THE ABILITY OF MARYLAND ENGLISH TEACHERS TO RATE HOLISTICALLY THE QUALITY OF STUDENT EXPLANATORY WRITING

by

Ronald Aaron Peiffer

Dissertation submitted to the Faculty of the Graduate School
of the University of Maryland in partial fulfillment
of the requirements for the degree of
Doctor of Education
1988

Advisory Committee:

Professor Richard K. Jantz, Chairman/Advisor
Professor Jean R. Hebeler
Professor V. Phillips Weaver
Professor Robert M. Wilson
Associate Professor Wayne H. Slater

ABSTRACT

The Ability of Maryland English Teachers to Rate Holistically The Quality of Student Explanatory Writing

Ronald Aaron Peiffer

Doctor of Education, 1988

Directed by: Professor Richard K. Jantz, Department of Education

The purpose of this study was to determine the accuracy of Maryland English teachers in using the Maryland Writing Test scoring criteria to place modified holistic ratings on student explanatory writing. The performance of eight expert raters, who had previously demonstrated 80% rating accuracy in training, was compared with the performance of six novice raters, who had not been required to demonstrate accuracy in their training. Accuracy was determined by analyzing error frequency and patterns in error size and direction. Scores were further analyzed to determine writing features, both internal and external to the Maryland Writing Test scoring criteria, that served as predictors of scores assigned by the two groups of raters.

Findings indicate that novice and expert raters were approximately 60% accurate in score assignments, with no significant difference in the accuracy level of the two groups. While scores assigned by both groups correlated highly, the size of their errors correlated moderately. Novice rater errors were more often one or more score points below the certified scores that compositions should have received while expert rater errors were equally distributed between overassessments and underassessments of writing quality.

The results of stepwise regressions showed certified scores as well as scores assigned by the two groups of raters to be predicted by the number of words in the composition and by the frequency of syntax errors. While 39% of the variance in certified scores was explained by the number of words, around 50% of the variances in novice and expert scores were explained by the same feature. Likewise, syntax error frequencies were slightly stronger predictors of rater scores than of certified scores, contributing 11% and 17% respectively to the variance in expert and novice rater scores. Of five features associated with the scoring guide, content was the strongest predictor of certified scores, explaining 99.4% of the variance in scores. However, organization was the strongest predictor of rater scores, explaining around 80% of the variance in scores.

DEDICATION

For
My Wife and My Son,
Karen and Tom

ACKNOWLEDGMENTS

This work would was made possible by the cooperation of the Maryland State Department of Education and members of the Maryland educational community. In particular, I would like to express my gratitude to the following:

Charles Allen, Baltimore City Public Schools
Douglas Coulson, University of Maryland
Owen Crabb, Maryland State Department of Education
Steven Ferrara, Maryland State Department of Education
Barry Gelsinger, Carroll County Public Schools
Gail Goldberg, Maryland State Department of Education
Robert Gabrys, Maryland State Department of Education
Members of the Maryland Writing Test Scoring Committee
Barbara Reeves, Baltimore County Public Schools
Lillian Rogers, Carroll County Public Schools
Patricia Williams, Maryland State Department of Education

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION TO THE STUDY

Introduction

Holistic scoring of written compositions has been used nationally in large-scale assessment programs and has been refined by researchers such that inter-rater agreement on scores approach 90% in some cases (White, 1985, p. 180). Maryland has modified the holistic scoring procedure to include a set of generic rating criteria to guide trained raters in scoring the Maryland Writing Test and has approached the high levels of accuracy observed in other large-scale assessment programs (Maryland State Department of Education, 1986, 1987c). In classrooms across the state, those criteria guide teachers on a routine basis in holistically scoring student work, but the effectiveness of such published criteria in standardizing teacher assessment of writing is yet unclear.

The Maryland Functional Writing Program constitutes a unique effort to impose on English teachers a statewide definition for quality of writing. English teachers historically have been divided over the definition of writing quality as evidenced by the diversity in what teachers teach and how they grade writing (Hirsch, 1977). Each English teacher has acquired a collection of writing features for which he or she looks in student compositions. The importance of each feature becomes transcribed into a self-set scale, thus producing a set of personalized criteria for scoring and grading--a personal definition of writing quality. The impact of a statewide definition of writing quality on teachers' rating decisions and the extent to which the state definition agrees and conflicts with individual definitions of

quality has been explored only in terms of the number of students passing the Maryland Writing Test.

Purpose

The purpose of this study was to examine the extent to which Maryland English teachers could effectively transfer modified holistic scoring procedures from a large-scale assessment design to a simulated classroom situation. The primary focus of the study was the effectiveness of training on rater accuracy. The study provided evidence about how closely training should resemble the kind of preparation raters receive in large-scale assessment programs. The impact of training was determined by analyzing teacher rating decisions for the relationship between scores and composition features within and outside of the scoring criteria. The determination of those rating predictors complemented an analysis of rater accuracy to help estimate the quality of feedback teachers would likely give to students about their writing.

Rationale

Holistic scoring of writing quality is a controversial practice, at best. Critics point out that research has not provided support for claims of validity and reliability made by proponents of holistic scoring (Charney, 1984). In many academic areas, assessment of student achievement has traditionally included multiple choice instruments that have historically enjoyed acceptance from the public and from educators alike. Unfortunately, multiple choice tests of writing quality have many of the psychometric benefits of other multiple choice tests except for validity. Such instruments simply cannot measure how well students compose (Godshalk et al., 1966; Culpepper & Ramsdell, 1982; Huntley, Schmeiser,

& Stiggins, 1979; Breland & Gaynor, 1979). Meanwhile, holistic scoring of writing quality has undergone intense refinement by state and national assessment agencies and researchers (McCready & Melton, 1981; White, 1985). Consequently, some standard practices have emerged that have produced, for large-scale assessments, accurate, defensible measurements of writing quality (Cooper, 1977; Maryland State Department of Education, 1986, 1987c, 1988).

In an effort to examine both the rating behaviors of trained raters and the characteristics of quality writing, much research has focused on the lexical and syntactic patterns of compositions and how those patterns related to the quality ratings compositions received (Cooper, et al., 1984; Freedman, 1979a, 1979b, 1984; Grobe, 1981; Nielsen & Piche, 1981; Nold and Freedman, 1977; Schmeling, 1970). A limited amount of additional research has gone further to pursue the relationship of rhetorical features of writing to quality ratings (Freedman, 1979a, 1984). But even less has been done to connect the rater's own schema with text features that influence score decisions (Freedman & Calfee, 1983).

