An Investigation of the Test/Retest Stability of the Stanford Achievement Test

Judy A. Misiurewicz

Follow this and additional works at: https://digitalcommons.brockport.edu/ehd_theses
Part of the Elementary Education Commons, and the Language and Literacy Education Commons

To learn more about our programs visit: http://www.brockport.edu/ehd/

Repository Citation
https://digitalcommons.brockport.edu/ehd_theses/1113

This Thesis is brought to you for free and open access by the Education and Human Development at Digital Commons @Brockport. It has been accepted for inclusion in Education and Human Development Master’s Theses by an authorized administrator of Digital Commons @Brockport. For more information, please contact kmyers@brockport.edu.
AN INVESTIGATION OF THE TEST/RETEST STABILITY
OF THE STANFORD ACHIEVEMENT TEST

THESIS

Submitted to the Graduate Committee of
Department of Education & Human Development
State University of New York
in Partial Fulfillment of the
Requirements for the Degree of
Master of Science in Education

by
Judy A. Misiurewicz
State University of New York
College at Brockport
Brockport, New York
July, 1968
SUBMITTED BY:

Judy A. Missinrewicz

APPROVED BY:

Peter Ann Balcer
Thesis Advisor
6/30/88

Robert B. Miller
Second Faculty Reader
6/29/88

Marvin Beers
Graduate Director
8/10/88
Abstract

This study investigated the stability of three subtests of the Stanford Achievement Test. It further analyzed reliability and validity assessment of standardized tests in general, with particular emphasis on their measurement of reading ability. It clarified the position of stability assessment within a total evaluation framework. The subjects of the study were 39 fourth grade students from an urban upstate New York Catholic school district. The Reading Comprehension, Word Study Skills and Vocabulary subtests of the Stanford Achievement Test, Intermediate Level I, Form E were administered to the subjects by their regular classroom teachers during the third week of May. Nine days later the same tests (same form) were readministered under similar conditions. The raw scores for each test were then paired and correlated for analysis utilizing a Pearson Product Moment procedure. Results of the study indicated that all three subtests were stable in their measurement of student performance over time. Positive, significant relationships existed between test/retest scores for each test. Student completion rate was found to have a major effect on the correlation index of the Reading Comprehension subtest.
Table of Contents

List of Tables

Chapter

I. Statement of the Problem
   Purpose
   Need for the Study
   Questions Addressed by the Study
   Limitations
   Summary

II. Review of the Literature
   Purpose
   Standardized Measurement
   Reading Measurement
   Reliability Assessment
   Validity Assessment
   Summary

III. Design of the Study
   Purpose
   Questions Guiding the Study
   Methodology
   Summary
<table>
<thead>
<tr>
<th>Tables</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading Comprehension Subtest</td>
<td>46</td>
</tr>
<tr>
<td>Word Study Skills Subtest</td>
<td>49</td>
</tr>
<tr>
<td>Vocabulary Subtest</td>
<td>52</td>
</tr>
</tbody>
</table>
Chapter 1

Statement of the Problem

Purpose

This study was initiated to investigate the stability of the Reading Comprehension, Word Study Skills and Vocabulary subtests of the Stanford Achievement Test, Intermediate Level I, Form E under local, actual use conditions.

Need for the Study

The valid use of any educational measurement cannot be established unless and until the measuring instrument can be shown to be reliable, that is, consistent in repeated measurements of the same concept or construct that it purports to measure. Various statistical methods may be employed to arrive at an empirical estimation of the degree of reliability of a test. The method chosen relates to the aspect of reliability under consideration. It is important for evaluators to determine a reliability estimation which yields the most accurate information.

More than thirty years ago The Psychological Corporation (Seashore, 1952) was cautioning educators that no test can claim to have one reliability coefficient, and that
reliability will vary with the range of talent within a group. Thus, it becomes "specific to the group on which it is estimated" (p. 3). This fact becomes crucial with regard to the use of norm-referenced, commercially available tests of achievement or ability. The range of talent of the norming population may be quite different from any given local population, thus producing a quite different spread in resulting scores. This variability is reflected in the coefficients obtained by any method. A wider range of talent within the tested group will yield much higher coefficients than will a more homogeneous range (Ebel, 1965). While current statistical theory permits the adjustment, or attenuation, of reliability estimates to accommodate skewed distributions, it would be implausible to suppose that any two populations match exactly. Information about the norming group is essential in order to correctly interpret reported reliabilities, but checking reliability with a school district's actual population will give much more accurate information.

While there is a general consensus of confidence in the reported reliability estimates of the Stanford Achievement Test, rechecking under local conditions seems to be a responsible approach. No matter how well-established or reputable, all measuring instruments need to be critically examined to ensure that decisions made regarding students
and instructional programs are based on accurate assumptions regarding all aspects of tests as they apply to specific groups of students. The Joint Committee on Standards for Educational Evaluation (1981) cautions evaluators to avoid the pitfall of "depending upon reliability results reported for a published instrument without considering the likely effects of differences in setting and sample" (p. 121), and Standard 6.7 in Standards for educational and psychological testing (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 1985) states that: "Test users should verify periodically that changes in populations of test takers, objectives of the testing process, or changes in available techniques have not made their current procedures inappropriate."

Burrill (1970) explains that the publishers of the Stanford Achievement Test have designed their standardization procedure to include attempts to guarantee that the sample of students tested is statistically random and is, in fact, a true cross-section of the general American student population. School districts are randomly chosen and then invited to participate in the standardization. This necessary process of voluntary participation allows the introduction of non-random, distorting effects of the distribution of students.
It becomes important for the testing service to give consumers accurate information about the actual representation of various groups within the norming population. The Stanford Achievement Test Norms booklet reports percentages of schools participating based on demographic factors. However, in their review of the Seventh edition of the test in Mitchell's Mental Measurement Yearbook, Subkoviak and Farley (1985) found no description of the "characteristics of the norm groups for particular grade levels" (p. 1451). Without checking reliability under local conditions, it is difficult for evaluators to determine the degree of confidence which can be placed in the scores.

In determining reliability estimates, Stanford Achievement Test publishers report using Kuder-Richardson Formula #20 to calculate coefficients for each skill area in all available forms and levels. These coefficients reflect the internal consistency of each test section. "Of the 280 Kuder-Richardson coefficients reported, 68% are above .90 and 97% are above .80" (Subkoviak, p.1451). These high correlations are computed by techniques based on a single test administration. This method will tend to somewhat overestimate the real consistency of a test (Zeller & Carmines, 1980). In addition, single sitting coefficients do not describe the stability of student scores (The
Psychological Corp., 1952). Reported alternate form coefficients are valid, but somewhat lower. 16% are above .90 and 81% are above .80. To arrive at these figures, one form is generally standardized and other forms equated to the one standardized. Standard errors of measurement and intercorrelations among tests are also reported. Stability over time coefficients are not reported.

Does actual test use at the fourth grade level in a large Catholic school district in upstate New York yield high reliability correlations? This study begins the process of computing local reliabilities by examining the reading subtests scores of the Stanford Achievement Test (SAT). Since one reliability coefficient doesn't address all considerations, it is important to refine the questions asked and to choose an appropriate statistical method to estimate the aspect of reliability being considered. If an evaluator asks, "Would this individual (or group) have achieved a similar score on this same test either a few days before or a few days after this administration"? concern is with stability of performance. The reliability coefficient derived from a test/retest method would most directly answer the question and is thus the focus of this study.

