils at selected points in the educational ladder since 1985. In the Spring of 1990, the Metropolitan Nashville Public School System changed from the Stanford Achievement Test Series to a state-wide testing program called the Tennessee Comprehensive Assessment Program (TCAP). The TCAP uses the 1989 Edition of the Comprehensive Test of Basic Skills Survey Test (CTBS/4). Additional subtests have

been added which measure specific skill objectives found in the Tennessee State Curriculum Guides. The CTBS/4 will be used to obtain norm-referenced information about the achievement levels of students. The additional subtests will be used to assess mastery of skills covered in the curriculum. It is important to review the characteristics of these tests.

Stanford Achievement Test Series

The Stanford Achievement Test Series, 7th Edition, contains ten battery levels which range from K.0 through 13.0. Test content was developed by the review of textbooks, state guidelines, and syllabi. The instructional objectives were developed from this pool. Test items were then developed to measure each of the objectives. The test was normed in 1982 providing empirical norms from the Fall and Spring semesters (Cannon et. al., 1985). Test authors used the 1970 School District Census from the U.S. Office of Education and statistically weighted test results to achieve a better approximation of demographic representation. Also data were collected concerning median family income and median years of schooling of adults over 25 years of age. The standardization sample included 250,000 students from 300 school districts for the Fall, and 200,000 students from 300 districts for the Spring sample (Gardner, et. al., 1985).

The Stanford research design called for traditional

methods of data analysis and the application of Rasch Model Techniques. A pool of tryout items was developed and administered to the standardization samples. Each item was judged according to the following criterion:

- a) How well does the question measure the particular objective for which it was written?
- b) How many in the tryout group answer correctly?
- c) How does the question distinguish between those who score high or low on the test?
- d) Do more students answer the question correctly at successively higher levels?
- e) How many students selected each option?

Students were administered adjacent levels of the test in order to place scores for all ten levels of the series on the same scale. Rasch item difficulties were calculated for each item in a domain, such as Mathematics, and a mean Rasch Item Difficulty was computed for each level in that domain. Then, the difference between the mean item difficulties of the levels was calculated and added in as an equating constant to convert the item difficulties of one test level to the scale of the next test level. The appropriate equating constant was then added to the Fall

standardization item difficulties of each test level to produce an equated Rasch ability scale (Gardner, et. al., 1985).

Validity data for the Stanford Achievement Test, 7th Edition, were in the form of a referral to the Stanford Index of Instructional Objectives and a suggestion that the user evaluate the validity of the test through careful examination of the test content. Reviews of the Stanford indicate it contains a comprehensive range of content suitable for a large number of schools. Item difficulties are provided for each item in the index. The technical manual also provides intercorrelations between the Stanford subtests and the Otis-Lennon School Ability Test. The intercorrelations were from a sample of 4,147 students at the beginning of grade 10. The correlation coefficients ranged from .70 to .80 (Gardner, 1985).

The Stanford Test of Academic Skills Level II (TASK II) is the 9.0-13.0 battery level of the Stanford Achievement Test Series. The Stanford TASK II was the Stanford Achievement Test that was administered to Hunters Lane High School students. As this study will only review data from Hunters Lane High School, the review will deal mainly with the Stanford Achievement Test (Level TASK II). This level of the Stanford Achievement Test was designed to evaluate those skills that are requisite to continued academic training. The

test is in a multiple choice format. The items are written so that the students perform at all levels of Bloom's Taxonomy. The test items are judged to be no better or worse than other achievement tests (Ory, 1985).

A reliability study was performed during the equating of forms program used to develop the final form of the test. Data were provided for internal consistency reliability, alternate-forms reliability, and measurement to show consistency over time. The internal consistency reliabilities were computed using the Kuder-Richardson Formula #20 reliability coefficients. They ranged from .88 to .96 for Form E and .90 to .96 for Form F in the Fall and from .86 to .96 for Form E and .89 to .96 for Form F in the Spring. Alternate-Forms reliability coefficients ranged from .84 to .92. Also, correlation coefficients were computed from the performance of students tested in both the Fall and Spring of the same school year. The Fall-Spring correlation coefficients ranged from .69 to .86 (Gardner, 1985). The test was reviewed by John C. Ory (1985) as having satisfactory reliabilities across subtests with reasonable standard errors of measurement.

Comprehensive Test of Basic Skills

The Fourth Edition of the Comprehensive Test of Basic Skills (CTBS/4) was published in 1989. There are eleven levels ranging from grades K.0 through 12.9 and

two forms of each level. Two new forms, the Survey form and the Benchmark form, are available for those who wish a quick survey but have no need for curriculum referenced information.

