8-1988

Assessing Writing Skills of Grades 3-6 in the Francis Howell School District

Ronald D. Brewer

ASSESSING WRITING SKILLS OF GRADES 3-6 IN THE FRANCIS HOWELL SCHOOL DISTRICT

BY

RONALD D. BREWER

Submitted in partial fulfillment of the requirements
for the Master of Arts in Education degree
Lindenwood College
August 8, 1988

Accepted by the faculty of the Department of Education, Lindenwood College, in partial fulfillment of the requirements for the Master of Arts in Education degree.

Gene Henderson
Advisor

Diane Denney
Reader

ABSTRACT

This project was an exercise in assessment of student writing in the Francis Howell School District. This district began a writing program in the fall of 1986, emphasizing 45 minutes of daily writing instruction and practice each school day. After approximately 44 weeks of exposure to the program, the assessment was conducted in the winter of 1988.

A published narrative essay test, the CAT Writing Assessment System, was given to 280 randomly selected students in grades three through six. Of these, 80 were used in the training of a rater team which then inter-scored the remaining 200 essays. They read each essay, gaining an overall impression of writing proficiency, and assigned a rank score of one (unacceptable) to four (good). These raters were guided in their scoring by the test publisher's provision of a scoring manual and anchor papers (sample essays chosen to illustrate the level of writing expected at each of the four categories, derived from experience in scoring tests from schools nation-wide). This type of grading is called holistic scoring.

The results of this writing assessment would mean little without the ability to compare them with either a pretest or the performance of a control group. Since

no pretest was systematically given before the writing program was begun, and all students in the district were treated in the writing program, it became necessary to create a theoretical control group, using the Kolmogorov-Smirnov one sample test. This is a statistical test of goodness of fit, which determines whether the scores in a sample can reasonably be thought to have come from a population having some specified theoretical distribution.

The 1986 achievement test stanines for language expression were recorded for each student in the sample, and used as an indicator of rank in written language proficiency. These rank-scores were compiled to develop a theoretical cumulative frequency distribution of student ability before treatment. That distribution was compared to the distribution of scores on the writing test using the K-S one sample test. Significant differences in distributions were found for all grades sampled except grade six (at the .05 level), indicating writing performances above expectations for grades three through five.

This project also demonstrated that holistic scoring can be done reliably and efficiently. This method of scoring was being practiced (to varying degrees) by most teachers in the district regularly in evaluating student writing; yet, without proper training, the majority found it laborious and

intimidating. A lack of district standards and a unified understanding of the relative importance to place on the various aspects of writing may have contributed to their insecurity with the system. The training methods and use of anchor papers described herein, along with the practice of inter-scoring essays for an averaged score, may provide a model for a more successful alternative in writing evaluation which could be used district-wide.

ACKNOWLEDGEMENTS

My most sincere thanks go to two people who helped make this project possible. Dr. Richard Schuppan freely gave guidance, inspiration, expertise and spiritual support in frequent times of need. Christine Wheeler provided the compositional and stenographic skills, not to mention time, needed to graphically present this thesis in a clear and attractive manner.

TABLE OF CONTENTS

CHAPTER I.

INTRODUCTION

Background

Donald Graves, one of the most noted and prolific pioneers in writing research, stated in 1978:

> Writing is the basic stuff of education. It has been sorely neglected in the schools. We have substituted the passive reception of information for the active expression of facts, ideas, and feelings. We now need to right the balance between sending and receiving. We need to let them write. (p.27)

Graves shared the sentiments of a diverse group of researchers, including Donald Murray (1968), Janet Emig (1971), and Lucy Calkins (1978), who have examined the writing process and how it is best taught in the schools. The general consensus of their research was that daily practice in actual writing improves student performance in not only composition skills, but all language arts skills, even reading. The act of expressing ideas on paper seems to contribute to one's ability to receive them in the same way.

In the last decade, much attention has been focused on incorporating the writing process into existing curricula. The most famous, perhaps, is the Bay Area Writing Project. Similar programs are the Vermont Writing Program and the New York State Board of Education Writing Program (Chew, 1982).

This growing enlightenment of the necessity and rewards of increased student involvement in the writing process led the Francis Howell School District, at the beginning of the 1986-1987 school year, to implement an approach to the teaching of writing based on the daily application of a writing process (prewriting, writing, rewriting, and publishing). Practice in both mechanical and expressive language skills based on textbook exercises was discouraged; instead, students would practice those skills by incorporating them in compositions, taking various forms (paragraphs, haikus, limericks, etc.) with a time allocation of 45 minutes daily.

Teachers were to informally evaluate these compositions. First, the application of whatever skill featured that day (e.g., proper use of capitalization) would be graded for a percentage-correct score. Beyond that, subjective evaluations would be periodically performed to determine a level of student performance in various forms of written expression.

The Problem

The adoption of this writing program was based on the belief that daily practice in writing would: 1) improve the student's ability to successfully apply various grammatical skills, 2) develop greater ability to express himself/herself in written form, and 3)

enhance learning of related language arts skills (e.g. spelling and vocabulary). In order to determine whether these goals were being met district-wide required a valid, reliable and comprehensive evaluation tool, uniformly administered and scored.

Traditional achievement tests, such as the CTBS, only indirectly evaluate the skills associated with writing. Individual, informal evaluations performed by various teachers lack any cumulative impact or meaning. A systematic approach to the assessment of writing skills was essential if the district was to determine the writing proficiency of students who had received the first year of writing instruction.

Research Design

It was the purpose of this study to examine student proficiency in narrative writing through direct assessment. A randomly selected sample of students in grades three through six were given appropriate levels of the CAT Writing Assessment System, a narrative writing test developed and published by CTB/McGraw-Hill, Incorporated. These narrative writing samples were then scored holistically by a trained team of Francis Howell classroom and special services teachers.

A true experimental design would require either a pretest or a control group to complement this writing

sample. However, no pretest was given, and all students in the district were treated in the writing program. Therefore, it became necessary to create a theoretical control group, using 1986 CTBS subtest scores in language expression and the Kolmogorov-Smirnov one-sample test.

The Kolmogorov-Smirnov one-sample test (Siegel, 1956) was appropriate for this type of study because it is a test of goodness of fit. It determines whether the scores in a sample can reasonably be thought to have come from a population having some specified theoretical distribution. The test involved specifying the cumulative frequency distribution which would occur under the theoretical distribution (based on 1986 CTBS stanine scores) and comparing that with the sample's cumulative frequency distribution. The theoretical distribution represents what would be expected under the null hypothesis. The test assumes that the variable under consideration has a continuous distribution and that observations are measured in an ordinal scale.

Null Hypothesis

For all students at each grade level, there is no significant difference in the expected frequencies and observed frequencies of narrative writing proficiency. Let the significant level be equal to .05.

Goal Statement

The goal of this study was actually two-fold. First, it was hoped that this assessment of elementary student writing performance would provide valuable information to parents, teachers, administrators, and school board members. Such data is essential in making judgements about the success of the current writing program, as well as determining whether revisions may be in order.

A second need may have been addressed by this project, as well. The most-voiced concern of elementary teachers and principals at recent writing workshops was over the apparent lack of a reliable, efficient evaluation strategy. The assessment method used in this project was chosen because it seemed to meet those needs. The implementation of this direct assessment technique, when combined with training for reliable holistic scoring, may be a valid and workable model for use at building or department levels.

CHAPTER II.

REVIEW OF LITERATURE

Graves (1978) summed up one of the most fundamental problems faced by anyone involved in implementing a writing program:

> The current emphasis on testing and documentation of pupil progress makes writing a stumbling block. Writing resists quantitive testing. A sixth grade teacher says, "I know why writing isn't emphasized more; it can't be tested. We are so hung up on nationally normed tests that we ignore teaching those areas where it can't be done. How do you say, 'Susie has improved six months in the quality of her writing'? We test them to death in reading and math and do some assessing in language conventions, but that's all." (p.14)

The question isn't whether children can make progress in developing writing skills; of course they can, and usually do. The question seems to be how to best measure that progress. Should direct or indirect methods be used? Can objective, standardized, multiple-choice tests be valid, or must writing sample evaluations be used? If essay tests are used, how should they be scored; atomistically, looking at isolated characteristics, or holistically? How can highest reliability be assured in test administration and scoring?