Kuder-Richardson Formula #20 (KR-20) coefficients appropriately report the degree of consistency throughout subtests of a battery. Comparisons between the KR-20
procedure and three other methods have justified its use with tests that show only moderate speededness (Swineford, 1973). Two relatively recent studies have addressed questions regarding the effects of speed on the Reading Comprehension subtest in the SAT battery. An extensive analysis of the validity and reliability of the SAT for use in the public schools of the Virgin Islands (Bliss, 1982, 1984) verified the comparability of the Stanford norming group and the local Virgin Islands school population. Adjusted reliability estimates for reading were numerically the same (Table 6). It was noted, however, that most students were not able to complete the Reading Comprehension subtest (1984). Another study in Palm Beach County, FL (Jolly, 1985) examined inconsistencies in student score profiles and determined that the speeded nature of the Reading Comprehension subtest was the apparent source of the discrepancies. The one reported score for comprehension does not allow for the separation of speed and power, and the published Stanford materials do not mention the measurement of rate. Since most students have difficulty completing the test in the time allowed, random marking, or guessing, seems to be distorting these scores, thus raising questions about their validity and reliability.

Twelve separate standards are identified by the AERA, APA and NCME (1985) as applying to reliability of
measurement. Standards 2.2 and 2.3 recognize that varying groups, conditions and situations influence reliability estimates, and the supporting comment recommends that coefficients not be accepted at face value, but interpreted based on very specific information about the relevant group (pp. 20-21). In order to more accurately assess the fit between the SAT reading subtests and the population of students in the schools of the Catholic Diocese of Rochester, and to more closely examine possible discrepancies in reading subtest scores, a local study of test reliability is indicated.

**Questions Addressed by the Study**

**Question 1:**

Are the Stanford Achievement Test Reading Comprehension, Word Study Skills and Vocabulary subtests consistent, or stable, in their measurement of achievement for the population of fourth grade students in the Rochester Catholic schools?

**Question 2:**

By how much does the stability vary from the accepted level ($r^2 .80$) to support test/retest reliability for each subtest?
Question 3:

Considering that student scores are only estimates of their true scores, and taking the reported Standard Error of Measurement for setting the limits of a student's true score, how many students exhibit a fluctuation in their two scores which extends beyond these limits? Are the same students showing these fluctuations across the three subtests?

Limitations of the Study

This study was limited to an examination of only the stability over time aspect of reliability assessment. To obtain a complete, more precise interpretation of reliability, information regarding test equivalence and item line analysis would also need to be collected for this population of students.

The small sample size (39 students) also limits the extent to which results may be generalized. A t-test to determine the significance of the correlation coefficient was, therefore, also applied to the data.

The assumption that the sample of students represented a normal distribution of intelligence was reasonable based on each student's level of functioning within the class, but IQ test scores for each student were not available.
Summary

In order for any test, or measuring instrument, to provide useful information about student performance, its reliability must be shown to be high enough to assure consistency of measurement from one situation to another. There is no one reliability estimate for a test and estimates vary with the tested group, even for reputable, well-designed standardized tests such as the Stanford Achievement Test (SAT). Reported correlation coefficients for SAT reflect either a measure of internal consistency, determined by a single administration, or alternate form reliabilities, determined by simultaneous administration of parallel forms. The first type of measure estimates how well the test is measuring the same thing throughout all items and the second gives information about the relative equivalence of forms and the effects of random error on the scores. Different statistical methods are employed to arrive at these coefficients. Stability of performance over time is not reported.

Measurement specialists have identified twelve standards which should guide the assessment of reliability, two of which recommend that reliabilities be interpreted based on information about the norming group and its
relation to the actual tested group. If reliabilities from actual test use are sufficiently high and closely resemble, or duplicate reported reliabilities, the test can be considered reliable, and possibly valid, for the local population.

The SAT subtests of Reading Comprehension, Word Study Skills and Vocabulary were chosen for investigation. Recent studies indicate that the speeded nature of the comprehension test raises questions about its validity and resultant scores may be distorting intercorrelations with other subtests. A test/retest procedure applied to the three subtests provides more accurate information about their stability for local interpretation, and begins the process of assessing the total test validity and reliability for Rochester Catholic schools use.
Chapter 2

Review of the Literature

Purpose

This study was initiated in order to examine the consistency with which reading ability is measured at the fourth grade level in the Catholic schools of the Diocese of Rochester by the Stanford Achievement Test subtests of Reading Comprehension, Word Study Skills and Vocabulary, Intermediate Level 1, Form E. An estimate of the stability of performance was sought through the use of a test/retest procedure. The literature review sought to examine the place of stability assessment within the total framework of reliability and validity of standardized reading measurement.

Standardized Measurement

Standardized measurement has been a part of the American educational process since the early part of this century. Beginning as an attempt to identify those with high intellectual ability and potential for college admission, this type of testing has broadened in scope to include
students at all levels of instruction. Even kindergarteners in some states are now being assessed and compared to an identified group or standard of performance.

Norm-referenced, standardized tests (NRT's) are the most common device used to measure student ability in the cognitive domain and are generally designed to assess either intelligence (aptitude) or achievement (the result of previous learning). However, it is "not possible to maintain rigorous theoretical distinctions between aptitude and achievement tests" (Mitzel, p.13). For this and other reasons standardization procedures include the administration of both intelligence and achievement tests to the population of students selected as the norming group. Procedures for constructing a NRT have been refined by new testing technologies and reflect updated curricular changes. They follow a pattern which has evolved since the 1920's and now includes: 1) a detailed analysis and summarizing of nation-wide curriculum; 2) identification of a set of content outlines specifying learning objectives; 3) the writing of test questions or items, which often includes several cycles of editing, rewriting and review by different experts; 4) development of experimental test forms; 5) selection of a random student population (the norm group) to represent the expected group of intended users of the test; 6) tryout administration of the test on the norm group; 7)
statistical analyses to determine reliability; 8) further item refinement; and 9) development of final forms (Burrill, 1970). The degree to which each step adheres to rigorous standards determines the resultant validity and reliability of the test.

Many major tests are prepared by commercial organizations and test publishers are assured a burgeoning market for their products. In a report prepared for the American Federation of Teachers, Ward and Gould (1980) state that testing has become a central issue in schools, and the use of tests of all kinds is increasing. This use is fueled in part by more federally funded and mandated programs, state-wide assessments of programs and minimum competency requirements. Surveys indicate that more than 87% of local school districts now have mandated tests. It can be safely assumed that most use standardized NRT’s as some part of their evaluation requirements (O’Neal, 1984). The use of such tests has consistently met with controversy. Criticisms have addressed the specification of content, the norming procedure, technical aspects of validity and reliability and psychological, emotional effects on students in addition to misunderstanding and misuse of results. The commercial distribution of these tests has assigned to school districts a role as consumers of a product designed for a broad national school market. Because of inevitable mismatches
between local curricula and test content, many studies have been conducted to determine the degree of similarity between content taught in school textbooks and content sampled by standardized tests. In an extensive analysis of four widely used tests and four textbook series Freeman, et al (1983) found that tests/text content was analogous "only when content is described at a relatively high level of generality" (p. 509). Variability in test content among publishers is evident and is not surprising since they are competing within the education market, and each claims to be the best representation of national school curricula. The Center for the Study of Evaluation (cited in Green, 1982) compiled a full documentation of the diversity of major test batteries available in the early 1970's in order to aid school districts in their search for a test which most closely matched local curricula.

The degree of importance attached to these NR measures remains very controversial while demands for their use continue. Ebel (1978) reflects the sentiment that achievements in learning which are not measured will not be recognized or rewarded, and failure to assess the progress of learning will likely cause learning to falter. This view has become ubiquitous among educators today and the current emphasis is toward providing more measurement data for more varied purposes. Normative measurement has been a powerful
means of communicating about the achievement of large groups, and most school districts still want this type of information. The scope of the conventional NR instrument is now being broadened to include objective referenced as well as survey data. The Stanford Achievement Test, as well as other NRT's, is now used to provide both types of interpretations. Some school districts have developed local, non-normed achievement tests and only periodically administer NRT's in order to equate their local tests to national percentiles to meet federal guidelines for specific programs, while at the same time assuring themselves of meeting local objectives (Christie & Conniff, 1981). Because the types of information requested are becoming more specific, test consumers are often attracted by publisher claims that one test can serve many purposes and yield in-depth prescriptive and/or interpretive information. Evaluators need to be wary not only of the possible non-equivalence of norm groups to their group, but also of differences in content, measuring approaches and the amount of tolerable error.