The test objectives were developed using textbooks and curriculum guides from state departments of education. The pool of items was reviewed by panels representing various ethnic groups and teachers. Comparisons were made of test items with other recently published CTB/McGraw-Hill tests (Linn, 1985). The items at each test level were categorized by subobjective (word attack) and by cognitive process, whether the item requires recall, understanding, inferential reasoning, or evaluation. There were new item types such as vocabulary and spelling being evaluated in context (Shepard, 1985). The test items reflected the need for higher order thinking skills such as: critical thinking, comprehension of the whole passage, and the ability to find, interpret, organize, analyze, and apply information for their own purposes.

The test provided norms for the Fall and Spring. The standardization samples consisted of students from the Northeast, Midwest, Southeast, and West regions that were stratified by region, community type (urban, suburban, rural), and size. The Fall sample consisted of 167,000 students in grades K-12 and 156,000 students from grades K-12 for the Spring sample. The samples

were composed of public school districts, Catholic Dioceses, and private non-catholic schools.

The CTBS/4 was constructed with the use of Item Response Theory used for item analysis, item bias studies, scaling, and estimation of standard errors of measurement. CTBS/4 used the Three-Parameter Logistic Model of Item-Response Theory. This model of IRT checks each item for item difficulty, item discrimination, and the probability of a correct response for a very low-scoring student. Students were administered adjacent levels of the test to form a continuous scale for each subtest. Each level of a subtest has a nominal range or a lowest obtainable scale score and a highest obtainable scale score. Each level has a range from a scale score at the 5th percentile of the Fall of the lowest target grade for that level to the 95th percentile of the Spring of the highest target grade. For example, Level 17/18 covers grades 6.6 to 9.2 and a score at the 5th percentile would be equivalent to the Fall of Grade 7. A score at the 95th percentile on Level 17/18 would be equivalent to the Spring of Grade 8. IRT scoring or number correct scoring can be used (Technical Bulletin, 1989).

Validity of the test was discussed in terms of content validity. The test authors refer the user to the Test Coordinators Handbook and the Class Management Guide for descriptive information about the test

content. Since the test was just recently published, no reviews were available to collect additional information concerning validity of the test. A review of the CTBS/3, Forms U and V, by Robert Linn (1985) indicated "scanty" evidence supporting the validity of that edition. Like the CTBS/3, the CTBS/4 technical manual lists pages of item location parameters, item difficulties, intercorrelations with the Test of Cognitive Skills, and proportion correct scores. Content validity was considered primary and the decision as to whether or not the test was valid for the intended population was left as a matter of judgement for the user.

Reliability, or the consistency of test results, was described using several kinds of data. The test was administered in the Spring and Fall of 1988 to the standardization sample. Students were given interlevel linking tests which, for example, contained half Level 11 items and half Level 12 items to form the Level 11/12 test. Hunters Lane High School students were administered the Level 19/20 Survey Test of the CTBS/4. Internal consistency reliability coefficients were computed for each subtest from one administration by using the Kuder-Richardson Formula #20 (KR20). The KR20 coefficients for Level 19/20 Survey, Form A, for grade 10 ranged from .73 to .87 in the Fall and from .74 to .94 in the Spring. The KR20 coefficients for Level

19 20 Survey, Form B, grade 10 ranged from .67 to .94 in the Fall and from .68 to .94 in the Spring. The test authors also refer to the use of the Standard Error of Measurement (SEM) as another aspect of reliability of the test scores. They mentioned the fact that measurement error is associated with every test score and that the SEM can be used to obtain a range within which a students true score is likely to fall. A Standard Error Curve was presented for each subtest in which a curve for each level is plotted. This was done in an effort to help the user identify the range within each level and within each test that provides the most accurate measurement.

Item Response Theory

The research design of both tests was based on Item Response Theory (IRT), pioneered by Frederick M. Lord. Item Response Theory became a dominant topic of study in the 1970's and is based on latent trait theory. According to Hambleton and Swaminathan (1985), examinee performance on a test can be predicted in terms of one or more characteristics referred to as traits. The traits must be estimated from observable examinee performance on a set of test items. Item Response Models are mathematical models which are based on specific assumptions about the test data. Tests built upon IRT provide an index which tells the precision with which each examinee's ability is estimated. This index

can vary from examinee to examinee, hence, the test is sample-free. Unlike standard testing methods, IRT produces tests that are sample-free and item-free. The scores are not a function of the items used to construct the test or the samples used to calibrate the tests.