Direct vs. Indirect Assessment

Indirect assessment involves the use of objective tests, such as achievement tests, to measure writing ability. Direct assessment requires the evaluation of actual student writing samples, such as essay tests. Arguing for indirect assessment, Noyes, Sale, and Stalnakes (1945) suggested that a student assessed using the essay format is in the position of a "gambler who risks all on a single throw of the dice" (p.9), while a multiple-choice test allows many throws.

Palmer (1961) supported the view that writing can be assessed with validity using objective tests. He investigated the development of the College Entrance Examination Board test in writing. He noted that the test began in 1904 as a largely essay exam, yet had evolved by 1961 into a largely objective test, due to problems with superficiality, time demands, scoring expense and unreliability when using the direct assessment format. Palmer conceded, however, that objective test items must be well written and clear, stressing sentence structure and proper word choice, rather than spelling or punctuation.

Wansor (1986) suggested that advocates of indirect assessment believe that objective tests are superior devices. Their opinion was that they are demonstrably more reliable, have predictive validity, and, importantly, are less expensive to administer and score

than direct assessments.

The question of cost is a serious one in most school districts, and a balance must often be struck between high test validity and reliability and the testing budget. Veal and Hudson (1983), in a study of Georgia high school testing procedures, compared direct and indirect writing assessment instruments from administration to scoring in terms of face validity, reliability, and per-pupil cost. Direct methods received the best overall score, with substantial reliability (.69 to .76) and surprisingly low cost (as little as $.39 per paper). Objective tests had slightly higher reliability, questionable face validity, and averaged $.53 per student using computer scoring.

In his comparison of the advantages and disadvantages of both direct and indirect writing tests, Stiggins (1982) suggested that the major costs of direct assessment occurred in scoring, while those of indirect, objective tests occurred in test development; therefore, if reuse of the instrument is possible, objective tests would prove cheaper with repeated administration.

On the questions of validity, however, Stiggins agreed with Veal and Hudson that direct assessment best ensures face validity. Stiggins used the analogy that, just as a driver's license is obtained only by

performance of required skill rather than correctly identifying examples of good driving, a student's writing skill can only be truly measured when it is performed and observed. Stiggins found a major disadvantage of objective tests in their "lack of fidelity to real world writing tasks" (p.111).

Reliability seems to be a strength for objective tests, yet a weakness for direct assessment. Akeju (1972) analyzed the holistic scoring of seven trained examiners who independently scored 96 essay tests for a high school in Ghana, West Africa. The scorers were trained by the West Africa Examination Council to encourage uniform marking and therefore higher inter-rater reliability. Akeju discovered that intercorrelations of the readers' markings ranged from .51 to .76, with an average of .69. This was below an acceptable level of .80 set by the W.A.E.C.. On the other hand, a similarly used multiple-choice test also developed by the W.A.E.C. reported test reliability ranging from .83 to .89 (these had not been used expressly for measuring writing ability, however). Akeju concluded that scorers were using different standards of measurement, causing reliability to suffer. The objective test format, according to Akeju, was more highly reliable and therefore, perhaps, more desireable as a future choice for writing assessment.

There do appear to be moderate correlations between direct and indirect test results. Breland, Conlon, and Rogosa (1976) found a correlation of .42 for a group of 96 college freshmen between a 20 minute essay test and the College Board Test of Written English (TSWE), a 50 item multiple-choice test.

Moss, Cole, and Khampalikit (1982) studied the performance of 84 students from three rural schools in Ohio at grades 4, 7 and 10. An objective language test was administered to each, along with two essay tests which were scored both holistically (one overall grade) and atomistically (a series of subscores for each aspect of writing measured - spelling, puctuation, syntax, etc.). Corrected correlations varied by grade level (grade 4, from .20 to .68; grade 7, from .60 to .62; grade 10, from .72 to .76). The authors concluded that correlations between tests drop as grade level drops. They suggested that, early in a student's development, writing skills may be less uniform, more fragmented than in later years, accounting to some degree for their result. As correlations drop, the question of validity once again becomes more important, since the various assessment forms are apparently not measuring the same skill to the same degree.

The choice, then, between direct and indirect assessment methods centers around two basic variables, reliability and validity. Direct assessment using

actual writing samples offers more obvious face validity, yet reliability in scoring seems difficult to achieve. Indirect tests offer proven reliability, but questionable validity. Of the two, validity is a paramount virtue in any test.

A prudent path to follow, it seems, would be to choose direct assessment, with sufficient provisions made to provide acceptable reliability. Given the apparently subjective nature of essay evaluation, this may seem an arduous, even perilous, task. Yet Cooper (1975) considered this challenge inevitable and quite necessary. He stated:

> We might as well give up wishing for an easily scored multiple-choice growth measure for writing ability, and we must resist those who want to push off on us one of the currently available ones. ...Only a human reader can make the delicate, complex, multivaried decisions about quality or ideas or coherence in a piece of writing. Only a human reader can sense a young writer's growth toward finding an authentic voice for himself. (p. 117)

Improving Reliability in Direct Assessment

In investigating reliability problems with direct assessment, French (1962) classified the sources of error as (1)"student error", (2)"task error", (3)"scale error", and (4)"reader disagreement"(p.8). Braddock, et al. (1963) used the terms (1)"writer variable", (2)"assignment variable", (3) "rater variable", and (4)"inter-rater variable"(p.2). The following is an

examination of these four variables in terms of what steps are necessary to minimize such variations.

Writer variables.

McColly (1970) stated that, while there were no research findings to support it, his position was simply common sense that more than one writing sample should be taken, at different times, to avoid accidently choosing an "off day" for the students involved. Braddock (1963) cited such possible interferences as illness, lawnmower noise outside the window, or family problems, any of which could temporarily cause the student to write below his capacity.

Most researchers agree that two writing samples should be assessed, rather than just one. Diedrich (1974) stated that, in pursuit of reliability, two essays are required from each student with some separation in time as well as topic. Cooper (1977) concurred, stating that to achieve reliable scores requires more than one piece of writing from each student, with two or more scorers per essay.

The assignment variable.

Cooper (1977) suggested that topic choice should

"stimulate students to write as well as they can within the narrowly defined kind of discourse to be examined"(p.41). The exercise must be within the legitimate range of knowledge of the student. Cooper preferred a wide-open subject, which allowed every writer to find a uniquely personal way to respond without confusion or frustration.

McColly (1970) agreed, stating that the choice of a writing test topic should provide students with an opportunity to have something to say; to "filter out" as much as possible, the problem of subject matter mastery, which would test a different objective than pure writing ability.

McColly suggested that meeting the afore-mentioned conditions would lead to a valid topic choice, and that if such a topic is used, only one writing sample would be required (provided other sources of error are removed or controlled). He stated:

> The matter of topic validity is perhaps more essential than any other aspect of essay testing in relation to research, since it would seem that the matter is amenable to experimental investigation, and so far there has been none conducted. (p.152)

Rater variables.

This aspect of possible error refers to the people in charge of reading and scoring the writing sample. They are referred to interchangeably as raters,

readers, scorers, and graders. Their task is to (a) have in mind a standard or set of criteria to judge each essay against and (b) assign a score to each essay based on how well it matches that standard.

Essays may be read and scored in various ways. The two most common are "holistic" and "atomistic". Holistic scoring, also called "general impression marking", refers to ranking a writing sample on the basis of some overall scale, usually running numerically 1-3, 1-4, 1-9, and so on, with the higher score indicating superior work. Charney (1984) indicated that there is no evidence in the research that one numerical range scale is any more reliable than another. Holistic scoring requires looking at several facets of writing ability (syntax, grammar, paragraph structure, etc.), weighing the relative strengths and weaknesses of each, and assigning a single score.