The proliferation of tests designed for specific or multiple purposes has often led to overtesting of students in some districts. Two recent approaches have attempted to alleviate this problem through consolidation of tests. Holmes (1986) reports that collected empirical information
demonstrates that the SAT and the California Assessment Program (CAP) evaluate common skills and unduly duplicate testing in California. Therefore, a set of multipurpose items were identified from the two tests and tried out to see if they would yield accurate NR information. Results showed that the procedure was feasible. It was recommended that a statewide content domain be established. Jolly (1983) documents the development of a procedure to yield objective referenced test data by combining subsets of curriculum relevant items from a NR test with items comprising a locally developed supplementary test. Results indicate a moderate relation to classroom performance.

These manipulations, combined with many other concerns, help to explain the current trend emphasizing the responsibilities of test users and the importance of communication between test publishers and users. Anastasi (1985) clarifies the functions of test users, those who choose tests and interpret scores to reach various decisions. In order to make informed interpretations these individuals need to be knowledgeable about both the "statistical properties of scores and the psychological characteristics of the behavior assessed by the tests" (p. iv). Thirteen standards have now been identified as specifically applying to test users (AERA, APA, NCME, 1985). These relate to all aspects of test use and underscore the
necessity for critical evaluation of all measuring instruments.

This search for precise measurement and interpretation of human behavior, which is extremely difficult to define in precise terms, has led to increasing concern with theoretical principles throughout the entire test development process. Statistical item analysis, for example, must also consider construct definition. Reliability coefficients must consider properly identified content domains. Multiple procedures need to be employed sequentially at various stages of test construction (Anastasi, p. xxvi). Published studies (Bliss, 1982; Jolly, 1985) examine aspects of validity and reliability concurrently. Other studies compare the various NRT's against each other (Amorilli, 1981), against other types of tests (Manning, G., 1985; Oliver & Arnold, 1978; Manning, M., 1985) or against local curriculum (Hawaii State Dept. of Education, 1984; Suydam, 1986). The findings of these and other studies indicate that standardized tests are moderately effective in assessing the attainment of local expectations, but do not address all evaluation considerations and should be used in conjunction with informal measures, teacher opinion or other formal devices. Very few studies recommend that their use be discontinued, primarily because, if they are properly used and understood,
they do provide a useful means of classifying individual students on a developmental scale. While this type of classification may not be attractive from a holistic educational viewpoint, its proponents praise the efficiency, accuracy and objectivity of standardized procedures, and claim that the provision of more indepth interpretive, prescriptive and technical information to school districts will help standardized survey tests to remain a major component in assessment systems. Ebel (cited in Bligh, 1979) recommends informed criticism of these tests as an impetus to their improvement. Critical concerns voiced by the National Education Association (NEA) (cited in Bligh, 1979) are mainly oriented toward content, but other validities of measurement are also questioned. Others such as Haney (1978) have recommended that more precise standards be developed to judge norming procedures and other technical qualities. Many of these recommendations have found concensus among measurement professionals.

Reading Measurement

The SAT has been cited by reviewers (Passow, 1978; Ebel, 1978; Subkoviak & Farley, 1985) as a model of an appropriately developed standardized achievement test battery, but critics continually decry the use of tests of
this type as adequate measures of reading ability. Current research directed toward discovering the intricacies of the acquisition of reading have caused the use of these tests to be even more vigorously debated. Some see their use as an aide to research. For example, by examining student NRT scores over a long time interval it was discovered that reading performance remains relatively constant, or stable, as students progress through school (Hopkins & Bracht; Amoriel, 1981; Hotard, 1983). Hopkins & Bracht and Amoriel showed that this stability of performance could be seen across various achievement test batteries, and Hotard found that early school functioning is a powerful predictor of later school functioning. If it can be assumed that these tests measure a general level of reading ability, they seem to do so consistently. Others claim that their use distorts instruction and shifts the emphasis of learning from an intrinsic individual pursuit to the pursuit of a given test rank or classification. Much of the debate seems to center around the question of exactly what is being measured, which leads ultimately to questioning the theoretical definition of reading which precedes the construction of items for any test. Two major theoretical models can be identified. One views reading primarily as a holistic process, a unified, interactive whole at all stages, and the other as an aggregation of discrete skills. Neither NRT's nor
criterion-referenced tests (CRT's) currently being used to assess achievement in reading are wholly satisfactory to proponents of either theory. Linn & Valencia (1986) reflect the sentiment of many by stating that there is a "relatively poor match between current theory and experimental research on the reading process and existing standardized tests of reading" (p. 35). Consequently, numerous other methods are often proposed as superior measures. These include such devices as Informal Reading Inventories (IRI's), teacher judgement, cloze procedures and direct measures of oral reading. Studies comparing the various techniques have had mixed results. Brown; Liebert; McCracken; and Sipay (cited in Coleman & Harmer, 1982) found wide variability in scores yielded by differing reading measures. Oliver & Arnold (1978) found standardized tests and teacher judgement to be significantly higher than IRI's in assessing third graders, tending to place students instructionally one to two years higher. Botel (1968) found that standardized tests overplaced more pupils than an IRI, but many were either correctly placed, or underplaced by the standardized silent reading tests. Smith (1980) found that the Metropolitan Achievement Test correlated closely with IRI and cloze procedures and appeared to accurately indicate instructional reading level. Layton (1983) found some teacher constructed cloze measurements to be as appropriate for measurement as
commercially prepared devices. Manning, G. (1985) found a statistically significant positive relationship between IRI and standardized test scores and a correlation between teacher opinion and all subtests of the SAT. Amoriell (1981) found significant discrepancies in grade level placement across four standardized tests (including SAT), and that more than 50% of instructional reading levels varied from two to five levels from those of an IRI. These and other studies indicate that no conclusive evidence can yet be assembled on the best device for measuring reading ability. Use of a combination of procedures is now favored.

An assessment of the reliability and validity of any reading instrument requires the consideration of theory at all stages. In an examination of curriculum based reading measures and standardized tests, Fuchs & Deno (1981) found that:

First, the degree of congruency between teacher placements and the curriculum-based placements varied with the instructional criteria used. Second, the extent of agreement between curriculum-based mastery measures and achievement test grade scores was different when different instructional criteria were employed. The degree of criterion validity of curriculum-based measures appeared to be dependent on the criteria employed in the measurement (p. 24).
Since publishers of NRT's are now providing CR interpretations of scores, the criterion validity of the NRT's needs to be more closely scrutinized. Cronin (1982) claims that since achievement and reading comprehension are "constructed" concepts, "it is impossible to make any claims about a measurement without some underlying theoretical network" (p. 2). In order to determine validity, a theoretical claim must be made and construct validity thus "subsumes all facets of validation (and reliability), as well as the procedures employed in constructing and revising assessment items" (p. 14).

Definitions of a reading universe, or domain, will derive from a publisher's theoretical perspective, but information about theoretical perspective is not usually provided by testing services. This is not surprising in light of the fact that tests attempt to mirror curriculum and an investigation by Durkin (1981) into comprehension instruction in five basal reader series uncovered evidence that none of the programs provided information about how their authors had conceived comprehension priorities. In fact, explicit instruction in comprehension is sparse. This area is coming under increasing scrutiny, and a two-year study by Langer (1982) investigated the connection between the process of constructing meaning (comprehension) and standardized tests. Using reading comprehension tests as
text with students of varied ages, it was possible to identify certain aspects of test language that can interfere with comprehension. These include density of ideas, over-reliance on assumptions and hypotheses, and assertions contrary to reader beliefs. The findings intensify the need to identify the purpose of testing. Standardized tests were shown to make adequate large scale discriminations between better and poorer school achievers, but they "do not measure the processes involved in the construction of meaning from a text (and do not) evaluate an individual's ability to manage those processes" (p. 33). Even when reading is viewed as a set of subskills, evidence does not suggest that standardized reading tests measure these skills well (Dole, 1981).