Though the list of possible writing features that might influence raters during holistic scoring is virtually endless, some have been more thoroughly studied than others. Sentence complexity, prime among them, has been widely studied following Kellog Hunt's (1965) initial work in the 1960's (Thomas & Donlan, 1980; Combs, 1976; O'Hare, 1973; Pedersen, 1977; Potter, 1967; Schmeling, 1970). While most researchers found sentence complexity to be characteristic of good writing, others found evidence that sentence complexity either was unrelated to writing quality or was related to poorer quality, often depending on the age of the writer

(Mellon, 1969; San Jose, 1972; Faigley, 1979; Hillocks, 1984, 1986, p. 75). Errors, on the other hand, have consistently related to writing quality, with spelling errors being among the most powerful indicators of poor writing (Baddely & Wing, 1980; Breland & Jones, 1984; Neilsen & Piche, 1981; Cooper, et al., 1984). Further, punctuation, capitalization, and usage mistakes have been nearly as frequent features of poor writing (Baddely & Wing, 1980; Bartholomae, 1980; Freedman, 1979a, 1979b; Gorrell, 1983; Williams, 1981; Cooper, et al., 1984).

Even composition length and vocabulary level have proven to be related to quality ratings (Grobe, 1981; Breland & Jones, 1984; Nold & Freedman, 1977). However, it is not clear if any such characteristics have influenced raters inappropriately or if they do indeed represent features of good writing. This is partly due to the absence of stated rating criteria in most research involving holistic scoring of compositions.

A considerable body of research on large scale assessment has helped to refine holistic scoring and to develop an understanding of its limitations. However, the classroom application of holistic scoring as an instructional tool has yet to reach the same level of refinement. In Maryland, where modified holistic scoring is used in a statewide assessment program, teachers are encouraged to implement holistic scoring into instruction. As such, modified holistic scoring requires raters to apply sets of generic narrative and explanatory rating criteria to Maryland Writing Test papers. Unlike other holistic rating procedures, these criteria can be applied to any narrative and explanatory composition, thus opening a unique opportunity for the researcher. Whereas prior research revealed only the appearance of text features in

relation to quality ratings, this study attempted to extend the line of inquiry to include the effects of rater training on scoring decisions.

Significance of the Study

Maryland classroom teachers are responsible for assessing student writing in preparation for the Maryland Writing Test in grade nine. Though at least part of the Maryland Writing Test scoring criteria are assumed to include a portion of all English teachers' personal writing criteria, teachers possess other personal criteria and place different levels of significance on them. In adapting Maryland's modified holistic scoring procedures to the classroom, the Maryland Writing Test scoring criteria form a critical link to the state test. The extent to which that link is obscured by a rater's personal writing assessment criteria or strengthened by training needs further definition.

A better understanding of the relationship of personal and Maryland Writing Test scoring criteria in assessment is needed to begin a search for intervention practices that would secure the accuracy of the classroom teacher's use of modified holistic scoring. The training procedures currently used in Maryland range from review of the scoring guides and annotations on scored papers to practice rating. This training is designed to assist the classroom teacher in applying the scoring criteria to student writing as a routine aspect of writing instruction. The use of rating scales and rubrics have proven a successful practice for improving writing performance in students (Hillocks, 1986, p. 156), and for Maryland English teachers, such practices have grown even more important. When Maryland English teachers use the state criteria to prepare students for the

Maryland Writing Test, accurate application of those criteria would be a necessity.

It is clear that the publication of the Maryland Writing Test scoring criteria and inservice activities associated with the Maryland Writing Test are intended to strengthen the rating accuracy of teachers. In effect, this means that such practices will need to narrow the focus of teachers who are scoring student papers to include a limited range of writing characteristics, at least at those times when teachers are preparing students to take the Maryland Writing Test. The degree to which teachers have assigned the same scores as trained raters indicates the ability of teachers to score accurately. However, teachers assigning differing scores may fully understand the criteria but may interpret the scale somewhat differently. Consequently, it was necessary to compare and contrast teacher scoring decisions in several ways to determine if rating differences were the result of disagreement with the Maryland Writing Test scoring criteria, imposition of other scoring criteria from the teachers' repertoire, or simply from rating more harshly or less harshly.

The relationship of the scores assigned by teachers and the characteristics of compositions would indicate the priorities of teachers when making rating decisions. Though the five features of the Maryland Writing Test scoring criteria should serve as the predictors of scores assigned by raters, other features, such as mechanical errors and handwriting, likely enter into rater decisions. The ability of training procedures to focus raters on the specified scoring criteria--in this case, the Maryland Writing Test explanatory scoring criteria--was determined by comparing the scores assigned by novice raters who were briefly trained

with scores assigned by expert raters who were extensively trained until they were able to assign accurate scores to 80% of the compositions they read. The degree to which accuracy declines after training was determined by comparing certified modified holistic scores with score assignments of expert raters, since they were trained several months prior to the scoring of papers in the study.

In considering activities or procedures to substitute for extensive rater training and monitoring for the classroom teacher, it has become important to determine if extensive training is needed to assure rater accuracy and if such training must occur immediately prior to scoring. It has also become critical to determine if any particular characteristics of writing such as mechanical errors and handwriting are more related to inaccurate score assignments than others and if the effects of training are more effective over time for some than for others.

The Problem

The study was designed to examine the ability of Maryland English teachers to perform modified holistic scoring in a simulated classroom assessment situation. Modified holistic scoring is a specialized skill for which Maryland teachers receive training. However, the effectiveness of such training activities in assuring rater accuracy was examined in terms of overall accuracy and, more specifically, in terms of the criteria with which teachers made score decisions. Both dimensions of scoring--accuracy and criteria--were examined in this inquiry in that both impact the kinds of feedback provided to students about their writing (Freedman, 1979a). In large-scale assessments, rater accuracy is developed through hours of training immediately prior to scoring . That training is focused on

helping each individual apply the same set of rating criteria to all compositions. In such assessment settings, raters can be trained to apply rating criteria with a high degree of success, but in the classroom, comparable training is not available and cannot be administered. Essentially, such training would be based on experts assigning correct scores to a set of sample papers from the collection of papers to be assessed (Maryland State Department of Education, 1988). Though such practices fortify the validity of large-scale assessment procedures, such authoritative intervention is not feasible for the classroom.