Atomistic scoring, also referred to as "analytic" scoring, is similar in that the reader again uses a numerical scale to rate the various aspects of writing; however, each category gets its own score. A series of subscores is then generated for each sample.

Many writing evaluation experts, including Diedrich (1974), Cooper (1975), Charney (1984), and Coffman (1971), recognized holistic scoring as superior. Two reasons for preferring holistic scoring were (a) it is

much faster and more efficient, and (b) it pulls all aspects of student writing ability together, whereas atomistic scoring is more complicated, time-consuming, and fractured in its assessment.

Regardless of scoring method used, essay scoring remains a qualitative act, and as such, causes problems with intra-rater and inter-rater reliability. If many raters score many essays, will all raters use the same criteria? Will each scorer have a fixed criteria, or will his/her judgements waver from essay to essay?

Diedrich (1974) reported on a study investigating scorer consistency. Sixty readers from six distinguished but different backgrounds (lawyers, English teachers, businessmen, etc.) rated 300 college freshmen essays. Correlations ran from .22 to .41, with English teachers earning the highest correlations. The results revealed the differences of opinion that result from uncontrolled grading, when each subgroup of reader is looking for different things in each essay (one gives more weight to spelling, while another emphasizes word usage).

Even the low correlations among English teachers were not surprising to Diedrich. He stated: "The average English teacher is capricious in judgement, full of prejudices that have no basis in anyone's system of grammar, rhetoric, or style" (1957, p.8).

Similar studies on the reliability of essay grading

(Cooper, 1977, Godshalk, et al., 1977, Odell & Cooper, 1980) concluded that highest reliability can be obtained if graders are first trained by studying descriptions of each grade on the holistic scale used, then using that scale to practice on sample essays.

There are many writing tests which have incorporated clearly defined scales, with examples of typical essays which would earn each score on the scale (often called "anchor papers"). Education Testing Service (ETS) and the College Entrance Examination Board (Diedrich, 1974) have used a scale of one to four and require self-training of scorers who must practice by grading, then trading and regrading, large numbers of essays in order to become "calibrated" to reach consensus. Reliability levels well over .80 have been achieved. The New York State Board of Education (Chew, 1982) commissioned Donald Graves in 1979 to develop a writing test for grades four to six, using two samples from each student and scoring on a predefined scale of one to four, with reliability running around .80. Each of these tests required varying degrees of training and practice by graders before essays were scored.

Training and the use of a clearly defined rating scale can dramatically increase reliability levels in scoring essays. Coffman (1971) trained four raters using a typical holistic scale, and had them score pairs of essays. Then, several weeks later, without

reminding the readers of their first scores, Coffman had the raters reread the same essays and rescore them, with a resulting intra-rater reliability rating of .87.

The importance of training and experience was underscored by the work of Sweedler-Brown (1985), who compared the scoring of well-trained and experienced "trainers" (those who train essay readers) and the scoring of regular readers, who had less experience and training. After analyzing the inter-scorer correlations of 36 college essays in which disagreements over scores required the scoring of both readers and trainers, it was discovered that trainers, using both holistic and atomistic scales to rate the essays, earned correlations between scales of .790, while readers' scores resulted in correlations of .566 overall.

Inter-rater disagreement.

In the quest for improved realiability, it goes without saying that the score of any essay would be more meaningful if not just one, but a least two, scorers independently scored it with similar results using the same scale. This procedure is so common that every direct assessment program researcher mentioned thus far endorsed or required the two-scorer format when assessing writing of student populations beyond classroom size.When two readers score within one point

of another, the procedure recommended by researchers like Coffman (1971) and Diedrich (1974) was to average the two for a final score.

What should be done when scorers disagree by more than one point? Such a discrepancy may indicate misunderstandings on the part of one of the scorers regarding criteria, so arbitration may be required. Godshalk (1966) found that in these situations reliability would be enhanced if a more experienced reader (perhaps a trainer) gave the disputed essay a third reading and assigned an "arbitration" score that cancelled out the previous score which was furthest from his/hers. Experience and extra training tend to increase expected reliability (Sweedler-Brown, 1985). Such strategies are used in the Writing Test for New York Elementary Schools (Chew, 1982) and the California Achievement Test of Writing Assessment developed by C.T.B./McGraw-Hill (1986).

Summary

In selecting the best method of assessing writing ability, test validity and reliability are major considerations. Indirect, objective tests have proven reliability, and predictive validity, yet lack face validity. Without this demonstrated validity, objective tests are less desirable choices for assessment.

Direct assessment using essays or other writing samples represents the most valid method of writing evaluation. The major disadvantage, due to the subjective nature of scoring essays, is a lack of reliability. This reliability problem can be dealt with through clearly defined rating scales, reader training and practice, and the availability of trainers to resolve scoring discrepancies.

CHAPTER III.

METHODOLOGY

Subjects

All elementary teachers in the Francis Howell School District were trained to implement the new writing program during workshops scheduled throughout the first quarter of the 1986-87 school year. The treatment (daily writing instruction) was implemented by the start of the second quarter of that year.

Given the slowly evolving nature of writing skills acquisition, it was determined that assessment of the students' writing proficiency should not occur until late in the first semester of the 1987-88 school year. Students in grades one and two were not included in this study because there were no expected proficiencies in composition at these grade levels.

Of the approximately 3,080 students in grades three through six in Francis Howell School District, 280 were randomly selected by computer without replacement. The only limitation placed on selection was that the sample come from students enrolled by the start of 1986-87 school year, when the writing program was initiated. The sample was stratified in the following ways: 1) each of the four grade levels was represented equally at 70 students per grade, and 2) the sample

included two to three students from each classroom (depending on class size) in the district, ensuring representation from all schools and learning environments involved.

A narrative writing sample was obtained from each student selected. Since the Francis Howell School District uses a "cycled" year-round enrollment arrangement which causes five staggered attendance calendars, multiple writing assessment dates were selected in order to equalize the treatment period. It was calculated that, regardless of cycle, each student should have received approximately 44 weeks of writing instruction prior to assessment.

Of the 280 students who submitted writing samples, it was later determined through teacher surveys that five should be excluded from the study since they were assigned to self-contained special service classrooms and had never received the treatment. These five, along with 75 other randomly chosen writing samples, were removed from the original 280 to be used for practice scoring during the later training of the scorer team. This reduced the actual subjects under study to 200 (50 per each of the grade levels three through six), yielding a confidence interval of $\pm$ 6.76.

Materials

To obtain student writing samples, it was

determined that highest validity would be achieved by using the students' own classroom environments for testing, under the direction of their classroom teacher. To accomplish this, each teacher was provided the materials and directions necessary to administer the California Achievement Test: Writing Assessment System (CTB/McGraw-Hill, 1986), Narrative. The following levels were used:

Grade	Level	Grade range
3	13	2.6-4.2
4	14	3.6-5.2
5	15	4.6-6.2
6	16	5.6-7.2

This test was developed in 1985 under the leadership of Dr. Barbara Cole, now with the Oregon State Department of Education, at the CTB/McGraw-Hill test development facility in Monterey, California. In a telephone interview, she provided information on the process used to norm the test.

According to Dr. Cole, 120 test prompts (essay questions) were originally piloted locally. Of these, 42 were tested in a national sampling involving school districts in 11 states, with over 200 student responses obtained for each prompt.

All writing samples were scored by trained evaluators at the company's Composition Evaluation

Center. Twice as many prompts as were used in the final instrument were tried out. Only those with the highest proven reliability measures were chosen.

Four different types of writing can be assessed with this test: descriptive, narrative, expository, and persuasive. For the purposes of this study, the narrative prompts ("tell a story about...") were used, since this was the type of writing most practiced in the writing program at all grade levels, and therefore should have possessed greatest face validity.