The Stanford Achievement Test is designed using the Rasch Model or One-Parameter Logistic Model. The Rasch Model assumes that all items have equal discriminating power and that guessing is minimal. Traub (1983), disagrees with the Rasch Model on the context that common sense and the history of testing show that guessing plays a part in the process of multiple choice items. However, the fact that the Rasch Model has fewer parameters lends it for easier application. There has been much research concerning the Rasch Model and it has been found to have fewer problems with parameter estimations than more general models (Hashaway, 1978).

The Comprehensive Test of Basic Skills was designed using the Three-Parameter Logistic Model. This design incorporated item discrimination and guessing as two significant factors in the development of a test. This model acknowledges that some test items may be more discriminating than others in predicting score consistency on items. The third parameter is called the pseudo chance level parameter. It accounts for item response data from low-ability examinees or the probability that a low-ability examinee answers an item

correctly. This model is said to have better "fit" between the model and the data which would lead to stronger results.

This research project will compare student scores on the Stanford Achievement Test and the Comprehensive Test of Basic Skills. Although both tests assess a broad range of achievement, these two tests are based on different models of test design. Therefore, the tests need to be compared to see which instrument better fits the population it was used to test.

CHAPTER 3

METHODOLOGY

Subjects

In the Fall of 1988, 494 Freshman were administered the Stanford Achievement Test (Level Task II). In the Spring of their Sophomore year, 361 students were administered the CTBS/4 (Level 19/20). Of these students, only 173 students had a score for every subtest on both the Stanford Achievement Test and the CTBS/4. Therefore, the sample of this research project consists of a population of 173 students.

Each student was assigned a case number, the names were not used. There was no contact with any of the students. The scores were obtained from the school data summary for the Stanford Achievement Test and from the individual student reports for the CTBS/4. (The individual student reports were not distributed to the students until the Fall of 1990.)

Instruments

The Stanford Achievement Test subtests and the CTBS/4 subtests were analyzed to find equivalent subtests to be correlated. Because the CTBS/4 Language Mechanics, Language Expression, Mathematics Computation, and Math Concepts/Applications subtests had no equivalent subtests on the Stanford Achievement Test, they were not used in this study. The following table, (Table 1), shows the Stanford Achievement Test and

CTBS/4 subtests that were equivalent and used in the study.

Table 1. Stanford and CTBS/4 Equivalent Subtests

Stanford Subtests	CTBS/4 Subtests
Reading Comprehension	Reading Comprehension
Reading Vocabulary	Reading Vocabulary
Total Reading	Total Reading
	Language Mechanics
	Language Expression
English	Total Language
	Math Comprehension
	M-Concepts/Application
Mathematics	Total Mathematics
Spelling	Spelling
Using Information	Study Skills
Science	Science
Social Studies	Social Studies

Procedures

The Stanford Achievement Test school data summary report and the CTBS/4 individual student reports contained scores for each student by subtest. Normal Curve Equivalents (NCE's), obtained from the tests manuals, were reported for each subtest raw score.

The equivalent subtests' Normal Curve Equivalents were correlated by a statistical computer program using

the Pearson Product-Moment Correlation Formula.

Analysis

Normal Curve Equivalents, which have a mean of 50 and a standard deviation of 16, were used for score comparison on both tests. The data were analyzed by comparing subtest correlations, means, and standard deviations of the sample group. Analysis of the correlation coefficients provides information concerning whether or not the students scored similarly on both tests. Analysis of the mean score of each subtest provides information as to how the sample performed as a whole on the subtest, which could indicate the sample difficulty of each subtest compared to the norm sample. Analysis of the standard deviation of each subtest provides information concerning the variability of the sample groups scores on each subtest.

CHAPTER 4

RESULTS

Table 2 lists the results of the correlation of equivalent subtests on the Stanford Achievement Test and the CTBS/4. The Total Reading subtest scores had the highest correlation with a .828, followed by the Total Mathematics' subtests with a correlation of .799. The rest of the subtests correlated moderately well with a range of .705 to .771 with the exception of one, Spelling, which had the lowest correlation of .681.

Table 2. Correlation Coefficients of Equivalent Subtests.