In choosing a reading test the purpose of testing should be included along with considerations of validity, reliability, norms and applicability, as well as the timed nature of the test. It is most often recommended that any standardized test be only one component of a comprehensive program which includes teacher observation, health and cumulative records, classroom work samples test samples, and case studies where indicated. Standardized testing becomes only one component in the entire process of reading evaluation, especially when the purpose is to measure growth or gains in ability. Marston (1983) found that greater
student gains were evident on a direct measure than on SAT scores. This is to be expected since NRT's primarily measure difference among individuals and relative position within a group and are not the best or most appropriate means to gauge individual changes.

Reading assessment requires the most rigorous attention to all phases of test reliability and validity including underlying theoretical construct or criterion assumptions which form the basis of item formation.

Reliability Assessment

The assessment of the technical qualities of a test has traditionally included an examination of reliability and validity. This assessment can be viewed as a process of examining separate components of a total framework which includes reliability aspects of stability, item analysis and equivalence or internal consistency, and validity aspects of content, criteria and constructs. Studies continue to extensively analyze each specific aspect, but always need to consider that other aspects impinge on and affect each other. Decisions regarding the usefulness of a test must account for all aspects. During the initial phases of test construction, validity concerns are paramount. During tryout
testing reliability is analyzed from collected scores and validity is reassessed as items are added or deleted. Once a test has been published, test users (consumers) can begin independent assessments by first examining reliability claims. To reflect reliable, consistent measurement a test must be as free as possible of random errors of measurement and "estimates of the reliability of a test should consider not only the relevant sources of error, but also the types of decisions anticipated to be based on the test scores" (AERA, APA, NCME, p. 20). A statistical coefficient needs to be considered in light of a subsequent decision. As noted previously, no test can claim to have one single reliability estimate. Estimates relate to specific groups and situations and may be generated to explain differing aspects of reliability. Without evidence of some degree of consistency of measurement across time and situation, a test cannot be valid for any purpose. This reliability is a property of a particular group of test takers and not of the test itself. It is a relative measure and does not depend upon the quality or validity of the test. It will vary with the set of test takers and may be distorted by test factors such as length or objectivity, or human variables such as motivation, attitude, physical/emotional condition and environment. These two sources, the test itself and the human participants, introduce error into the measurement. To
gauge the degree of this error, sets of scores are compared. Methods of comparison will vary depending upon the aspect of reliability under examination.

The assessment of stability involves administering the same test to the same group at different times (test/retest) and correlating the resultant pairs of scores to determine if there is evidence of consistent measurement over time. Because of error factors, the correlation is never expected to be perfect, but a relatively high correlation will give evaluators more confidence in the test/retest reliability of the test. Some types of tests may exhibit more stable measurement than others. Cook (1982) found that estimates obtained for aptitude test data had a higher degree of stability than those estimated for achievement tests. Periodic re-calibration of achievement test items may be necessary to produce more stable results. The SAT, as well as other reputable NRT's, is standardized and correlated concurrently with an aptitude test (The Psychological Corp., 1983). SAT publishers use the Otis-Lennon Mental Abilities Test.

Studies by Dearborn & Rothney; Peel; and Kreit (cited in Eichelberger, 1973) indicate that test scores tend to increase when a standardized instrument is repeatedly administered. Eichelberger (1973) investigated the assumption that remembering specific test items is a major
determinant of these increases by administering the Otis-Lennon Mental Abilities Test at three month intervals to 145 students. Resulting data tended to support the assumption that items were remembered, while other test-taking skills did not appear to significantly predict changes in scores. Zeller & Carmines (1980) refer to the problems that affect test/retest stability measurement, including memory effects, which lead to inflated reliability estimates, and reactivity (the measurement process itself inducing change in what is being measured), which may lower correlation estimates. Many studies have examined over time reliability and a review of this literature may be found in Wheaton, et al (cited in Zeller & Carmines, 1980). Studies such as the present one involve short time intervals and remembering needs to be considered in analyses.

The reliability aspect of equivalence examines "multiple indicators of a concept measured at a single point in time" (Zeller & Carmines, p. 54). With regard to achievement tests, each item in a given domain is considered a separate, equivalent measure of the underlying trait. The test can then logically be split into halves and the two halves correlated as if they were single representations of the measurement. This process is part of the standardization procedure for NRT's and produces the correlation coefficients for forms which contain homogeneous items.
Theoretically, however, it cannot be established with certainty that any given item measures only one construct or criterion. Usually more factors (smaller basic categories of the construct or criterion) are present, requiring further analysis to determine overlaps. Identifying, extracting and correlating these factors becomes a critical step in the assessment process and involves validity interpretations. A study of reliability becomes a comment on validity and requires a re-clarification of underlying theoretical conceptualizations. Equivalent reliability coefficients are an interpretation of scores and only an indirect indication of the theoretical equivalence of the items composing the test. High coefficients indicate high correlations between separate items and suggest high consistency of measurement. SAT publishers employ internal consistency methods and resultant coefficients are high (Subkoviak & Farley, 1985). Toole (1970) and Bliss (1982) found similarly high coefficients when independently analyzing SAT. In fact, SAT has been considered for many years as a model for standardized NRT's. Many studies use scores generated from its use as bases for other investigations (Kilgallon & Mueller, 1986; Manning, M., 1986; Manning, G., 1985). As early as 1957, of 94 school districts using achievement batteries, 45% used SAT (Goodman, 1960). In 1965 the state of California adopted SAT for statewide use. A survey by
O'Neal (1984) indicates that many districts now use more than one test and the California Achievement Test is the most popular. SAT, however, is still respected and used by a great many districts.

Validity Assessment

From a general viewpoint a test may be considered valid for a particular purpose if accumulated evidence supports the "appropriateness, meaningfulness, and usefulness of the specific inferences made from the test scores. The inferences regarding specific uses of a test are validated, not the test itself" (AERA, APA, NCHE, p. 9). Test users need to make decisions based upon responsible examination of test validity for their group of students. In addition, current concerns regarding the legal aspects of testing have led to more indepth scrutiny of the technical qualities of popular standardized tests. In response to the enactment of state laws requiring disclosure of information to test takers, and subsequent court cases, Brown (1980) compiled a background report summarizing the pro's and con's of the issue. He asserts that it is virtually impossible to verify validity to the satisfaction of everyone involved in test use, and consensus has never existed, even within the testing community about exact standards, or the ability of test makers to actually meet the standards already defined.
Reliability standards can be met with greater precision, but the underlying assumptions of the tests themselves are subject to criticism and, therefore, so are interpretations of stated correlation coefficients. Evaluators must operate within the context of degrees of certainty.

Numerous studies have been undertaken to ascertain degrees of test validity. Many involve investigation of the uses of a measuring instrument in situations similar to those in which it was tried out by the sample, or norming, group. Decisions can then be made as to whether validity assumptions may be generalized. The aspects most often examined are content, criteria and construct validity, although demarcation lines between the three are often vague. Content validity refers most often to a test’s accurate sampling of a particular domain of content. To initially construct a test a researcher or publisher must "first, specify the full domain of content that is relevant to the particular measurement situation; second, sample specific (knowledge) from this collection; and finally, put (the knowledge) in a form that is testable" (Carmines, p. 20). Limitations are inherent in this process, beginning with inadequate definitions of the universe of content to be measured, and including the difficulties of determining to what extent attainment of the relevant knowledge can be measured. Most often educational studies have approached the
issue by examining the degree of similarity between local curricula and test items. Bliss (1982) found an acceptable match between the curriculum of schools in the United States Virgin Islands and content sampled by the SAT. Other studies found mismatches in the area of mathematics (Hawaii State Dept. of Education, 1984; Suydam, 1986). Attempts to effect a closer match between textbooks and tests have fueled vigorous debate about the existence of and advantages/disadvantages of a homogenized national curriculum. Freeman et al (1983) examined the match between textbooks and four widely used standardized tests and found a small core of commonality, but a greater extent of diversity. The diversity has undoubtedly contributed to further proliferation of other types of testing considered to be more content valid. Nelson (1970) suggested that if a student is tentatively placed for instruction at a given level based on an achievement test score, and subsequently is able to handle assignments at that level, his success indicates a "partial vindication for the validity of the test" (p. 108). Content validity may also be established if student scores are reasonably high and normally distributed.