To promote reliability, each teacher administered the test using the following:

1. Prescribed directions (e.g. Today we are going to take a writing test. I am going to give you a writing book with a number on it. Do not open your book until I tell you to do so.).
2. A writing book for the student which contained a narrative writing assignment (e.g. for grade 3, "Imagine that a dinosaur came to your school. Write a story about what happened.") and lined space for writing.
3. Writing time of 25 minutes with an indication from the teacher when five minutes remained for writing.

Scoring

Each student's writing sample was holistically scored by two trained evaluators. The evaluators were

classroom teachers who had served on the district's Writing Curriculum Committee and were trained by this study's author based on guidelines for training prescribed by the CAT Writing Assessment System. The teachers were provided with released time at district expense to perform the assessment (see Appendix A).

The author also directed the scoring sessions which involved the following procedures:

1. The evaluator impressionistically read the sample in one sitting.
2. The writing sample was rated good (4), acceptable (3), below average (2), or unacceptable (1), based on guidelines for each category (CTB/McCraw-Hill, Writing Assessment Guide, p.10):

> 4 The paper creates and develops a structured series of events. The sequence of events is clear, and the events are connected by effective transitional signals. A consistent story line is established. The writer uses various narrative reatures to give variety to details. The mechanics (including sentence structure and punctuation, precise word choice, precise word use, spelling and capitalization) are basically sound, and communication is clear.
>
> 3 The paper creates a series of events. The story is clear but not always focused on a unified event. There may be digressions. Occasional transitional signals create coherence and flow. The details lack variety but are basically appropriate. Mechanical errors do not influence communication.
>
> 2 The paper may be a list of events generated from the topic. The sequence is confusing. The use of details is not controlled. The use or lack of transitional signals often impedes the flow or undermines the coherence.

Mechanical errors negatively affect communication.

1 The paper is not focused on a single event. Details are extraneous or missing, and there is not control of narrative featurs. Transitional signals are not used to create flow or coherence. Mechanical errors greatly interfere with communication.

An anchor paper was provided for each category which was selected from student papers in a nationwide sample. For example, the anchor paper for narrative writing, Grade Three, category four (good):

> One day I went to school and saw that there was a dinosaur there. It was huge! It could hardly fit in the school. I thought it was funny. It broke my desk! He was almost bigger than the school! My teacher got mad at him. He got his name on the baord with five checks.

3. A second evaluator read and rated the writing sample, independent of the first evaluator's rating.

4. If the two ratings were the same, the final score matched them. If the two ratings were one point apart, they were averaged for the final score. If the two ratings were more than one point apart, a third reader (a trainer) evaluated the paper, unaware of the two discrepant ratings and assigned a final score (e.g. the first evaluator, two, second evaluator, four, final score anywhere from two to four, as determined by the third reader.).

Rater Training

The CAT/McGraw-Hill Writing Assessment System

provided a manual which detailed the process to be followed for training raters. There were three goals:

1. To develop familiarity with the rating scale and published anchor papers at each test level.
2. To practice ranking, then interscoring essays in comparison to the anchor papers.
3. To develop group uniformity in the understanding and application of scoring criteria in order to achieve acceptable reliability levels.

The training process took slightly over half a day, with raters actually taking the test and scoring each other's work. Next, scoring criteria and the rating scale were introduced and examined. The anchor papers were then discussed, along with their role as models for each rank on the scoring scale, derived through scoring samples nation-wide. Following that, each level's scoring team began ranking, trading, then interscoring "range-finders" (student essays drawn from the original sample to be used for practice in evaluation). A total of 20 range-finders were used in the practice session at each of the four levels. Interscoring and discussion of discrepancies gradually built a stable consensus among the raters.

CTB/McGraw-Hill offered reliability figures to be expected upon completion of the training procedure. They used Pearson correlations between the ratings of pairs of sample raters before a third evaluation was

involved. For narrative writing, the overall rating for a perfect match in scores for levels 13-16 typically range from .60-.68. However, differences of one point or less should occur from 97.1% to 99.9% of the time. A third (arbitration) evaluation should be required for no more than 1.5% of the essays scored.

Design

All of the elementary students in Francis Howell School District have received the treatment in question (writing instruction), and no pretest was systematically administered before the treatment began. These factors made the implementation of a true experimental design to test the effectiveness of the new writing program impossible without a control group to use as a basis for comparison. Such a control group was hypothetically constructed using the Kolmogorov-Smirnov one-sample test.

The primary author of the test, Andrey Kolmogorov (1941), is considered to be one of the 20th century's most influential Soviet mathematicians. He specialized in probability theory based on measure theory. He was prolific in the development of many statistical tests, among them the Kolmogorov-Smirnov (K-S) one-sample test.

As described in Chapter I, this is a test of goodness of fit, which compares the cumulative

distribution of a set of sample values (observed scores) with some specified theoretical distribution. The test was conducted by first specifying the cumulative frequency distribution which would occur under the theoretical distribution (student achievement tests were used to compute this) and comparing that with the observed frequency distribution computed using writing assessment scores). The point of greatest divergence between these two distributions was then calculated. The size of that divergence determined whether any differences between the two distributions was likely on the basis of chance.

The K-S one-sample test has been used in many ways. Warner and Buford (1941) used the test to confirm their hypothesis that American black people seem to have an heirarchy of preferences regarding shades of skin color. Grandy and Stahmann (1974) compared parents' personality types (expressed through occupational choices) to off-springs' personality types (same method) to determine the amount of parental influence parents exert on personality development. They used the K-S one-sample test to determine whether the matched types were due to chance.

The previously described writing assessment instrument employed in this study used ordinal measurement to rank student performance on a scale of one to four (1, 1.5, 2, 2.5, etc.) against nationally

derived scoring models called anchor papers. In order to satisfy the needs of the K-S test and this experiment, it was necessary to build a theoretical cumulative distribution of a control group with identical characteristics which was not exposed to the writing program treatment. This cumulative frequency distribution would have to be ordinally measured and reflect a ranking of expressive language ability compared to national norms or standards.

The most readily available source of data which provided the necessary information while meeting the previously described requirements was the 1986 CTBS achievement subtest scores for expressive language for each student in the sample. These tests were administered by the district in the spring of 1986, while the writing program began in the fall of 1986.

The 1986 CTBS expressive language subtest indirectly measured the student's ability to express himself/herself in written language, and assigned each a nationally normed score. An ordinal measurement of rank was assigned in the form of a stanine. These stanines could be used to develop a theoretical frequency distribution of student achievement in direct narrative writing assessment, suggesting where the student would be likely to rank on such a test without the treatment (exposure to the writing program).

There do appear to be positive correlations between

direct and indirect test results. Breland and Gaynor (1979) reported correlations for college freshmen of .63 using a comparison of the College Board test of written English and an essay test. On the elementary school level, Hogan and Mishler (1980) gave the Metropolitan Achievement Test-Language Instructional Test (MAT-LIT) and 20 essay tests to approximately 140 third graders and 160 eighth graders, with correlations of .68 and .65 respectively.

The major obstacle to overcome in building a theoretical distribution of direct writing assessment results using the 1986 CTBS stanines concerned the difference in the number of possible ranks in which a student could be placed. The direct writing instrument produced ranks of 1, 1.5, 2, 2.5, 3, 3.5, and 4, a total of seven categories. Of course, stanines provided nine. It was reasoned that the most equitable modification was to collapse stanines one and two, as well as eight and nine, creating a sta-seven rank order. The lowest and highest categories now comprised all who ranked in the top or bottom 11% of the national test population.

Using the K-S one-sample test and the CAT/McGraw-Hill Writing Assessment System instrument, the null hypothesis was tested.

Null Hypothesis

For all students at each grade level, there is no significant difference in the expected frequencies and observed frequencies of narrative writing proficiency. Let the significant level be equal to .05.

CHAPTER IV.