Stanford Subtest	CTBS/4 Subtest	Correlation
Reading Comprehension	Reading Comprehension	0.711
Reading Vocabulary	Reading Vocabulary	0.771
Total Reading	Total Reading	0.828
English	Total Language	0.722
Mathematics	Total Mathematics	0.799
Spelling	Spelling	0.681
Using Information	Study Skills	0.739
Science	Science	0.736
Social Studies	Social Studies	0.705

The correlations showed that for the most part students scored similarly on both tests. A student that scored high, average, or low on the Stanford Achievement

Test tended to score similarly on the CTBS/4. The students scored most similarly on the Total Reading subtest and the Spelling subtest obtained the most dissimilar scores. Since both tests were designed using school curriculum guides as guidelines for the development of their items, they tended to tap or evaluate similar abilities.

Table 3 lists the equivalent subtests and their means based on the sample group scores.

Table 3. Equivalent Subtests and Means

Stanford Subtest	Mean	CTBS/4 Subtest	Mean
Reading Comp.	51.803	Reading Comp.	55.127
Reading Voc.	46.335	Reading Voc.	51.306
Total Reading	48.832	Total Reading	53.428
English	52.775	Total Language	52.393
Mathematics	54.717	Total Mathematics	54.272
Spelling	52.272	Spelling	52.647
Using Information	54.775	Study Skills	54.491
Science	50.214	Science	49.468
Social Studies	52.312	Social Studies	51.734

The Stanford Achievement Test means ranged from 46.335 on the Reading Vocabulary subtest to 54.775 on the Using Information subtest. An average of the means listed on the Stanford equals an overall test mean of 51.559.

The CTBS/4 means ranged from 49.468 on the Science subtest to 55.127 on the Reading subtest. An average of the means listed on the CTBS/4 equals an overall test mean of 52.763.

The Hunters Lane High School sample performed, as a whole, "slightly" higher than the norm group on both tests with overall test means over the expected norm group mean of 50. The students overall test mean was higher on the CTBS/4 than the Stanford.

Table 4 lists the equivalent subtests and the standard deviations obtained for each subtest based on the sample group scores.

Table 4. Equivalent Subtests and Standard Deviations.

Stanford Subtest	S.D.	CTBS/4 Subtest	S.D.
Reading Comp.	15.133	Reading Comp.	17.299
Reading Voc.	16.377	Reading Voc.	18.934
Total Reading	15.126	Total Reading	17.363
English	14.865	Total Language	18.142
Mathematics	15.672	Total Mathematics	18.863
Spelling	14.071	Spelling	19.709
Using Information	16.751	Study Skills	21.047
Science	15.896	Science	19.832
Social Studies	15.788	Social Studies	18.240

The Stanford Achievement Test standard deviations range from 14.071 on the Spelling subtest to 16.751 on

the Using Information subtest. An average of the standard deviations listed on the Stanford equals an overall test standard deviation of 15.520.

The CTBS/4 standard deviations range from 17.299 on the Reading Comprehension subtest to 21.047 on the Study Skills subtest. An average of the standard deviations listed on the CTBS/4 equals an overall test standard deviation of 18.825.

The Stanford and CTBS/4 standard deviations, for the most part, were in acceptable ranges in comparison with the expected standard deviation of 16 as in the norm group. However, the CTBS/4 Study Skills subtest with a standard deviation of 21.047 may have questionable discrimination capabilities. This standard deviation indicates a great deal of intrasubtest variability.

CHAPTER 5

DISCUSSION AND SUMMARY

The hypothesis that there would be a great deal of score variance due to the two different test designs proved to be incorrect. The equivalent subtests correlated moderately well with the Spelling subtest obtaining the lowest correlation.

The lower correlation between the Spelling subtests could be due to the fact that the CTBS/4 used a new format on the Spelling subtest. Instead of choosing the correctly spelled word out of a series of spellings of the same word, the CTBS/4 tested spelling words in context. Although Hunters Lane High School Students performed better than the norm group on both subtests, the standard deviation of the CTBS/4 Spelling subtest indicated it had more intrasubtest variability. The low correlation indicates that the two subtests may be tapping different abilities. However, this study did not research the individual abilities measured by each subtest; therefore, this topic may need to be followed up at another time.

Comparison of the individual subtest means of the CTBS/4 indicated that the students scored above the norm group on all the CTBS/4 subtests with the exception of one, the Science subtest. The Science subtest mean of 49.468, which is below the norm group mean of 50, indicated that the sample had difficulty with this

subtest.

Comparison of the individual subtest means of the Stanford indicated that the students scored above the norm group on all the Stanford subtests with the exception of one, the Reading Vocabulary subtest. The Reading Vocabulary subtest mean of 46.335, which was below the norm group mean of 50, indicated that the sample had difficulty with this subtest. The Total Reading Score was below the norm group mean as well because of the low mean on the Reading Vocabulary subtest.