An analysis of teachers' modified holistic score assessments was necessary to determine the extent to which scoring inaccuracy is a problem for Maryland teachers. The results provide some insights into the level of training needed to assure rating accuracy and assist in identifying some specific aspects of writing that predict the kinds of scores raters assign.

Research Questions

1. What is the relationship between certified modified holistic scores, scores assigned to compositions by expert raters undergoing complete scoring training several months prior to scoring, and scores assigned to compositions by novice raters who have experienced only brief inservice programs in modified holistic scoring within the last year?

    a. To what extent do the scores assigned by the two groups correlate with each other and with the certified Maryland Writing Test scores?

    b. To what extent do the errors in the scores assigned by the two groups of raters correlate?

c. What is the mean difference in the number of accurate scores assigned by the two groups of raters?

d. What is the mean difference in the size of errors made by the two groups of raters?

e. What is the mean difference in the direction of errors made by the two groups of raters?

2. Which of the following characteristics or combination of characteristics of writing are predictors of the certified modified holistic scores, scores assigned to compositions by expert raters undergoing complete training several months prior to scoring, and scores assigned by novice raters who have experienced inservice programs in modified holistic scoring within the last year?

a. sentence complexity (as indicated by mean t-unit length)

b. number of words per composition

c. mean raters' assessments of composition handwriting quality

d. number of punctuation errors per 100 words

e. number of spelling errors per 100 words

f. number of capitalization errors per 100 words

g. language usage errors per 100 words

h. number of syntax errors per 100 words

3. Which of the following analytic ratings of writing features from the Maryland Writing Test scoring criteria are predictors of certified modified holistic scores, scores assigned by expert raters undergoing complete training several months prior to scoring, and scores assigned by novice raters undergoing inservice programs in modified holistic scoring within the last year?

a. organization

b. content

c. attention to audience

d. sentence formation

e. mechanical errors that interfere with meaning

Definition of Terms

Modified holistic scoring- the practice of assigning overall quality scores of 1 to 4 to writing on the basis of specific pre-set criteria and sets of anchor papers. In this study, the pre-set criteria used by raters were the Maryland Writing Test scoring criteria for explanatory writing (Maryland State Department of Education, 1987b, 1987d).

Maryland Writing Test scoring criteria- a set of general descriptions of compositions at each of the four score points stated in terms of content, organization, audience, sentence formation, and conventions errors. See Appendix A for a copy of the criteria for explanatory writing, which was used in the study.

Maryland Writing Test Scoring Committee- a group of twenty-four Maryland educators , working under the direction of assessment specialists from the Maryland State Department of Education to construct anchor papers and training materials for use by contracted raters in the scoring of the annual Maryland Writing Test.

Certified modified holistic scores- composition scores determined by the Maryland Writing Test Scoring Committee under the direction of assessment specialists from the Maryland State Department of Education. In this study, certified modified holistic scores were considered correct scores for all compositions.

Expert raters- eight members of the Maryland Writing Test Scoring Committee, all of whom were English teachers. As scoring committee members, they had undergone extensive scoring training in modified holistic scoring and had demonstrated 80% accuracy in scoring student compositions.

Novice raters- six Maryland English language arts teachers who had participated in at least one brief inservice experience reviewing the Maryland Writing Test criteria in the fall of 1987 as mandated by Maryland State Department of Education. These teachers had not demonstrated a pre-set level of scoring accuracy in their training.

Sentence complexity- the degree to which writing, on the sentence level, displays subordination of ideas and thoughts through the use of phrases and clauses. Generally, sentence complexity is measured by counting the mean number of words per t-unit for a composition.

Handwriting quality ratings- ratings of 1,2, or 3 assigned by members of both rater groups participating in the study. A score of 1 represented the lowest quality with 3 representing the highest quality. Ratings represented raters' perceptions of the quality of the handwriting of each composition in comparison with the handwriting of the other papers in the set that they holistically scored.

Analytic ratings of writing features- ratings of 1 to 4 assigned by a panel of three writing assessment specialists to each composition for each of the five features identified in the Maryland Writing Test explanatory scoring criteria--content, organization, attention to audience, and sentence formation and convention errors that interfere with meaning.

Limitations

1. Each of the two groups of teachers participating in the study included only a small number of subjects. The number of expert raters (eight) was limited by the number of teachers who were members of the 1988 Maryland Writing Test Scoring Committee Narrative Subcommittee. The number of novice raters (six) was subsequently limited to a comparable level and involved one high school English department.

2. Teacher participation in the study was limited to an English department from one Maryland high school. Teachers included in the study fell into one of three categories:

   a. teachers who at least one class period per day taught writing instruction to students preparing to take the Maryland Writing Test in grade nine

   b. teachers who provided state-mandated remediation instruction in writing to students in grades ten and eleven who have previously failed the state test and were preparing to retest

   c. teachers who used the Maryland Writing Test criteria in classroom assessment of student writing, but whose students had previously passed the Maryland Writing Test

3. Student writing to be used in the study was composed of samples of explanatory writing only. Maryland State Department of Education specialists in writing assessment and instruction had observed that teachers' definitions of explanatory writing quality seemed to vary among teachers more than their definitions of narrative writing. Consequently, it

appeared as if explanatory writing would allow a greater opportunity to observe those influences on rater scoring decisions.

4. The set of compositions was selected from actual compositions written in response to the 1988 administration of the Maryland Writing Test. However, the compositions were not randomly selected. Though compositions were characteristic of student work, they were ones judged by Maryland State Department of Education assessment specialists as being clearly representative of each of the score points. This also allowed a wider variety of compositions for the study since the ninth grade pass rate for the 1988 test was 82.2%. A random sample would have produced a large number of papers at the 3 and 4 score points.

5. Since expert raters who participate in the study represent several different school systems from across the state, materials were sent to them by mail, and those subjects completed activities independently. However, for five of the novice raters, the activities at the participating high school were completed in two single sessions with teachers present in one room, working at their own pace; two teachers worked in one session, three teachers worked in another session. One novice rater who was absent on the day data were collected completed the packet independently without monitoring.

6. Modified holistic scores and analytic scores (Appendix A) used in this study were both limited to a four-point scale range. Two of the analytic scores--errors in sentence formation and conventions--were further limited to a three-point scale. Attention to audience, a fifth feature, was rated on a two-point scale.

Assumptions

1. The compositions used in the study were typical of what the teacher sees in the classroom.

2. The performance of teachers approximated their behavior in rating their own students' compositions in their own classrooms. The teachers in the study were typical of their peers elsewhere in the state. As a consequence, teachers did not use sets of anchor papers in scoring compositions since anchor sets are not feasible in most classroom writing assessment situations.