The assessment of criterion validity requires the collection of evidence to show either that "early test data can be used to estimate criterion scores that will be obtained in the future" or that a measure obtains
"prediction and criterion information simultaneously" (AERA, APA, NCME, p. 11). Essentially, the test will be used to estimate a behavioral outcome, so there will be a different validity for each identified criterion. Limitations on this type of assessment relate to the inability to effectively measure abstract criteria. How is a criterion such as "success" at any level precisely measured? Webb (1984) contrasted the predictive validity of three standardized tests and found all of them to be adequate. The best of the three, when combined with class rank, accounted for 68% of the variance in reading-related grades. Many years of use by numerous school districts has established SAT scores as good predictors of general school ability, but in the search for more precision, tests labeled as CRT's now compete with more traditional NRT's and different approaches are used to assess the validity of each. However, Haladyna (1976) argues that the use of the term criterion-referenced measurement is paradoxical because "an analysis of the distinctions commonly drawn between CR and NR measurement, coupled with accumulating test data, suggests that there is only one measurement construct with two functions, NR and CR" (p. 3) and almost any test can have a CR interpretation. Hambleton (1972) acknowledges that a test "may be classified as NR according to one definition, CR according to another" (p. 2). Since tests like SAT, CAT and IOWA contain subscales of
homogeneous items which measure common skills, each item is in essence a mini-test and reliability and validity assessment focuses on subscale scores. Criterion and content validity intertwine because, according to Popham & Husek (cited in Hambleton, 1972) "content validity is determined by a carefully made judgement based on the test's apparent relevance to the behaviors legitimately inferable from those delimited by the criterion" (p. 8). For tests identified primarily as NR, criterion validity assessment begins with item analysis, which leads in turn to an examination of the underlying theoretical assumptions regarding the criterion and the development of a large pool of homogeneous items which purport to measure the criterion. Theoretical underpinnings in essence determine the validity.

This is especially true regarding construct validity which "focuses primarily on the test score as a measure of (a) psychological characteristic" (AERA, APA, NCME, p. 9). A conceptual framework, even if imperfect, defines the meaning of the construct under examination and helps determine how it should best be measured. Andrulis (1972) attempted to identify within an achievement test the presence of constructs which might relate with individual performance. Statistical procedures of factor and multiple discriminant analysis give clues as to the validity, but theoretical foundations need always to be considered. Statistical
procedures provide estimates for reasonable judgements regarding valid use of an instrument in specific situations. Green (1984) examined construct/content validity together but found it necessary to first clarify his theoretical conceptions. He links growth and development to normed ability or trait construct, and his findings suggest that when a NR trait involves growth, content areas that do not grow do not belong in its measurement. This further complicates the process for test publishers who must define a trait or construct and then determine how best to measure it. Validity rests upon the truth of the assumptions on which the procedures rest (Zeiler & Carmines, 1980). Standards 1.8 and 1.10 (AERA, APA, NCME, p. 15) require that evidence related to the specific constructs be presented whenever inferences are made regarding construct measurement and that test scores can be shown to be closely associated with theoretical variables. Careful validity assessment is the responsibility of both test publishers and test users.
Summary

Standardized, norm-referenced measurement, comparing student achievement and abilities against a predetermined standard or a specified group, remains a central consideration in American schools today. Demand for its use continues despite controversial opinions. Procedures for constructing standardized tests are guided by identified standards which apply to all test aspects, especially validity and reliability. Since more and varied uses are being made of collected data, test users must also be guided by standards and must assume the responsibilities of being knowledgeable about the technical and theoretical bases of these tests, as well as about student development and performance in the areas being measured. While standardized tests have been shown to be moderately effective in assessing local school district objectives in content areas, and while they do provide efficient, objective measurement, they are only one part of the student evaluation process and test users must consider the amount of tolerable error they exhibit. Informed criticism may provide an impetus to their improvement.

Norm-referenced tests (NRT's) are often criticized as inappropriate devices for reading measurement. Some see their use as distorting instruction and point to other measures such as informal reading inventories, close
procedures, direct oral measures, etc. as more accurate measures. Others see their use as spurring needed research. Differing reading theories lead to differences in the identification of a valid content domain for test item construction, especially in the area of reading comprehension where there is no consensus on a definition. Some studies have even identified test factors that appear to interfere with comprehension. Theory must be considered at all stages of test construction and assessment. Numerous studies comparing various measures have produced mixed results and no conclusive evidence exists as to the best reading measure, or to what is actually being measured. Multiple procedures are most often recommended for effective evaluation. The purpose for the testing becomes critical in choosing the proper measure. Standardized survey instruments such as Stanford Achievement Test (SAT) are designed to measure a general level of achievement and variability among students. They are not properly used to analyze changes in individual students, or to diagnose difficulties.

Decisions regarding the usefulness of a test must account for all aspects of validity and reliability. Any test must be acceptably reliable in order to be valid for any purpose. Reliability estimates reflect the amount of random error introduced into the measurement by the test itself and by the human participants. The two major aspects
of reliability are stability, consistent measurement over time, and equivalence, equal measurement across all items. Aptitude tests tend to be more stable than achievement tests and are often administered concurrently as part of the standarization process for reputable NRT's. For local assessment of a published test, a test/retest procedure may be used to determine the stability. Problems with this procedure include memory effects, if the time interval between tests is short, and reactivity effects which are present as participants relate and react to the testing situation. Equivalence is assessed by administering the test at one sitting, splitting it into halves and correlating the halves. The major problem with this procedure involves the degree of certainty that all items are truly homogeneous. Reliability is affected by the validity assumptions of each item. As an adequately developed model of a NRT, the SAT demonstrates high internal consistency, or equivalence, and is widely used and respected for its measurement of general achievement.

A test is valid if the inferences made from the test scores are appropriate, meaningful and useful. It is virtually impossible to verify validity to the satisfaction of all users, and is often difficult to identify and meet standards for validity, but degrees of certainty can be attained for standardized tests when they are used in
situations and with populations similar to the norming group. The three aspects of validity most often examined are content, criterion, and construct. Content validity implies an adequate definition of a domain of knowledge and an acceptable match between local curriculum and test sample items. The diversity among tests and differing theories about the acquisition of knowledge preclude the existence of a perfect curriculum/test match, but evaluators need to ascertain that the match is as close as possible. Criterion validity assessment requires that it be shown that a test score adequately predicts future scores in the measurement of a given criterion. NRT's are often used to provide criterion related information, so the underlying theoretical definitions and construction of criterion items need to appropriately apply to the purpose for which the test will be used. Test user responsibility continues as construct validity is assessed. Constructs, psychological characteristics, are complex and abstract and are not simply or easily defined. Their measurement requires careful, informed decisions about acceptable definitions.

All of these aspects of test use and construction impinge on and affect each other. Critical evaluation and attention to accepted standards will assure more valid, reliable measurement of student ability and achievement.
Chapter 3

Design of the Study

Purpose

This study sought to investigate the degree of stability of the Reading Comprehension, Word Study Skills and Vocabulary subtests of the Stanford Achievement Test, Intermediate level I, Form E, which are used to assess students in the Catholic schools of the Diocese of Rochester. The determination of stability of performance over time aids in the process of determining total test reliability.