Comparison of the sample's performance on the equivalent subtests of the Stanford and CTBS/4 indicated that the sample scored higher in the areas of Reading and Spelling on the CTBS/4 than the Stanford. In contrast, the sample scored higher in the areas of English/Language, Math, Using Information, Science, and Social Studies on the Stanford than the CTBS/4.

Hunters Lane High School students, as a whole, scored better than the norm group on both tests. Comparison of their overall test means indicated that the students scored higher on the CTBS/4 than the Stanford based on their averages.

The CTBS/4 subtests had higher standard deviations than the Stanford subtests, which means the CTBS/4 measured wider individual differences between the students scores than the Stanford. The CTBS/4 Study

skills subtest scores showed the most variability with a standard deviation of 21.047, while the Stanford Spelling subtest scores showed the least variability with a standard deviation of 14.071. The standard deviations were in acceptable ranges when compared with the norm group's standard deviation of 16.

Whether one test was better than the other could not be proven by the data gathered in this study. The data show that the students did not score significantly better or worse on either test. The decision as to which test would be the better fit with the students requires further study. A follow up study needs to be done that correlates the objectives covered in the curriculum with the objectives covered by the tests.

As Metropolitan Nashville Public Schools have already adopted the CTBS/4, some advantages of the new test should be mentioned. The CTBS/4 has more recent norms and the short Survey form enables school systems to gather normative data and add criterion referenced items designed to assess the individual school systems needs.

As in any form of assessment, individual or school-wide, important decisions about an individual student or program should never be based on only one assessment tool. There are many variables that can affect test scores that have nothing to do with the test. All of the variables need to be considered when making any

decisions. Accountability must work both ways, the schools need to assess to evaluate progress, but the ones who look at the test scores to make decisions should know how to interpret the scores and be accountable for the reforms that come from their decisions.

BIBLIOGRAPHY

- Airasian, P.W. (1987). The consequences of high school graduation testing. *NASSP Bulletin*, 71, 54-68.
- Baker, Eva L. (1988). Can we fairly measure the quality of education? *National Education Association*, January 1988, pp. 9-14.
- Brandt, R.L. (1989). On misuse of testing: a conversation with George Madaus. *Educational Leadership*, 46, 26-29.
- Brown, Rexford (1989). Literacy and Accountability. *The Journal of State Government*, pp. 68-72.
- Cannon, Tom et. al. (1989). 1989-90 catalog of tests and services. *Tennessee department of education office*, Nashville, TN, (pp. 31-40).
- Cohen, Michael il. (1988). Designing State Assessment Systems. *Phi Delta Kappan*, 69, 583(6).
- CTBS/4 Technical Bulletin, (1989). *CTB Macmillan/McGraw Hill School Publishing Company*, Monterey, CA.
- David, J.L. (1988). The use of indicators by school districts; aid or threat to improvement? *Phi Delta Kappan*, 69, 499-504.
- Gardner, Eric F. et. al. (1985). Stanford Achievement Test Technical Data Report. *The Psychological Corporation*, Harcourt Brace Jovanovich, Inc.
- Hambleton, Ronald K. and Swaminathan, Hariharan (1985). *Item Response Theory*. *Kluwer-Nijhoff Publishing*, Hingham, MA.

- Hashway, Robert M. (1978). Objective Mental Measurement. Praeger Publishers, Praeger Special Studies, New York, N.Y..
- Linn, R.L. (1985). Review of the comprehensive test of basic skills. In J.V. Mitchell (Ed.), The ninth mental measurement yearbook: Vol. 2, (pp. 381-385). University of Nebraska-Lincoln: The Buros Institute of Mental Measurements.
- Macmillan/McGraw-Hill (1990). CTB Catalogue. CTB Macmillan/McGraw-Hill School Publishing Company, (pp. 10-20).
- Ory, J.C. (1985). Review of stanford test of academic skills. In J.V. Mitchell (Ed.), The ninth mental measurement yearbook: Vol. 2, (pp. 1468-1469). University of Nebraska-Lincoln: The Buros Institute of Mental Measurements.
- Pipho, Chris (1989). Accountability Comes Around Again. Phi Delta Kappan, 70, 662(2).
- Shepard, L.A. (1985). Review of comprehensive test of basic skills. In J.V. Mitchell (Ed.), The ninth mental measurement yearbook: Vol. 2, (pp. 386-389), University of Nebraska-Lincoln: The Buros Institute of Mental Measurements.
- Traub, R.E. (1983). In R.K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.