RESULTS

Narrative Writing Assessment

Following 44 weeks of daily writing practice and instruction, the sample of 200 students grades three through six were administered the CTB/McGraw-Hill narrative writing test. Those tests were separated by grade level and scored by a pair of trained raters who ranked, exchanged essays, then reranked each test sample according to nationally derived standards (anchor papers) on a scale of one (unacceptable) to four (good). If the two scores differed by one point, an average of the two was assigned. A two point or more disagreement required the regrading of the essay by the trainer and the assignment of an arbitration score. The results of this process are shown in Table 1.

Rater Agreement

Despite the fact that the scoring of the writing test involved ordinal measurement, the nature of the dual-rater format used herein made it possible to test for interrater reliability using the Pearson Product Moment Correlation test. This is the test of correlation predominantly used by researchers mentioned

Table 1

Narrative Writing Assessment Rank Scores

	Test levels			
Scale	13	14	15	16
4	20	20	22	20
3.5	10	10	10	9
3	13	14	9	15
2.5	4	4	4	2
2	2	0	4	3
1.5	1	0	0	1
1	0	0	0	0

thusfar in determining rater reliability. Pearson correlation coefficients for the four scorer teams were:

Grade:	Coefficient:
3	.688
4	.521
5	.694
6	.768

Grade four's scorer team had the lowest correlation, perhaps due to the lack of fourth grade teaching experience by either rater on the team (one was an Learning Disabilities specialist, while the other had only taught third grade). A team more familiar with typical fourth grade student writing performance and expectations might have had a higher

correlation. All other teams were composed of teachers practicing at the grade level to which they were assigned, and, therefore, more familiar with evaluating the level of writing proficiency which might be encountered (see Appendix B).

The test publisher provided Pearson correlation coefficients for holistic scoring, based on the performance of their trained evaluators. Their correlations ranged from .60 to .77, suggesting that the reliability levels of the Francis Howell scorer teams were largely comparable to those achieved by professional raters.

The value of these results was dependent on the reliability of the raters who determined the scores. Perfect agreement in assigned scores occurred in 72% of the sample. One point or less disagreements occurred 99% of the time. Two point disagreements, requiring a third party arbitration score, occurred in only two cases, or 1% of the total scored. These statistics indicated rater performance at or above CTB/McGraw-Hill expectations, and suggested that the essay test scores have reliability levels at or above the publisher's anticipation.

Achievement Tests

Once the direct writing assessment was completed, the existing test records of the students in the sample

were examined in order to find expressive language achievement test scores from the spring of 1986 to use in implementing the Kolmogorov-Smirnov one-sample test. Upon investigation, it was determined that, of the 200 students in the sample, three had no achievement test records for 1986; one in fifth grade and two in fourth. These students' essay test scores were then removed from the sample, leaving 197. Four others in the sample had no CTBS scores, but were given the Iowa Test of Basic Skills in the spring of 1986. Since that test had a Language Usage subtest which was similar in design and method of measurement to the CTBS, the stanine scores from those tests were included as indicators of those students national rank in expressive language skills prior to the fall of 1986. The distribution of the sample's national stanine ranks in expressive language skills prior to the start of the writing program appears in Table 2.

According to Siegel (1956), "the statistic most appropriate for describing the central tendency of scores in an ordinal scale is the median" (p.25). Since both tests used in this study provide ordinal data, median scores for each were determined. Since both test scores have been calculated in seven categories of rank, a comparison is possible if each scale is renamed by category:

Rank Category	1	2	3	4	5	6	7
Essay test	1	1.5	2	2.5	3	3.5	4
CTBS stanine	1-2	3	4	5	6	7	8-9

Table 2

1986 CTBS Language Expressing Stanines

	Grade			
Stanine	3	4	5	6
8-9	19	19	13	15
7	2	0	1	9
6	12	13	6	14
5	10	6	12	4
4	6	9	14	4
3	1	0	2	2
1-2	0	1	1	2

Thus, a student who had a stanine of five on the 1986 CTBS and a student who received a 2.5 on the essay test both fall in category four, or middle in rank when compared to national norms (CTBS) or a nationally derived and applied ranking scale (narrative essay test). The median score for the narrative writing assessment was in a higher category at all grade levels than for the CTBS, as can be seen in Table 3.

Table 3

Comparison of Test Medians

Grade	CTBS Stanine	Essay Score
3	6 (5)	3.5 (6)
4	6 (5)	3.5 (6)
5	5 (4)	3.5 (6)
6	6 (5)	3.5 (6)
Overall	6 (5)	3.5 (6)

Note - The numbers in parentheses indicate rank categories (out of seven possible) in which each median falls.

Further comparison of the numbers of students in each rank category revealed similarities between the two test results. By far the most students ranked in category seven on the CTBS, followed in number by category five. This was also true for the narrative essay test. Conversely, the fewest students ranked in category one on both tests, with category two consistently being second from the bottom in numbers of students ranked there.

Kolmogorov-Smirnov One-Sample Test Implementation

In order to employ the K-S one-sample test, these stanines were used to establish a theoretical seven-category cumulative frequency distribution for

each grade level. This distribution reflected how many of the sample would theoretically fall in each rank-order category on a given written language expression test if nationally normed or scored according to national standards, such as CTB/McGraw-Hill does through its development and use of anchor papers.

Once a theoretical control group was established in this way, the observed scores of the writing assessment test were transformed into an observed cumulative frequency distribution and compared to the previously computed theoretical distribution. The difference between the two was determined at each category for each grade level. The largest difference for each grade level was underlined and designated the maximum deviation for that test. Given the student population size of the Francis Howell School District and the size of the sample, a maximum deviation of .192 or higher would indicate significance at the .05 level. The results are presented in Table 5.

It can be seen that, at each grade level, the maximum deviation marked the point where the sample's scores were most skewed in their distribution. In each case, the observed scores' cumulative frequency distribution were skewed in bulk to the right of the maximum deviation point, or higher in rank than one would have expected, given the theoretical cumulative frequency distribution based on previous achievement

test performance. All grade levels' rank-scores were skewed above the maximum deviation point, yet only in grades three, four, and five were those differences considered significant.

Summary

As shown, in all grades except grade six, there was a significant difference between expected and observed frequency distributions of student expressive written language ability. Thus, test results suggest that in grades three, four, and five, the null hypothesis that there is no significant difference between expected and observed frequencies of narrative writing proficiencies must be rejected. Only in grade six did test evidence support the null hypothesis.

CHAPTER V.

DISCUSSION

Test Results

Since the fall of 1987, the elementary teachers of the Francis Howell School District have remained enthusiastic in their interest in and teaching of the daily writing program. The general concensus is that children do seem to understand and apply concepts involved in developing written language skills more efficiently and completely when those skills are practiced daily in actual student writing. The results of this study might reinforce, even increase, that enthusiasm. With a median of 3.5 on the CAT writing test and a significant difference in cumulative frequency distributions between 1986 language achievement and 1988 essay test scores in three out of four grade levels sampled as determined by the K-S one sample test, it may appear at first glance that the students seem to have, indeed, improved in their ability to express themselves in writing.