3. The score assignments made by the writing assessment experts (certified modified holistic scores) were assumed to be accurate for all compositions.

4. Though results of the study would specifically apply to the preparation of students for the Maryland Writing Test, these results could apply to the application of scoring criteria other than the Maryland Writing Test criteria whenever scales and scoring criteria are used in instruction.

Overview of Method

Two groups of teachers used the Maryland Writing Test explanatory scoring criteria to score 35 compositions written by students for the 1988 Maryland Writing Test. One group of eight teachers consisted of expert raters in that they had qualified as raters at a training activity conducted by Maryland State Department of Education. A second group of six teachers consisted of novice raters in that they had been trained in a less intense, shorter training experience in which they were not required to reach a qualifying level of accuracy in scoring compositions. A third set of scores--certified Maryland Writing Test scores--provided a set of scores that were

assumed to be accurate scores with which to compare the scores assigned by novice and expert raters. Rater-assigned scores from both groups were analyzed to determine accuracy as well as predictors of score assignments.

Summary

The inservice activities used in Maryland to train teachers in modified holistic scoring have differed from the training that is provided to raters who score papers in large-scale holistic scoring operations in several ways. The training is generally conducted using prompts that are not later used in the classroom, is generally not followed up by rater monitoring, and does not require teachers to reach a qualifying level of accuracy. Training is designed to assist teachers in applying an imposed set of writing assessment criteria that often differs from personal criteria. Personal criteria can include emphases on content and organizational features of writing, but can also include a focus on mechanical errors such as spelling and punctuation. Scoring accuracy and the predictors of scores assigned by teachers were investigated to determine the impact of training on the scoring decisions made by teachers.

CHAPTER 2

REVIEW OF THE LITERATURE

Introduction

This chapter examines the current research in writing assessment, particularly as it relates to modified holistic scoring used in conjunction with the Maryland Writing Test. It compares and contrasts the Maryland test with other direct writing assessment procedures that have been developed in the field, and it examines the problems associated with holistic scoring, including criteria and threats to validity and reliability. The cognitive activity of rating compositions is examined in this chapter through an information processing model as a means of understanding better the influences that affect raters. Research on those influences, in turn, are examined in more detail, particularly as they emerge from the composition text. From those influences, certain ones pose threats to the validity and reliability of holistic scoring and are discussed in terms of the controls applied in many assessment procedures to minimize their effects.

Problems in Writing Assessment

The task of building a good yardstick of writing skills and of having it make valid and reliable measurements has been arduous. Flower and Hayes (1980a; 1980b), in their think-aloud protocols of writers as they worked, have helped to build a theoretical model to explain what is happening in the writer's mind. The validity of Flower's and Hayes' specific claims may be questioned in that having writers orally articulate their thoughts as they move through the writing process may reconfigure the actual chain of events involved in writing. However, this research has been successful in identifying the complexity of the task of writing. Clearly,

writing is a recursive activity requiring the interplay of multiple skills. Meanwhile, other researchers, for various reasons, explored the relationships of a host of other variables to writing and confirmed that at least reading skills played a part in writing measurements (Grobe & Grobe, 1977; Hiebert, Englert, & Brennan, 1983; Shannahan, 1984; Shannahan & Lomax, 1986). Consequently, any measurement of writing would not be a discrete measurement of writing skills alone.

Where educators have chosen to assess writing skills with multiple choice tests, simple, reliable measurements of isolated skills have resulted. They have been shown to correlate with written composition test scores at least at a moderate level (Godshalk, et al., 1966; Culpepper & Ramsdell, 1982; Huntley, Schmeiser, & Stiggins, 1979; Smith, et al., 1980; Breland & Gaynor, 1979). Further, some of those studies determined that some indirect measurements of writing skills had a significant level of predictive validity in terms of in-class performance at the college level (Gorrell, 1983; Breland & Gaynor, 1979). Similar relationships between multiple choice tests and composition tests emerged in studies of younger students (Hogan & Mishler, 1980). However, Moss, Cole, & Khampalikit (1982) were not able to replicate those findings in their study.

Researchers seem to have discovered that there are, then, statistically significant relationships between multiple choice tests of writing skills and composition ratings. Further, some have found that such indirect assessments of writing skills have demonstrated predictive and concurrent validity and are, above all, reliable. However, there are questions about the practice of simply correlating essay scores with standardized test scores and course grades. Clemson (1978), in her work

with the Basic Skills Assessment, found the essay portion of that writing test to correlate weakly with the multiple choice portion but concluded that the importance of having students experience the writing portion of the test was important enough to retain it in practice.

Direct writing assessments continue to be important in writing instruction for more powerful reasons than having students experience the opportunity to write. Assessing writing skills by having students produce writing samples gives the teacher an opportunity to examine both the parts and the whole. While multiple choice tests allow an opportunity to examine some component subskills such as those related to the mechanics of writing, they do not allow the student to compose. Consequently, the opportunity to examine a student's ability to organize ideas and to communicate meaning is not available through indirect assessment (Odell, 1981, p. 107). But such opportunities are possible with direct assessments. The problem, however, is in producing valid and reliable measurements of writing--of the whole.

In sitting down to assess a written composition the teacher is faced with numerous distractions that can make it difficult to measure how well the student has communicated information to the reader. One choice, in assessing student writing, has been to count the frequencies of specific writing features such as errors and to assess quality on the basis of this information (Odell, 1981, p. 119). An alternative has been to read the composition for the purpose of determining the quality of the sample. Such qualitative assessments are not as reliable, but they can produce broader statements about the writer's performance (Stiggins, 1982).

Pressures on state education agencies and national testing companies have driven some of the research in this area. Especially where assessments have been tied to program decisions such as graduation or promotion, the demands for valid, reliable assessments have been strong. Indirect assessments have sometimes been chosen by a few states for their writing assessment programs, but the vast majority have elected direct assessments, and those have been qualitative in nature (McCready and Melton, 1981). The cost and time constraints associated with processing large numbers of writing samples have limited the choices for agencies and companies and have produced some procedures, which are difficult to replicate in the classroom (White, 1985, p. 68).

Large-scale testing has produced some state-of-the-art standards for writing assessment, and some refinements in direct writing assessment have gained the support of some writers (White, 1985, p. 16). This chapter reviews the work of researchers to understand the cognitive base for the rating of writing quality for both large-scale scoring settings and for the classroom. This chapter also reviews research efforts to identify factors operating during the rating process to cue the rater to the writer's skills and to distract the rater from seeing the valid evidence regarding those same skills. From research in both areas may come the development of better scoring procedures, more effective rater training, or scoring devices that may assist in reducing the activity of scoring distractors in the qualitative assessment of compositions.