Questions Guiding the Study

Question 1:

Do each of the three subtests under investigation measure achievement consistently at different points in time for the fourth grade population in Rochester Catholic schools?

Question 2:

By how much do the computed estimates vary from the accepted level (r $\geq .80$) to support validity?
Question 3:

By how much do score estimates vary from the range predicted by the Standard Error of Measurement (SEM), which sets the limits to encompass each student's true score? What factors may be identified as contributing to the variation?

Methodology

Students in two classrooms were randomly chosen from among the 80 classes (1600 students) comprising the fourth grade population in the Rochester Catholic Diocese schools. The 39 participating students were assumed to represent a normal distribution of intelligence. Classes are heterogeneous groupings and include an approximately equal mix of male/female and inner city/suburban students. The students participated in the regular spring group administration of the Stanford Achievement Test Battery, Intermediate Level I, Form E, including the three subtests of Reading Comprehension, Word Study Skills and Vocabulary, which were the focus of this study. The tests were administered on a single day and according to the instructions in the manual provided by the publisher. These instructions include completion of a sample item for each test, attention to time limits, proper matching of each item number in the test booklet to numbers on the answer sheet, and responding by darkening an appropriate bubble.
Time limits and number of items for the three subtests were:

- **Reading Comprehension**: 30 min./60 items
- **Word Study Skills**: 35 min./60 items
- **Vocabulary**: 30-40 min./36 items

The time limits are purportedly designed to assure completion. However, nine students failed to complete the entire Reading Comprehension test. All students responded to all items in the Word Study Skills test and two students did not complete the final item in the Vocabulary test.

Nine days later the Reading Comprehension test (same Form E) was readministered under similar conditions and using the same instructions as the first administration. Four students finished the test in 15 minutes and all were able to complete all items in 28 minutes. The following day Word Study Skills and Vocabulary tests were readministered under similar conditions. All students were able to finish each test more quickly than the first time.

The number of correct responses (raw scores) were tabulated for each test and the sets of scores from the two administrations were paired and correlated using the Pearson product-moment formula. Further analysis located the scores creating variance. The time interval between the first test and the retest was viewed as short enough to prevent any
significant new learning from occurring and long enough to somewhat alleviate memory effects.

**Summary**

A group of thirty-nine students from two classrooms comprised the sample population for this study. Through a test/retest procedure using the same form of the three subtests examined, scores were generated by tabulating the number of correct responses for each student. Sets of scores were then paired and correlated for each test by using the Pearson product-moment formula. The data was then analyzed to discover the sources of instability.
Chapter 4

Analysis of the Data

Purpose

This analysis was undertaken to examine the stability of the Reading Comprehension, Word Study Skills and Vocabulary subtests of the Stanford Achievement Test, Intermediate Level 1, Form E, as they are used by fourth grade students in the Rochester Catholic schools.

Questions Guiding the Analysis

Question 1:

Does the strength of the relationship between test/retest scores for each subtest indicate consistent, stable measurement of the skills or abilities tested?

Question 2:

What is the correlation coefficient variance from the accepted level of 64% (r²=.64).

Question 3:

How varied is the fluctuation in student score estimates from one test to the other beyond the range
predicted by the SEM for each subtest? Are any causes evident?

Findings and interpretations

Reading Comprehension:

A strong positive relationship was found between the test/retest scores for the Comprehension subtest, indicating acceptable stability ($r = .84$). 70% of the variation in the retest scores can be explained by knowing the first test scores ($r^2 = .70$). The retest score estimates of 27 of the students in the sample group fell within the range predicted by the Standard Error of Measurement (SEM) reported for the SAT norming group. 12 pairs of scores apparently account for the major portion of the instability. Of these 12, 5 pairs were from 4 to 8 points above the group means for both tests, and the remaining seven pairs were from 1 to 9 points below the group means. Twenty-nine students had a greater number of correct responses on the retest and 5 students duplicated their original score.

One factor appears to have had a decided lessening effect on the correlation coefficient. Nine students were unable to complete all 60 items in the allotted time of 30 minutes during the first administration of the test. One student did not respond to the last 20 items, two did not
complete the last 11 items, and the remaining six left an average of 4 final items incomplete. The raw scores of five of these students were dramatically higher on the retest when all students completed all items in the same 30 minute time allotment. The SEM for this subtest is ± 3.0 points and these five students exhibited variations ranging from 4 to 15 points. These wide fluctuations outside of the range predicted by the SEM give the appearance of a lesser degree of stability. The cause seems to be the speeded nature of the test. When the sample group is adjusted to include only those students who completed all items on both tests (30 students) a much higher degree of stability is evidenced (r = .91).

Since the one score reported for this subtest does not separate speed (rate) from power of comprehension, inferences made from the scores should consider completion factors. The fact that four students completed the retest in 15 minutes and all were able to complete it within 28 minutes indicates that remembering of former responses increased the completion rate and probably led to more correct responses.
Test   | Retest
------|------
Number | 39   | 39   
Mean   | 47.74| 50.46|
Sta.Dev.| 7.16 | 6.94 |

\[
r = 0.839 \\
r^2 = 0.704 \text{ variation explained} \\
1 - r^2 = 0.296 \text{ variation unexplained}
\]

\[
t(0.05) = 9.427 \text{ (at 37 df)}
\]

*SEM = 3.0

*for the SAT norming group for this subtest
Reading Comprehension

![Graph showing the relationship between First Test Raw Scores and Re-test scores. The graph indicates a positive correlation with a scatter plot and a line of best fit.]
Word Study Skills:

All students completed all 60 of the syllabication and vowel/consonant items comprising the Word Study Skills subtest. Correlation of the resultant pairs of scores showed a very strong positive relationship ($r = .89$). Stable performance over time is indicated. 79% of the variation in retest scores could be explained by knowing the first test scores ($r^2 = .79$). While only 20 of the 39 retest scores fell within the range predicted by the SEM of 2.4, an additional 12 score estimates were within 1 or 2 points of the range. Fluctuations were not very wide. 13 of the 19 pairs of scores outside of the expected range were very near, or slightly above, the group mean, and of the remaining 6 pairs, only 2 were well below the mean. 23 students had a greater number of correct responses on the retest, 1 had the same number, and 15 had fewer. For this group, this test appears to measure consistently over time.
<table>
<thead>
<tr>
<th></th>
<th>Test</th>
<th>Retest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>39</td>
<td>39</td>
</tr>
<tr>
<td>Mean</td>
<td>46.21</td>
<td>46.95</td>
</tr>
<tr>
<td>S.D.</td>
<td>6.69</td>
<td>7.41</td>
</tr>
</tbody>
</table>

\[ r = 0.89 \]
\[ r^2 = 0.792 \text{ variation explained} \]
\[ 1 - r^2 = 0.208 \text{ variation unexplained} \]

\[ t_{(0.05)} = 11.558 \text{ (at 37 df)} \]
\*SEM \* = 2.4

*for the SAT norm group for this subtest*
Word Study Skills

Retest Raw Scores vs. First Test Raw Scores
Vocabulary:

A marginal degree of stability was found for the SAT Vocabulary subtest for this sample of students ($r = .79$). Only 62% of the variation in retest scores could be accounted for by knowing the first test scores ($r^2 = .62$). 12 of the 39 retest scores fell outside of the range predicted by the SEM of 3.2. All but 2 students responded to all items in the allotted time for both administrations, and those 2 left only the final item incomplete. 5 scores were duplicated, 20 retest scores showed a greater number of correct responses, and 14 had fewer correct responses. Of the scores that fluctuated beyond the expected range, only 3 were at or above the group mean. The remaining 9 were 1 to 9 points below the means for both tests.