However, several factors must be weighed in interpreting the results of this study. The most basic question concerns how closely the CTBS and CAT tests matched in terms of the skills actually tested. The description of both tests and the skills examined are

Table 4

Kolmogorov-Smirnov One-Sample Test

	Grade 3						
Test scale	1	1.5	2	2.5	3	3.5	4
No. essay scores	0	1	2	4	13	10	20
Theo. cum. freq.	0	.020	.140	.340	.580	.620	1.000
Obs. cum. freq.	0	.020	.060	.140	.400	.600	1.000
Max. deviation	0	0	.080	.200	.180	.020	0

	Grade 4						
Test scale	1	1.5	2	2.5	3	3.5	4
No. essay scores	0	0	0	4	14	10	20
Theo. cum. freq.	.021	.021	.208	.333	.604	.604	1.000
Obs. cum. freq.	.000	.000	.000	.083	.375	.583	1.000
Max. deviation	.021	.021	.208	.250	.229	.021	0

	Grade 5						
Test scale	1	1.5	2	2.5	3	3.5	4
No. essay scores	0	0	4	4	9	10	22
Theo. cum. freq.	.020	.061	.347	.592	.715	.734	1.000
Obs. cum. freq.	.000	.000	.082	.163	.347	.551	1.000
Max. deviation	.020	.061	.255	.429	.367	.183	0

	Grade 6						
Test scale	1	1.5	2	2.5	3	3.5	4
No. essay scores	0	1	3	2	15	9	20
Theo. cum. freq.	.040	.080	.160	.240	.520	.700	1.000
Obs. cum. freq.	.000	.020	.080	.120	.420	.600	1.000
Max. deviation	.040	.060	.280	.120	.100	.100	0

Note: Maximum deviation of .192 or more indicates a significant difference at .05 level.

quite similar. The CTBS language subtest was designed to measure a student's ability to express himself/herself using written language (subject-verb agreement, coherent and complete sentences, proper sequence, syntax, etc.). The narrative version of the CAT essay test examined largely the same skills in its description of holistic scoring. Despite the fact that the students were telling a story, the raters were cautioned not to allow themselves to be swayed in their scoring by the level of student creativity, penmanship, or tastefulness--none of these things should have been part of the scoring. Yet, despite the findings of other studies mentioned in Chapters I and II which suggested that there are positive correlations to varying degrees between direct and indirect assessment, the degree to which these tests actually agree in what they measure has not been quantitatively determined.

Another question concerns the degree of validity in comparing stanines on the CTBS and the rank-scores of the writing test. A justification for this comparison was offered in Chapter IV; however, even though both scores were derived from national ranking procedures (norming and anchor papers), there is again no statistical evidence that students who ranked in the fifth stanine on the CTBS were precisely equivalent in their mastery of writing skills to those who scored a 2.5 on the CAT. The CTBS stanines were simply used as

approximate indicators in placing each student in a theoretical frequency distribution of writing ability in comparison to a national sample, while the CAT essay test scores indicate where the students would score if their papers were ranked if scored with others nation-wide.

It should also be noted that because the reporting of all data was in the ordinal, rather than interval, scale, also implied was an underlying continuum of achievement on both the CTBS and writing test. Of the students in the fifth stanine on the CTBS, some scored almost at the bottom, while others were within a percentage point of scoring in the sixth stanine. If the essay tests which received a three were compared, it would be seen that some were barely good enough to be scored in that category, while others were almost as well done as some of the essays in category four. Some may see the use of ordinal data as an indicator of vagueness in this study's findings. However, in measuring something as multi-faceted as student writing, especially when employing the holistic scoring method, rank-order or ordinal measurement seemed to be the most reliabile scale to use and, therefore, was the choice of this study. Likewise, all direct writing assessment tests using holistic scoring mentioned in Chapter II generated ordinal data.

Holistic Scorer Performance

Perhaps the most rewarding aspect of this project was the success experienced in training the rater team and their subsequent performance. With the guidance of the CAT scorer training manual and the use of anchor papers, the speed and ease with which the eight teachers and the trainer gained predicted uniformity in rating (according to test publishers) was truly surprising. After approximately four hours of discussion, examination of anchor papers and scoring criteria, and rating practice involving 80 actual student essays (20 per each two-member team), the raters felt confident that they shared a common understanding of what indicators of student writing to look for, how much relative importance to give each, and what distractors to avoid (penmanship, creativity, etc.). The results reported in Chapter IV indicated that holistic scoring can be reliable and efficient (scoring time averaged one minute or less per essay).

Once the results were known, the most pleased and surprised of all at the performance of the rater team were the members of the team themselves. Most had come to the meeting reluctantly, lacking confidence in their ability to reliably and efficiently score so many essays. Afterward, all were enthusiastic about the prospect of developing a district-wide writing assessment program to ease the predominant problem of

assigning student quarterly grades in writing.

Summary

In Chapter I, two goals were expressed for this study. The first was that the writing assessment performed in this study would "provide valuable information to parents, teachers, administrators, and board members" about the state of student writing proficiency following 44 weeks of daily writing instruction. Given the limitations mentioned, the value of the data generated must be determined by the reader. The results may be, at the least, encouraging to those who have worked so diligently to make the writing program a success.

The second goal, the development of "a reliable, efficient evaluation strategy" for assessing student writing proficiency at the building or department level, was most surely achieved.

When this prospect was discussed with the rater team following the scoring session, they seemed sure similar success could be achieved at each grade level. They envisioned periodic schoolwide essay tests graded by teacher teams. Workshops initiated by the language arts committee and staffed by trainers from the original rater team could be provided to train all teachers in the district in holistic scoring. In each school, rating teams could be formed to score periodic

essay tests with reliability and efficiency, doing away with the need for the laborious and frustrating practice of individual teachers frequently scoring their own students' work throughout a quarter without another teacher's opinion, unaware of the criterion used by peers, and assigning a final grade based on questionable evidence.

If nothing else, this study has made some facts concerning the writing process and its assessment clearer. Writing skills are many and intertwined. A student's proficiency in writing is something complicated and resistent to definition or measurement. Progress in the development of writing skills is slow and equally elusive in one's attempt to assess it.

On the other hand, it was also demonstrated that direct assessment can be scored with acceptable reliability and efficiency. The largest problem voiced by those who teach daily writing can perhaps be better addressed: how does one give a grade on student writing with any sense of confidence, with reasonable expenditures of time and effort? The members of the rater team felt they have found a way. The development and acceptance of a similar assessment strategy district-wide with similar success would make the time and effort this project required well spent.

APPENDIX A

Essay Test Score Sheet

CAT Test Level: 13

Scorer: Grade Three

Date: 2-4-88

Scorer:	A	B	Av.
1.	4	3	3.5
2.	3	4	3.5
3.	4	4	4
4.	4	4	4
5.	3	3	3
6.	3	3	3
7.	3	2	2.5
8.	4	4	4
9.	3	3	3
10.	3	2	2.5
11.	3	4	3.5
12.	3	3	3
13.	2	2	2
14.	4	4	4
15.	4	4	4
16.	3	4	3.5
17.	3	3	3
18.	4	4	4
19.	4	3	3.5
20.	2	2	2
21.	3	3	3
22.	4	3	3.5
23.	3	3	3
24.	3	4	3.5
25.	4	4	4

Scorer:	A	B	Av.
26.	3	3	3
27.	4	4	4
28.	3	3	3
29.	3	4	3.5
30.	3	3	3
31.	3	2	2.5
32.	4	4	4
33.	3	2	2.5
34.	4	4	4
35.	4	4	4
36.	4	3	3.5
37.	1	2	1.5
38.	3	4	3.5
39.	3	3	3
40.	4	4	4
41.	4	4	4
42.	4	4	4
43.	3	3	3
44.	3	3	3
45.	4	4	4
46.	4	4	4
47.	4	4	4
48.	4	4	4
49.	4	4	4
50.	4	4	4

Key: 4 = good
3 = acceptable
2 = below average
1 = unacceptable

Essay Test Score Sheet

CAT Test Level: 14

Scorer: Grade Four

Date: 2-4-88

Scorer:	A	B	Av.
1.	4	3	3.5
2.	3	3	3
3.	3	3	3
4.	4	3	3.5
5.	3	3	3
6.	4	4	4
7.	4	4	4
8.	3	3	3
9.	2	4	3 Arbitration
10.	4	4	4
11.	3	2	2.5
12.	4	3	3.5
13.	3	3	3
14.	2	3	2.5
15.	3	3	3
16.	2	3	2.5
17.	4	4	4
18.	3	3	3
19.	4	4	4
20.	4	4	4
21.	4	4	4
22.	3	3	3
23.	4	4	4
24.	3	3	3
25.	3	3	3