Qualitative Direct Writing Assessment

Some current state-of-the-art qualitative writing assessments may seem somewhat removed from the traditional methods used by English teachers to grade compositions. This is due in part to the refinements in assessment and to a current interest in measuring various writing skill competencies (Odell, 1981, p. 107). In the English class, the teacher will often circle errors and write comments in the margin about specific syntactic or rhetorical features of the writing, and he or she may have generalized criteria for grading. However, the teacher may not have attempted to specify criteria exactly and to apply those criteria evenly through the set of compositions. At the same time, accountability assessment programs have worked for specificity in criteria and precision in instrumentation (White, 1985, p. 24; Maryland State Department of Education, 1988).

Holistic Scoring Procedures

Teacher criteria arise from the definition the rater has in mind for writing, and it is assumed that each rater has that definition cognitively activated as he or she reads the writing sample. One type of qualitative writing assessment, holistic scoring, requires the rater to apply a specific definition of writing quality consistently to each writing sample. Further, primary trait scoring, a form of holistic scoring, permits the rater to measure quality in terms of a limited set of traits that are peculiar to the writing task required of the student (Lloyd-Jones, 1977, p. 32).

Similarly, the Maryland Functional Writing Program has employed a modified holistic scoring scheme (Maryland State Department of Education, 1987c). This sets the Maryland Writing Test apart somewhat

from other tests that employ scoring schemes that use pure holistic scoring designs. In pure holistic scoring schemes, the scoring criteria essentially emerge from the collected body of papers. A set of anchor papers is selected as representative of each of the score points, and scoring proceeds on the basis of those self-set criteria.

While this procedure is also a part of the modified holistic scoring method for Maryland, state officials have established a specific set of criteria for narrative writing and a slightly different set of criteria for explanatory writing. These two sets of criteria remain fixed from year to year and form the basis for the selection of the anchor papers for the annual test (Maryland State Department of Education, 1987c). This added step is essential in a large-scale assessment program to assure that the standards do not drift from year to year.

In building the test for each annual administration, new writing topics are developed. Consequently, after the first state test booklets are returned, specialists select hundreds of compositions from those written by students and identify from among them those compositions that clearly are representative of the kind of writing raters will be seeing. Maryland State Department of Education specialists work with the Maryland Writing Test Scoring Committee--a panel of 24 teachers and supervisors from across the state--to place scores on a limited number of those papers, using the Maryland Writing Test scoring criteria. Following lengthy concensus sessions, the committee is able to identify the compositions that will be used to train contracted raters and other compositions that will serve as a reference for raters--the anchor papers.

The scoring procedures remain a critical part of test construction and administration, but the writing tasks themselves are equally important. In the design of the writing prompts or topics, educators remain divided on the most desirable characteristics of ideal topics. McColly (1970) found that if writing prompts were broad, allowing for a wide variety of responses, the ability of the reader to provide a valid assessment of the writing sample would be impaired. Consequently, he recommended that subjects be forced to respond to very narrowly defined prompts. Meanwhile, another declared that at least some latitude should be allowed for the student to utilize personal experience in writing (Lloyd-Jones, 1977, p. 41). The lack of a personal knowledge base would severely limit the ability of a writer to respond to the topic.

Clearly, the designers of writing prompts would be hard-pressed to produce test questions that adhered to both restraints. In designing prompts for the Maryland Writing Test, Maryland State Department of Education has indicated that of about one hundred prompts submitted by prompt writers to the agency, only one or two pairs are deemed suitable for use in live testing situations. The screening process eliminates prompts that are invalid, culturally biased, or socially inappropriate and which do not perform well in field testing (Maryland State Department of Education, 1987a).

While the large-scale assessment programs generally pay very close attention to prompt design, some research has indicated that such concern may be at least in part unwarranted. In a comparison of essay topics that included considerable detail in outlining the writing task with less specific versions of the same topics, there was no significant

difference in the scores on student responses (Brossell, 1983). In that study, it was supposed that the increased attention paid to construction of the prompt, and particularly to providing the writer with detailed specifications for the task would enhance the students' ability to respond. However, students did not appear to respond to the additional assistance provided by the text of the prompt. Similarly, a study by Karen Greenburg (1981) showed that students who were allowed to choose their own response designs did not write significantly better.

Consequently, if prompt design, response design, and the ability of a test question to engage the personal experience of a writer do not operate as powerful variables in affecting student performance, the classroom teacher appears to have much freedom in writing assessment. He or she should, with relative ease, be able to produce for students writing tasks that result in samples of writing from which valid measurements of skills can be taken. Though the Brossell (1983) and Greenburg (1981) studies might indicate that classroom assessment design can be somewhat forgiving, no such flexibility is implied for large-scale assessments. Those who design writing tests for administration to large numbers of students with diverse backgrounds often employ intensive procedures for screening prompts about which it might be difficult for some students to write (White, 1985, p. 108).

The rater probably brings to the scoring situation more threats to validity and reliability than are contained in some instruments. For large-scale assessment programs, the rater is trained to understand the criteria and is checked periodically to see if he or she applies the criteria consistently to all writing samples (Maryland State Department of

Education, 1986, 1987c). The classroom teacher often operates without benefit of specified criteria, training, or monitoring. Consequently, the defenses against threats to validity and reliability are weak. In order to understand the sources of those threats, the testing situation and especially the rating procedures need scrutiny.

Holistic Scoring Criteria

Since direct qualitative assessment seems to be a principal choice for informal classroom writing assessment, the charge to the classroom teacher may in some ways be more serious than it is to administrators of large-scale assessments. The teacher is faced with identifying the precise criteria for assessment, with electing a valid and reliable scoring procedure, and with assuring that the student understands that he or she is being rated against the criteria. Further, the criteria need to be valid. In many reported studies of holistic scoring, the criteria are not reported, but in others, researchers have worked extensively to identify rating scales and criteria that have proven research or theoretical bases (Cooper, 1977, p. 15; Lloyd-Jones, 1977, p. 49; Breland and Jones, 1984). The criteria, in Maryland, can be the Maryland Writing Test criteria, which have recently been clarified and distributed to state English language arts teachers (Maryland State Department of Education, 1987b). The scoring procedure can be holistic scoring and patterned after that described by numerous researchers and agencies (White, 1985, p. 19; Cooper, 1977; Maryland State Department of Education, 1986, 1987c; 1988). Though it is possible for one individual rater to score papers holistically, the teacher will not likely be able to attain in isolation the reliability that is possible with multiple raters (Breland, 1983).