Since score estimates did not correlate at or above the accepted reliability level of .80, error factors creating the instability need to be investigated. One reasonable explanation may be that students' attention is more easily distracted due to the listening nature of the test. Questions and answers are read aloud by the examiner and students may more frequently guess at answers if they miss hearing some of the content. Other factors which may have contributed to the added instability include the possible lack of content validity of the tested vocabulary for this sample of students and/or student fatigue.
<table>
<thead>
<tr>
<th></th>
<th>Test</th>
<th>Retest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>39</td>
<td>39</td>
</tr>
<tr>
<td>Mean</td>
<td>27.23</td>
<td>28.28</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>4.82</td>
<td>4.71</td>
</tr>
</tbody>
</table>

\[ r = 0.79 \]
\[ r^2 = 0.624 \text{ variation explained} \]
\[ 1-r^2 = 0.376 \text{ variation unexplained} \]

\[ t_{(0.05)} = 7.9 \text{ (at 37 df)} \]
*SEM = 3.2

*for the SAT norming group for this subtest*
Vocabulary

First Test Raw Scores

Retest Scores
Summary

Correlated student scores indicate that the Reading Comprehension and Word Study Skills subtests of the SAT, Intermediate Level I, Form E are acceptably stable in their measurement of student performance over time for this sample of students. Comprehension scores correlated at $r = .84$ and Word Study Skills scores at $r = .89$. It was noted, however, that the correlation index might have been much higher for the Comprehension test if all students had completed all items. The speeded nature of the test appears to be distorting student true scores to a greater extent than would be normally expected, and the distorted scores are creating a greater degree of instability.

The Vocabulary subtest scores correlated at $r = .79$, showing only a marginal degree of stability. Error factors of the test, test administration, or students themselves are causing scores to fluctuate more than would normally be expected.

The paired, correlated scores for each of the three subtests were found to be positively and significantly related.
Chapter 5

Conclusions and Implications

Purpose

This study investigated the stability of the Reading Comprehension, Word Study Skills and Vocabulary subtests of the Stanford Achievement Test, Intermediate Level I, Form E as it was utilized by fourth grade students in the Rochester Catholic schools. The investigation further clarified the place of stability assessment within the total framework of reading measurement.

Conclusions

The results of this study indicate that the Reading Comprehension and Word Study Skills subtests of the SAT form and level investigated exhibit an acceptable degree of stability when used by the fourth grade sample of students from the Rochester Catholic schools. They appear to measure achievement consistently over time. The respective correlation coefficients of .84 and .89 exceeded the accepted level ($r \geq .80$) necessary to support, but not alone sufficient to make the case for validity. An even stronger
A degree of confidence might be placed in the stability of measurement of the Comprehension subtest if rate and power of comprehension were not combined in a single score. Wide fluctuations in the scores of students who did not complete the first test in the allotted time tended to depress the computed coefficient for this test.

The Vocabulary subtest scores contained slightly more instability than is generally accepted. Further analysis is indicated in order to identify the contributing error factors. The coefficient of .79 is, however, very close to the level for acceptable reliability. If the test can be shown to be valid for its intended uses, this stability estimate should not prevent its use.

Retest scores for Reading Comprehension showed that the number of correct responses for the majority of students increased. This was not true for Word Study Skills and Vocabulary, where an almost equal number of scores decreased as increased. This would seem to indicate that students tended to guess more on the latter two tests, thus increasing the errors of measurement. Although the stability of Word Study Skills was very strong, it too might have been higher if the guessing factor were diminished.

Twelve students had scores which placed them outside of the range predicted by the SEM for the Reading Comprehension subtest. Five students had scores which were considerably
higher on the retest. Word Study Skills scores did not fluctuate to such a great extent. Although nineteen students scored outside of the predicted range, twelve of them were within one or two points of the range. Twelve of the thirty-nine retest scores for Vocabulary fell outside of the predicted range.

The "memory effect" was evident in this study based on the fact that all students finished the retest more quickly than the first test. Other test/retest contamination effects were not so obvious, but can be presumed to be present in some degree.

Individual student scores did not fluctuate consistently across the three subtests, but the scores on all three do reflect the presence of factors which cause student performance to vary when assessed at different points in time. The relative strength of these factors does not cause the subtests to be unreliable, but is strong enough to warrant caution on the part of teachers and other evaluators who might make inferences based on these scores. They are sufficiently stable if all other evidence of equivalence and content, criterion, construct validity point to their appropriate use. These results appear to coincide with other studies which have found standardized reading tests to be moderate indicators of student achievement.
Implications

The degrees of instability of performance noted for the three subtests suggest that these test scores be used as only one component in the assessment of student reading achievement and interpreted as only an estimate of a general level of functioning. Communication between those who use the Stanford Achievement Test and the test publisher concerning the effects of speed on Comprehension test scores would hopefully facilitate alterations in the number of items or revisions of the time allotment so that student scores would not be distorted by completion rates.

Given the position of stability estimation within the total framework of reliability and validity assessment of reading measurement devices, the results of this study should induce further analysis and investigation of equivalence and all validity aspects for these subtests. Only when viewed in combination with all other evidence can these results infer the appropriateness or usefulness of these tests for local purposes.
Summary

Based on analysis of the data in this study, the Reading Comprehension and Word Study Skills subtests of the SAT, Intermediate Level I, Form E were found to be stable measuring devices for fourth grade students in Rochester Catholic schools. The Vocabulary subtest was found to be marginally stable. The Comprehension scores were markedly affected by completion rates and Word Study Skills and Vocabulary scores may reflect random guessing. However, if other evidence supports the valid use of these tests, the results of this study support their stability. In combination with the high estimates of their consistency reported for the norming group, they may be considered reliable measures.
References


Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 076 657)


Hopkins, K. D. & Bracht, G. H. (no date). A longitudinal
study of constancy of reading performance. No publisher cited. (ERIC Document Reproduction Service No. ED 069 782


Langer, J. A. (1982). The construction of meaning and the


Education, New Orleans, LA. (ERIC Document Reproduction Service No. 074 095)


Appendix A

Student Raw Scores

**Reading Comprehension** | **Word Study Skills** | **Vocabulary**
---|---|---

<table>
<thead>
<tr>
<th>Student</th>
<th>#</th>
<th>Test</th>
<th>Retest</th>
<th>Test</th>
<th>Retest</th>
<th>Test</th>
<th>Retest</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>44</td>
<td>45</td>
<td>45</td>
<td>44</td>
<td>21</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>54</td>
<td>55</td>
<td>50</td>
<td>48</td>
<td>35</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>(4) *</td>
<td>27</td>
<td>27</td>
<td>28</td>
<td>18</td>
<td>21</td>
</tr>
<tr>
<td>4</td>
<td>36</td>
<td>(5)</td>
<td>38</td>
<td>42</td>
<td>48</td>
<td>25</td>
<td>24</td>
</tr>
<tr>
<td>5</td>
<td>53</td>
<td>53</td>
<td>47</td>
<td>50</td>
<td>29</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>54</td>
<td>56</td>
<td>41</td>
<td>48</td>
<td>33</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>52</td>
<td>58</td>
<td>49</td>
<td>50</td>
<td>33</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>52</td>
<td>57</td>
<td>47</td>
<td>53</td>
<td>32</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>42</td>
<td>(6)</td>
<td>48 (1)</td>
<td>52</td>
<td>47</td>
<td>27 (1)</td>
<td>29</td>
</tr>
<tr>
<td>10</td>
<td>44</td>
<td>43</td>
<td>31</td>
<td>33</td>
<td>25</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>43</td>
<td>45</td>
<td>47</td>
<td>51</td>
<td>22</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>47</td>
<td>49</td>
<td>45</td>
<td>47</td>
<td>30</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>53</td>
<td>55</td>
<td>46</td>
<td>48</td>
<td>32</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>46</td>
<td>(2)</td>
<td>50</td>
<td>40</td>
<td>42</td>
<td>31</td>
<td>30</td>
</tr>
<tr>
<td>15</td>
<td>52</td>
<td>55</td>
<td>50</td>
<td>52</td>
<td>28</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>56</td>
<td>58</td>
<td>55</td>
<td>56</td>
<td>31</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>50</td>
<td>53</td>
<td>46</td>
<td>45</td>
<td>29</td>
<td>29</td>
<td></td>
</tr>
</tbody>
</table>