Scorer:	A	B	Av.
26.	4	4	4
27.	3	4	3.5
28.	4	4	4
29.	3	3	3
30.	4	4	4
31.	3	4	3.5
32.	2	3	2.5
33.	3	4	3.5
34.	3	4	3.5
35.	4	4	4
36.	4	3	3.5
37.	4	4	4
38.	4	4	4
39.	4	3	3.5
40.	4	4	4
41.	4	3	3.5
42.	4	3	3.5
43.	4	3	3.5
44.	3	3	3
45.	4	4	4
46.	4	4	4
47.	4	4	4
48.	4	4	4
49.	4	4	4
50.	3	3	3

Key: 4 = good
3 = acceptable
2 = below average
1 = unacceptable

Essay Test Score Sheet

CAT Test Level: 15

Scorer: Grade Five

Date: 2-4-88

Scorer:	A	B	Av.
1.	4	4	4
2.	2	2	2
3.	2	3	2.5
4.	2	2	2
5.	3	3	3
6.	2	2	2
7.	3	3	3
8.	3	3	3
9.	3	4	3.5
10.	4	4	4
11.	4	4	4
12.	3	4	3.5
13.	3	3	3
14.	4	4	4
15.	4	4	4
16.	2	2	2
17.	2	3	2.5
18.	4	4	4
19.	4	3	3.5
20.	4	3	3.5
21.	4	4	4
22.	3	3	3
23.	3	2	2.5
24.	3	3	3
25.	4	4	4

Scorer:	A	B	Av.
26.	4	4	4
27.	4	4	4
28.	4	4	4
29.	4	4	4
30.	4	4	4
31.	4	4	4
32.	4	4	4
33.	4	4	4
34.	3	3	3
35.	4	3	3.5
36.	4	2	3 Arbitration
37.	3	4	3.5
38.	4	4	4
39.	3	3	3
40.	4	3	3.5
41.	4	4	4
42.	4	4	4
43.	4	4	4
44.	3	4	3.5
45.	3	4	3.5
46.	4	4	4
47.	4	3	3.5
48.	2	3	2.5
49.	4	4	4
50.	4	4	4

Key: 4 = good
3 = acceptable
2 = below average
1 = unacceptable

Essay Test Score Sheet

CAT Test Level: 16

Scorer: Grade Six

Date: 2-4-88

Scorer:	A	B	Av.
1.	4	4	4
2.	3	4	3.5
3.	3	3	3
4.	3	2	2.5
5.	3	3	3
6.	4	4	4
7.	2	1	1.5
8.	4	4	4
9.	4	4	4
10.	4	3	3.5
11.	3	4	3.5
12.	3	3	3
13.	3	3	3
14.	3	3	3
15.	4	3	3.5
16.	4	3	3.5
17.	2	2	2
18.	2	2	2
19.	4	4	4
20.	3	3	3
21.	3	3	3
22.	4	4	4
23.	4	4	4
24.	3	4	3.5
25.	3	3	3

Scorer:	A	B	Av.
26.	3	3	3
27.	4	4	4
28.	4	3	3.5
29.	4	4	4
30.	4	3	3.5
31.	4	4	4
32.	4	4	4
33.	3	2	2.5
34.	3	3	3
35.	3	3	3
36.	3	3	3
37.	3	3	3
38.	2	2	2
39.	4	4	4
40.	4	3	3.5
41.	4	4	4
42.	4	4	4
43.	4	4	4
44.	4	4	4
45.	4	4	4
46.	4	4	4
47.	4	4	4
48.	4	4	4
49.	3	3	3
50.	3	3	3

Key: 4 = good
3 = acceptable
2 = below average
1 = unacceptable

APPENDIX B

WRITING ASSESSMENT: PEARSON PRODUCT MOMENT CORRELATION

GRADE: 3

STUDENT	RATER A	RATER B	A SQUARED	B SQUARED	A * B
1	4	3	16	9	12
2	3	4	9	16	12
3	4	4	16	16	16
4	4	4	16	16	16
5	3	3	9	9	9
6	3	3	9	9	9
7	3	2	9	4	6
8	4	4	16	16	16
9	3	3	9	9	9
10	3	2	9	4	6
11	3	4	9	16	12
12	3	3	9	9	9
13	2	2	4	4	4
14	4	4	16	16	16
15	4	4	16	16	16
16	3	4	9	16	12
17	3	3	9	9	9
18	4	4	16	16	16
19	4	3	16	9	12
20	2	2	4	4	4
21	3	3	9	9	9
22	4	3	16	9	12
23	3	3	9	9	9
24	3	4	9	16	12
25	4	4	16	16	16
26	3	3	9	9	9
27	4	4	16	16	16
28	3	3	9	9	9
29	3	4	9	16	12
30	3	3	9	9	9
31	3	2	9	4	6
32	4	4	16	16	16
33	3	2	9	4	6
34	4	4	16	16	16
35	4	4	16	16	16
36	4	3	16	9	12
37	1	2	1	4	2
38	3	4	9	16	12
39	3	3	9	9	9
40	4	4	16	16	16
41	4	4	16	16	16
42	4	4	16	16	16
43	3	3	9	9	9
44	3	3	9	9	9
45	4	4	16	16	16
46	4	4	16	16	16
47	4	4	16	16	16
48	4	4	16	16	16
49	4	4	16	16	16
50	4	4	16	16	16
SUM	170	169	600	597	591

NUMERATOR (N*SUM/A*B)*(SUM/A*SUM/B)

29550 28730 820

DENOMINATOR

SR((N*SUM/A SQR)-(SUM/A)SQR)*SR((N*SUM/B SQR)-(SUM/B)SQR)

30000 28900 1100 29850 28561 1289

33.16625 35.90265 1190.756

CORRELATION COEFFICIENT 0.688637

WRITING ASSESSMENT: PEARSON PRODUCT MOMENT CORRELATION

GRADE: 4

STUDENT	RATER A	RATER B	A SQUARED	B SQUARED	A * B
1	4	3	16	9	12
2	3	3	9	9	9
3	3	3	9	9	9
4	4	3	16	9	12
5	3	3	9	9	9
6	4	4	16	16	16
7	4	4	16	16	16
8	3	3	9	9	9
9	2	3	4	9	6
10	4	4	16	16	16
11	3	2	9	4	6
12	4	3	16	9	12
13	3	3	9	9	9
14	2	3	4	9	6
15	3	3	9	9	9
16	2	3	4	9	6
17	4	4	16	16	16
18	3	3	9	9	9
19	4	4	16	16	16
20	4	4	16	16	16
21	4	4	16	16	16
22	3	3	9	9	9
23	4	4	16	16	16
24	3	3	9	9	9
25	3	3	9	9	9
26	4	4	16	16	16
27	3	4	9	16	12
28	4	4	16	16	16
29	3	3	9	9	9
30	4	4	16	16	16
31	3	4	9	16	12
32	2	3	4	9	6
33	3	4	9	16	12
34	3	4	9	16	12
35	4	4	16	16	16
36	4	3	16	9	12
37	4	4	16	16	16
38	4	4	16	16	16
39	4	3	16	9	12
40	4	4	16	16	16
41	4	3	16	9	12
42	4	3	16	9	12
43	4	3	16	9	12
44	3	3	9	9	9
45	4	4	16	16	16
46	4	4	16	16	16
47	4	4	16	16	16
48	4	4	16	16	16
49	4	4	16	16	16
50	3	3	9	9	9
SUM	174	173	626	613	611

NUMERATOR	(N*SUM/A*B)*(SUM/A*SUM/B)		
	30550	30102	448

DENOMINATOR					
SR((N*SUM/A SQR)-(SUM/A)SQR)*SR((N*SUM/B SQR)-(SUM/B)SQR)					
31300	30276	1024	30650	29929	721
	32			26.85144	859.2460
CORRELATION COEFFICIENT					0.521387