In the classroom, the teacher is left essentially unmonitored to score papers with the only guidance being provided by an explicit scoring rubric. Whereas, in the Maryland Writing Test scoring rubric, not only is the rater trained to identify papers at each score point and in the identification of characteristics related to the criteria, raters are also periodically checked to identify those who are displaying drift, assigning scores that are dissimilar from the scores assigned by other raters to the same papers (Maryland State Department of Education, 1986).

Sources of Influences on Raters

Freedman and Calfee (1983) have categorized sources of influence on score assignments that can apply to both large-scale efforts and to the classroom. They felt that one cluster of rating influences arises from the text of the writing sample. However, they also surmised that another realm of rating influences is actually carried with the rater to the scoring session cognitively in the form of the rater's schema. They felt that the third cluster of influences arise from the physical and psychological context within which the rating activity occurs. The complex task of unravelling the scoring process, then, would center on developing a better understanding of the cognitive activity associated with rating and on those influences affecting scoring decisions.

While Flower and Hayes (1980b; Hayes and Flower, 1980) have worked to explore cognitive activity of writers who are producing written text, little parallel work has been done to structure the cognitive activity of raters who are attempting to assign a quality rating to that text. That is not to say that researchers have not explored quality ratings thoroughly. In fact, considerable work has produced an increasingly clearer picture of the

characteristics of writing that has been judged by raters to be of both poor and exemplary quality. Unfortunately, that research has largely focused on relatively simple syntactic and lexical features of writing and on how those features related to quality rating assignments (Cooper, et al., 1984; Freedman, 1979a, 1979b; 1984; Grobe, 1981; Neilsen & Piche, 1981; Nold & Freedman, 1977; Schmeling, 1970).

Cognitive Processes in Rating

Freedman (1979, 1984) also investigated the relationship of content and organization to scoring papers, but little research has explored the actual process. Freedman and Calfee (1983) proposed a model for rating essays based on three component processes--creation of a mental image of the text following reading, evaluation of that image, and articulation of the stored impression of the rating. Freedman and Calfee justified this search for the substructure of rating as congruous with the ideas of Simon (1981), who in The Sciences of the Artificial proposed the concept of decomposable hierarchy. This dissection of the structural process, according to Simon (1981, p. 106), would lend insight into the intercomponent interactions. The Freedman and Calfee schematic of cognitive activity proposed three subcomponent processes, but the actual process may be more complex.

In the absence of a specific body of research on decision-making as it occurs in holistic scoring, some insight may be derived from principles associated with signal detection theory (SDT). The theory originated from work that was first published in 1948 by Shannon (1949) in which the author produced a complex set of twenty three mathematical theorems relating to the transmission and reception of electronic signals. As an

engineering advance, Shannon's theorems resulted in one set of research activity in the electronic field. However, simultaneously, another body of work began in the field of human communication when Weaver (1949) suggested that signal detection theory could aptly be applied to the processes underlying the sending and receiving of information in humans.

In revisiting the original papers of both Weaver and Shannon, Ritchie (1986) attempted to clarify the differences between Shannon's theorems and Weaver's interpretations for the social scientist. In Ritchie's discussions, he was less concerned with the direct application of each mathematical theorem to actual physical human activity. This arose, in part, from the inability of SDT theorems to work when applied to auditory and visual discrimination tasks in the laboratory. Ritchie suggested, instead, that the theory be applied only when three conditions prevail: where problem situations can be decomposed into subproblems that fit Shannon's theorems, where problem situations resemble Shannon's theoretical problems but do not specifically agree with his assumptions, or where problem situations can be attacked using hypotheses emerging from Shannon's assumptions.

Ritchie went further to propose that more important, perhaps, than Shannon's theorems might be his method. He viewed the problem of noise obscuring the signal in information transmission with an intent to understand better the mathematical relationship of the two to accurate reception of the signal. In applying this problematic view of information transmittal between sender and receiver to the task of assessing the quality of writing, some parallels seem to emerge. Consequently, it might be of use to consider using the simple signal detection model as a tool in

conceptualizing what occurs cognitively when the teacher decides on a quality rating for a student composition.

If one considers the text of the student writing sample to be the source of information in this model, that transmission might consist of a signal and accompanying noise. The signal, of course, could represent the writing characteristics that are related to the criteria the rater is using to assess the quality of the composition. Error in detection of the signal, and in turn of judgment, is related to the noise accompanying the signal that is transmitted as well as noise introduced in the processing of the signal (Lindsay, 1970, p. 154). The processing, or rating of the paper, could entail the application of the identified scoring criteria to the signal. The noise introduced at that point could be identified from the Freedman and Calfee (1983) model as emerging from the context associated with scoring and from the text itself. However, noise in the rating process also might arise from the activation of extraneous scoring criteria that are a part of the rater's schema, but not necessarily part of the scoring criteria that the rater is expected to be using at that time. Consequently, error threats that diminish the integrity of the judgment process seem to emerge from points external to and internal to the rater.

In the Flower and Hayes (1980b) model of composition construction, cognitive activity, during writing entails sorting information in the writing as it is read, and storing in short term memory certain elements of the composition in terms of the text to be composed. Selecting information for rating may mirror some of that process. By viewing decision-making about the quality of a sample of writing through a signal detection model, it may be possible to understand how the rater detects

the correct signal before applying the criteria to the writing. However, the decision-making process in writing assessment is not simplistic but rather is more similar to multichannel processing situations (Lindsay, 1970, p. 157).

As the rater reads the written text, he or she is completing a complex task that is much more than passing a string of words through memory, as it may appear on the surface. In Maryland, the text is simultaneously evaluated for five discrete characteristics (Maryland State Department of Education, (1988), which could be treated as five different channels of signals entering, occasionally being stored in short term memory. Memory becomes the site at which the writer not only collects details, but also builds comprehension of the passage and provides the basis for recursive decision-making as multiple facets of the writing are considered.

Lindsay (1970, p. 24) proposed multiplexing of information as a possible model for processing some kinds of data. He saw the signals entwining and overlapping and then being decoded through a filtering process. Unfortunately, the detection of such multichannel signals produces a higher range of error than where one simple signal is to be identified from a background of noise (Moray, Bates, and Barnett, 1965).