*number of final items incomplete
<table>
<thead>
<tr>
<th>Reading Comp.</th>
<th>WSS</th>
<th>Voc.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test</strong></td>
<td><strong>Retest</strong></td>
<td><strong>Test</strong></td>
</tr>
<tr>
<td>18</td>
<td>52</td>
<td>54</td>
</tr>
<tr>
<td>19</td>
<td>51</td>
<td>52</td>
</tr>
<tr>
<td>20</td>
<td>58</td>
<td>58</td>
</tr>
<tr>
<td>21</td>
<td>51</td>
<td>54</td>
</tr>
<tr>
<td>22</td>
<td>38</td>
<td>(20)</td>
</tr>
<tr>
<td>23</td>
<td>32</td>
<td>35</td>
</tr>
<tr>
<td>24</td>
<td>53</td>
<td>50</td>
</tr>
<tr>
<td>25</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>26</td>
<td>54</td>
<td>54</td>
</tr>
<tr>
<td>27</td>
<td>43</td>
<td>(11)</td>
</tr>
<tr>
<td>28</td>
<td>41</td>
<td>46</td>
</tr>
<tr>
<td>29</td>
<td>51</td>
<td>53</td>
</tr>
<tr>
<td>30</td>
<td>55</td>
<td>51</td>
</tr>
<tr>
<td>31</td>
<td>51</td>
<td>55</td>
</tr>
<tr>
<td>32</td>
<td>52</td>
<td>52</td>
</tr>
<tr>
<td>33</td>
<td>52</td>
<td>(2)</td>
</tr>
<tr>
<td>34</td>
<td>54</td>
<td>51</td>
</tr>
<tr>
<td>35</td>
<td>45</td>
<td>(4)</td>
</tr>
<tr>
<td>36</td>
<td>45</td>
<td>52</td>
</tr>
<tr>
<td>37</td>
<td>38</td>
<td>35</td>
</tr>
<tr>
<td>38</td>
<td>57</td>
<td>60</td>
</tr>
<tr>
<td>39</td>
<td>38</td>
<td>(11)</td>
</tr>
</tbody>
</table>
Appendix B

Student Raw Scores in Descending Order

<table>
<thead>
<tr>
<th>Student</th>
<th>Reading</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>Comprehension</td>
<td>Word Study Skills</td>
<td>Vocabulary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>Retest</td>
<td>Test</td>
<td>Retest</td>
<td>Test</td>
<td>Retest</td>
<td>Test</td>
<td>Retest</td>
<td>Test</td>
</tr>
<tr>
<td>20</td>
<td>58</td>
<td>58</td>
<td>26</td>
<td>56</td>
<td>53</td>
<td>34</td>
<td>35</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>57</td>
<td>60</td>
<td>16</td>
<td>55</td>
<td>56</td>
<td>2</td>
<td>35</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>56</td>
<td>58</td>
<td>27</td>
<td>55</td>
<td>56</td>
<td>6</td>
<td>33</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>55</td>
<td>51</td>
<td>22</td>
<td>54</td>
<td>54</td>
<td>7</td>
<td>33</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>54</td>
<td>56</td>
<td>9</td>
<td>52</td>
<td>47</td>
<td>30</td>
<td>33</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>54</td>
<td>55</td>
<td>29</td>
<td>51</td>
<td>54</td>
<td>38</td>
<td>33</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>54</td>
<td>54</td>
<td>32</td>
<td>51</td>
<td>54</td>
<td>26</td>
<td>32</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>54</td>
<td>51</td>
<td>28</td>
<td>51</td>
<td>49</td>
<td>13</td>
<td>32</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>53</td>
<td>55</td>
<td>30</td>
<td>51</td>
<td>49</td>
<td>8</td>
<td>32</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>53</td>
<td>53</td>
<td>18</td>
<td>51</td>
<td>49</td>
<td>16</td>
<td>31</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>53</td>
<td>50</td>
<td>15</td>
<td>50</td>
<td>52</td>
<td>20</td>
<td>31</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>52</td>
<td>58</td>
<td>2</td>
<td>50</td>
<td>48</td>
<td>14</td>
<td>31</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>52</td>
<td>57</td>
<td>38</td>
<td>49</td>
<td>56</td>
<td>12</td>
<td>30</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>52</td>
<td>57</td>
<td>35</td>
<td>49</td>
<td>54</td>
<td>28</td>
<td>30</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>52</td>
<td>55</td>
<td>20</td>
<td>49</td>
<td>53</td>
<td>5</td>
<td>29</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>52</td>
<td>54</td>
<td>33</td>
<td>49</td>
<td>52</td>
<td>17</td>
<td>29</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>52</td>
<td>52</td>
<td>7</td>
<td>49</td>
<td>50</td>
<td>33</td>
<td>29</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>51</td>
<td>55</td>
<td>34</td>
<td>49</td>
<td>47</td>
<td>31</td>
<td>28</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>51</td>
<td>54</td>
<td>21</td>
<td>48</td>
<td>49</td>
<td>35</td>
<td>28</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>Reading Comp.</td>
<td>WSS</td>
<td>Voc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-----</td>
<td>-----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>51</td>
<td>53</td>
<td>39</td>
<td>48</td>
<td>45</td>
<td>15</td>
<td>28</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>51</td>
<td>52</td>
<td>8</td>
<td>47</td>
<td>53</td>
<td>9</td>
<td>27</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>50</td>
<td>53</td>
<td>11</td>
<td>47</td>
<td>51</td>
<td>32</td>
<td>27</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>47</td>
<td>49</td>
<td>5</td>
<td>47</td>
<td>50</td>
<td>21</td>
<td>27</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>46</td>
<td>50</td>
<td>36</td>
<td>47</td>
<td>39</td>
<td>26</td>
<td>27</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>45</td>
<td>52</td>
<td>13</td>
<td>46</td>
<td>48</td>
<td>19</td>
<td>27</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>45</td>
<td>48</td>
<td>17</td>
<td>46</td>
<td>45</td>
<td>18</td>
<td>26</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>45</td>
<td>45</td>
<td>12</td>
<td>45</td>
<td>47</td>
<td>10</td>
<td>25</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>44</td>
<td>45</td>
<td>1</td>
<td>45</td>
<td>44</td>
<td>4</td>
<td>25</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>44</td>
<td>43</td>
<td>19</td>
<td>44</td>
<td>47</td>
<td>37</td>
<td>23</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>43</td>
<td>58</td>
<td>23</td>
<td>43</td>
<td>40</td>
<td>39</td>
<td>23</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>43</td>
<td>45</td>
<td>4</td>
<td>42</td>
<td>48</td>
<td>11</td>
<td>22</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>42</td>
<td>48</td>
<td>37</td>
<td>42</td>
<td>41</td>
<td>27</td>
<td>22</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>41</td>
<td>46</td>
<td>6</td>
<td>41</td>
<td>48</td>
<td>1</td>
<td>21</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>38</td>
<td>53</td>
<td>14</td>
<td>40</td>
<td>42</td>
<td>23</td>
<td>21</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>38</td>
<td>47</td>
<td>24</td>
<td>40</td>
<td>41</td>
<td>22</td>
<td>20</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>38</td>
<td>35</td>
<td>31</td>
<td>36</td>
<td>34</td>
<td>29</td>
<td>20</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>36</td>
<td>38</td>
<td>10</td>
<td>31</td>
<td>33</td>
<td>24</td>
<td>20</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>32</td>
<td>35</td>
<td>25</td>
<td>29</td>
<td>25</td>
<td>25</td>
<td>19</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>30</td>
<td>3</td>
<td>27</td>
<td>28</td>
<td>3</td>
<td>18</td>
<td>21</td>
<td></td>
</tr>
</tbody>
</table>