WRITING ASSESSMENT: PEARSON PRODUCT MOMENT CORRELATION

GRADE: 5

STUDENT	RATER A	RATER B	A SQUARED	B SQUARED	A * B
1	4	4	16	16	16
2	2	2	4	4	4
3	2	3	4	9	6
4	2	2	4	4	4
5	3	3	9	9	9
6	2	2	4	4	4
7	3	3	9	9	9
8	3	3	9	9	9
9	3	4	9	16	12
10	4	4	16	16	16
11	4	4	16	16	16
12	3	4	9	16	12
13	3	3	9	9	9
14	4	4	16	16	16
15	4	4	16	16	16
16	2	2	4	4	4
17	2	3	4	9	6
18	4	4	16	16	16
19	4	3	16	9	12
20	4	3	16	9	12
21	4	4	16	16	16
22	3	3	9	9	9
23	3	2	9	4	6
24	3	3	9	9	9
25	4	4	16	16	16
26	4	4	16	16	16
27	4	4	16	16	16
28	4	4	16	16	16
29	4	4	16	16	16
30	4	4	16	16	16
31	4	4	16	16	16
32	4	4	16	16	16
33	4	4	16	16	16
34	3	3	9	9	9
35	4	3	16	9	12
36	4	3	16	9	12
37	3	4	9	16	12
38	4	4	16	16	16
39	3	3	9	9	9
40	4	3	16	9	12
41	4	4	16	16	16
42	4	4	16	16	16
43	4	4	16	16	16
44	3	4	9	16	12
45	3	4	9	16	12
46	4	4	16	16	16
47	4	3	16	9	12
48	2	3	4	9	6
49	4	4	16	16	16
50	4	4	16	16	16
SUM	172	173	618	621	613

NUMERATOR (N*SUM/A*B)*(SUM/A*SUM/B)

30600 29756 844

DENOMINATOR SR((N*SUM/A SQR)-(SUM/A)SQR)*SR((N*SUM/B SQR)-(SUM/B)SQR)

30900 29584 1316 31030 29929 1121

36.27671 33 48134 1214.592

CORRELATION COEFFICIENT 0 694883

WRITING ASSESSMENT: PEARSON PRODUCT MOMENT CORRELATION

GRADE: 6

STUDENT	RATER A	RATER B	A SQUARED	B SQUARED	A * B
1	4	4	16	16	16
2	3	4	9	16	12
3	3	3	9	9	9
4	3	2	9	4	6
5	3	3	9	9	9
6	4	4	16	16	16
7	2	1	4	1	2
8	4	4	16	16	16
9	4	4	16	16	16
10	4	3	16	9	12
11	3	4	9	16	12
12	3	3	9	9	9
13	3	3	9	9	9
14	3	3	9	9	9
15	4	3	16	9	12
16	4	3	16	9	12
17	2	2	4	4	4
18	2	2	4	4	4
19	4	4	16	16	16
20	3	3	9	9	9
21	3	3	9	9	9
22	4	4	16	16	16
23	4	4	16	16	16
24	3	4	9	16	12
25	3	3	9	9	9
26	3	3	9	9	9
27	4	4	16	16	16
28	4	3	16	9	12
29	4	4	16	16	16
30	4	3	16	9	12
31	4	4	16	16	16
32	4	4	16	16	16
33	3	2	9	4	6
34	3	3	9	9	9
35	3	3	9	9	9
36	3	3	9	9	9
37	3	3	9	9	9
38	2	2	4	4	4
39	4	4	16	16	16
40	4	3	16	9	12
41	4	4	16	16	16
42	4	4	16	16	16
43	4	4	16	16	16
44	4	4	16	16	16
45	4	4	16	16	16
46	4	4	16	16	16
47	4	4	16	16	16
48	4	4	16	16	16
49	3	3	9	9	9
50	3	3	9	9	9
SUM	172	166	612	578	589

NUMERATOR	(N*SUM/A*B)*(SUM/A*SUM/B)			
	29450		28552	898

DENOMINATOR

SR((N*SUM/A SQR)-(SUM/A)SQR)*SR((N*SUM/B SQR)-(SUM/B)SQR)					
30600	29584	1016	28900	27556	1344
	31.87476			36.66061	1168.548

CORRELATION COEFFICIENT 0.768474

BIBLIOGRAPHY

Akeju, S. (1972). The reliability of general certificate of education examination English composition papers in West Africa. Journal of Educational Measurement, 9(2), 175-179.

Braddock, R., Lloyd-Jones, R., & Schoer, L. (1963). Research in written composition. Champaign, Il.: National Council of Teachers of English.

Breland, H., Conlon, G., & Rogosa, D. (1976). A preliminary study of the test of standard written English. Princeton, N.H.: Educational Testing Service.

Breland, H., & Gaynor, J. (1979). A comparison of direct and indirect assessment of writing skill. Journal of Educational Measurement, 16(2). 119-128.

C.A.T. Writing Assessment System. (1986). Monterey, Ca.: CTB/McGraw-Hill

Calkins, L. (1978). Children write--and their writing becomes their textbook. Language Arts, 55(7), 804-810.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. Research in the Teaching of English, 18(1), 65-80.

Chew, C. (1982). The writing test for new york state elementary schools: Development, form, implications. U.S. Dept. of Education, National Institute of Education.

Coffman, W. (1971). On the reliability of ratings of essay examinations in English. Research in the Teaching of English, 5(1), 24-35.

Cooper, C. (1975). Measuring growth in writing. English Journal, 64(3), 111-119.

Cooper, C. (1977). Evaluating writing: Describing, measuring, judging. Urbana, Il.: National Council of Teachers of English.

Diedrich, P. (1957). The problem of grading essays. Princeton, N.J.: Educational Testing Service.

Diedrich, P. (1974). Measuring growth in English. Urbana, IL.: National Council of Teachers of English.

Emig, J. (1971). The composing process of twelfth graders. Urbana, Il.: National Council of Teachers of English.

French, J. (1962). Schools of thought in judging excellence of school themes. Princeton, N.J.: Educational Testing Service.

Godshalk, F., Swineford, S., & Coffman, W. (1966). The measurement of writing ability. New York, N.Y.: The College Entrance Examination Board.

Grandy, T., & Stahmann, R. (1974). Types produce type: An examination of personality development using holland's theory. Journal of Vocational Behavior, 13(5), 231-239

Graves, D. (1978). Balance the basics: Let them write. New York, N.Y.: Ford Foundation.

Hogan, T., & Mishler, C. (1980). Relationships between essay tests and objective tests of language skills for elementary school students. Journal of Educational Measurement, 17(3), 219-227

Kolmogorov, A. (1941). Confidence limits for an unknown distribution function. Annals of Mathematical Statistics, 9(12), 261-463.

McColly, W. (1970). What does educational research say about the judging of writing ability? The Journal of Educational Research, 64(4), 148-156.

Moss, P., Cole, N., & Khampalikit, C. (1982). A comparison of procedures to assess written language skills at grades 4, 7, and 10. Journal of Educational Measurement, 19(1), 39-47.

Murray, D. (1968). A teacher teaches writing. Boston: Houghton-Mifflin Co.

Noyes, E., Sale, W., & Stalnaker, J. (1945). Report on the first 6 tests in English composition. New York: The College Board.

Odell, L., & Cooper, C. (1980). Procedures for evaluating writing: Assumptions and needed research. College English, 42(1), 35-42.

Palmer, O. (1961). Sense or nonsense? the objective testing of English composition. English Journal, 50(5), 314-320

Siegel, S. (1956). Non-parametric statistics for the behavioral sciences. New York: McGraw-Hill Book Co.

Stiggins, R. (1982). A comparison of direct and indirect writing assessment methods. Research in the Teaching of English, 16(2), 101-114.

Sweedler-Brown, C. (1985). The influence of training and experience on holistic essay evaluations. English Journal, 74(5), 49-55

Veal, L., & Hudson, S. (1983). Direct and indirect measures for large scale evaluation of writing. Research in the Teaching of English, 17(3), 290-296

Wansor, C. (1986). Assessing writing ability: What are the issues, approaches? NASSP Bulletin, 70(4), 67-73

Warner, W., & Buford, H. (1941). Color and human nature. Washington, D.C.: American Council on Education.