Decision-Making in Holistic Scoring

Treating each of the four scale points (in the Maryland rubric) as categories for each of the five separate criteria, the reader appears to assign a score to each criterion before subsuming those five criterion-based decisions into one overall score. The decision, then, is more complex than those discussed by McNicol (1970) where raters expressed

certainty of decisions on a rating scale. At each criterion, the four-point scale infers four decisions. The rater must accept or reject each score point separately as representative of the criterion. Four such sets of five decisions, then, precede a final decision where the overall quality is judged on the basis of the twenty independent choices.

The preliminary decision-making is not verbalized, nor is it recorded. When formalized, such scoring resembles procedures associated with analytic scoring scales (Cooper, 1977, p. 24). However, by not formalizing the weighting of features, the rater is able to attend to the overall effectiveness of the writing in communicating meaning to the reader. At the same time, holistic scoring can allow that general impression score to become suffused with bias from the rater or can allow the rater to drift from the intended criteria (Charney, 1984).

The difficulty in applying signal detection research to written composition is seated in the complexity of human language. Much research applying signal detection theory to human communication has focused on simplistic, essentially single-channel information transmission; researchers have examined how well human subjects are able to perform visual and auditory discrimination activity involving single tones, numbers, or words (Ritchie, 1986). Further, decision-making in rating written text is far beyond simple detection and discrimination tasks in that the rater is asked to comprehend complex meaning from the text. However, the reader is also required to assess the entire piece of writing using a multi-faceted rubric that specifies in precise language what should be important to the rater at the moment of decision.

Signal detection theory may be useful to the researcher in identifying the kinds of validity and reliability threats made by the rater's own schema, by the context, and by the text itself (Freedman and Calfee, 1983). The advantage, then, of analyzing the rating process through such a model helps in classifying the so-called noise sources into clusters of scoring influences in a more exact fashion than can be accomplished through the Freedman and Calfee theoretical model.

These clusters are not discrete, are not entirely external to the criteria-related text characteristics, and are not capable of being isolated, in many instances. While, in practice, it may be difficult for a reader to set aside personal bias in rating and to disregard writing characteristics that are not appropriate at the time, it should be possible to measure the extent to which those characteristics perform as influences in the scoring process.

Text Features That Influence Raters

It is tempting to identify a number of qualitative and quantitative measures of text that could be touted as scoring influences. Many of those measures, which emerge from the literature on assessment practices, are artificial and merely represent characteristics of the writing that are easy to count. The real influences are likely less finite and less discrete. These characteristics operate in chorus with other characteristics and, in some cases, do not have a name. However, some measures, such as lengths of words, sentences, and compositions, are cues to the reader of both related and unrelated features in writing. The meaning of such features in a study of influences on raters may be less obvious to the researcher and needs to be viewed from several vantage points.

Sentence Complexity

Though a rater does not read text while using a mental yardstick, the size and complexity of syntactic structures often seem to correlate significantly with quality ratings of writing (Thomas and Donlan, 1980; Combs, 1976; O'Hare, 1973; Pedersen, 1977; Potter, 1967; Schmeling, 1969). This is not surprising in that better writing would seem to include some of the more convoluted sentence structures typical of mature writers. However, complex writing is not necessarily the benchmark of quality. Some of the many studies of sentence complexity and quality ratings contained questionable procedures in which assessment criteria were not clearly delineated. Further, other studies found little relationship between complexity and composition quality (Mellon, 1969; San Jose, 1972; Faigley, 1979).

Complexity of writing cannot be measured directly, but rather must be measured indirectly through counting features of the writing. The standard measure of complexity, the t-unit, was developed by Kellog Hunt (1977, p. 93). The t-unit is generally defined as a unit of thought or an idea consisting of a main clause and all of its supporting ideas. Such units are often followed by a terminal punctuation mark, thus the "t" in t-unit. Though not all such lengthy structures are effective conveyors of information, it was perceived that where the word count for the average t-unit was larger, the likelihood was greater that the writer had subordinated thoughts under the main idea of the independent clause. Thus, a measure of the mean t-unit length would give a secondary reading of the level of sentence complexity in student writing. More precisely, the t-unit mean size is a limited measurement of the ability of the student to determine what the central

idea of a sentence should be and to support that idea with other appropriate information.

The role of syntactic complexity and its relationship to the development of the young writer's skills are not indigenous to the English language. Studies of Dutch students (Reesink et al., 1971) have confirmed that younger writers produce simpler structures than their older peers and that students increase their sentence length with experience and age. The sentence complexity patterns that had been observed by Hunt (1965) in American children were clearly replicated in the Dutch students despite the structural differences between the two languages.

Other researchers testing the effects of Hunt's sentence combining curriculum at various grade levels found it to be successful in increasing the length of compositions and in building sentence complexity (Miller and Ney, 1968; Hunt and O'Donnell, 1970). Crowhurst (1980) found that younger students displayed no significant quality differences when such ratings were examined in relation to sentence complexity, essentially, that more complex writing was not of a better quality. However, in that same study, the highest quality compositions of high school students proved to display more complex syntactic structures. The link between syntactic maturity and writing quality remains elusive in the running contradictions experienced by researchers, including Hillocks (1984, 1986, p. 75) who examined collected studies in writing instruction in his meta-analysis. However, the t-unit remains important as a tool for measuring complexity of writing and has remained as a standard unit of length used in many quantitative studies of writing quality.

In t-unit studies, it is often unclear as to what basis upon which quality judgments were made. Where criteria were not specified, it is difficult to ascertain the role that syntactic maturity might have played in quality judgments. Conceivably, the complexity of syntactic structures may serve to cue rater bias or to activate irrelevant criteria in the decision-making stage of assessment. Though it may be impossible to determine the extent to which syntactic complexity should be expected to occur in Maryland Writing Test compositions, studies of composition quality should continue to examine relationships between sentence complexity and composition quality in order to assist in clarifying the currently clouded issues. Certainly, the t-unit is an established unit of measure and seems to provide an indirect measure--more accurate than sentence length--of sentence complexity. Further, the contradictory evidence regarding the relationship of syntactic complexity and writing quality leaves that question open for exploration.

Mechanical Errors

The t-unit measure has become a standard means of quantifying writing for the purpose of measuring other characteristics of writing such as frequencies of errors. Gary Grobe (1981), in replicating a prior study done with Stewart (Stewart and Grobe, 1979), used t-unit length, clauses per t-unit, and various error counts per t-unit in a study of influences on teacher rating in the New Brunswick Writing Assessment. The t-unit served as a reasonable quantitative base from which to characterize the tendency toward errors in writing samples and assisted in identifying relationships with other factors. In both of Grobe's studies, spelling errors appeared less frequently in papers with higher quality ratings.

Throughout the literature, researchers have reported that spelling errors appeared more frequently in poorer quality writing (Baddely and Wing, 1980; Neilsen and Piche, 1981). Breland and Jones (1984) found spelling errors to correlate more highly with quality ratings than any of forty other characteristics of writing. These findings were echoed in a study of Stanford freshmen (Cooper, et al., 1984). It is an especially troublesome phenomenon in that rating criteria such as those for the Maryland Writing Test generally direct the reader to ignore spelling errors where they do not interfere with the meaning of the passage. Yet, the prolific appearance of such errors in the writing of less effective writers serve to cue the reader, engaging bias or activating the spelling error criterion in cases where it is not to be applied to writing as a quality measure. Spelling, too, is an especially thorny feature because it is an important criterion for judging certain other types of writing exercises. Consequently, it is a suitable feature to include in a study of lexical and syntactic influences on ratings.

Other types of writing errors are not as easily classified as quality criteria. That situation probably results from pedagogical influences on the teaching of writing. From the earliest stages of composition instruction, one of the primary modes of student feedback on writing performance has been error analysis. When the definition of writing is centered primarily on mechanical perfection, errors are often circled, red-lined, and counted, with a grade assigned on the basis of error patterns. Error frequency is often incorporated in research today such as that by Grobe (1981) and others as a means of documenting error frequency and not necessarily as a means of supporting the practice of teaching mechanics.

Traditionally, a classroom emphasis on errors has often permeated instruction (Applebee, 1984), but not without impeding the productivity and stifling the learning progress of some students. In studies of writers who performed considerably below expectations, it was discovered that error-centered pedagogies had contributed significantly to writer apprehension (Rose, 1983; Smith, 1984). Further, Mina Shaughnessy's (1977, p. 5) study of basic writers at the learning center of the City University of New York revealed that a large volume of student writing was riddled with errors. Her work there credited much of the writing incompetence of those basic writers to the misguided efforts of the earlier teachers of those students. She felt that the emphasis on errors in the instruction of developing writers was not only overdone but was also wrongly structured.

Shaughnessy called for teachers to go beyond the superficial levels of error identification when evaluating student writing. She felt that students would benefit much more by classifying errors by their inherent causes and that instructional programs should be centered less on eliminating the errors in student writing and more on assisting the student in moving toward writing proficiency.

These findings raise important questions about the use of error counts in assessment as well as instruction. If assessment is to provide a valid measure of skills achieved, then it must match what is taught. The use of error-based instruments may only serve to focus instruction on syntactical and mechanical problems and would not enable students to understand the extent to which broader writing skills are needed in quality writing. Such limited information on the writing product might lead the

teacher to overlook other unassessed, but critical writing characteristics when delivering instruction.

The definition of writing competency, according to the Maryland Writing Test scoring criteria, deemphasizes mechanical errors (Maryland State Department of Education, 1987b) It directs the reader to consider errors only where they become impediments to the conveyance of meaning. The treatment of errors as a secondary characteristic of good writing does not release the student from the responsibility of assimilating a solid base of ground rules for constructing written text. Rather, it is in chorus with the views of Shaughnessy and others.

The Maryland test, because it measures a minimal standard of writing (i.e. basic or minimal competency) is, by virtue of its purpose, an assessment component of a basic skills instructional program. The instrument was designed to assure that high school graduates could perform simple writing tasks such as completing job applications and writing business letters. As such, it was not designed to assure that graduates could produce perfect prose. Further, the student is asked to write only a second draft of each composition for the test, but is not required to put the writing in final form. Consequently, readers were directed, both in training programs and in the scoring rubric to be forgiving of mechanical errors as long as such errors did not obscure intended meaning (Maryland State Department of Education, 1986; 1987c; 1988).

Much of the research on writing quality has been performed at the post-secondary level and thus has involved a somewhat higher standard in measuring writing quality. But studies at that level and lower have always shown spelling errors to be closely linked with writing quality

(Baddely & Wing, 1980; Bartholomae, 1980; Freedman 1979a, 1979b; Gorrell, 1981; Williams, 1981). The Stanford study of college freshmen (Cooper et al., 1984) had examined spelling as well as other error types, but Cooper warned that gross error counts produce unclear data about writing. Whereas spelling errors are often simply the result of an inability to recall the correct spelling, the nature of specific syntactic errors is more complex.

The placement of punctuation, capitalized words, and the construction of sentences call on a variety of skill and knowledge bases. An examination of rating influences would need to include mechanical error counts (Cooper et al., 1984) to tie with prior research. The reason for counting errors might be less to view the specific features of writing that are judged to be of high quality and more to determine the priorities of raters. A rater who was directed in training and in scoring rubrics to diminish the importance of spelling errors might be capable of separating out spelling errors as an isolated error type. However, some error types never fully fall outside or inside of the stated criteria for a particular writing quality assessment. It is important for raters to know when to activate certain criteria relating to mechanics.

When rubrics omit error frequencies, the raters need to score compositions accordingly. Sweedler-Brown (1983) raised those same concerns about raters. However, she found in her study of university level composition scoring, both trainers and raters seemed to properly base their holistic scores on quality-related features of the rating. She also felt that experience and training were inherently responsible for focusing

raters on the proper criteria and for keeping them from straying from the rubric into personal criteria.

Alternative Error Analysis Taxonomies

The scheme for error analysis in studies can be selected from among many different models, most of which seem to have a logical base, but not necessarily a theoretical nor instructional rationale. For instance, Cooper et al. (1984, p. 24) elected to focus on punctuation errors (terminal and within-sentence), spelling mistakes, and usage errors (tense and agreement only), recognizing that there was a virtual galaxy of error types that could have been included. However, he gave no clear reason why only these were selected. All of the error types selected by Cooper provide data on quality that could be of use in instruction, but are based on direct, quantitative measures.

In contrast, Nystrand (1982, p. 66) suggested examining error in terms of "textual cognition"--a term defined as the writer's awareness of the need to set a stage for the reader. The inability to set this stage would result in misleading the reader or providing the reader with either too much or too little information. These three kinds of error were applied to the handwriting and the lexical, syntactic, textual, and contextual levels of the writing. This taxonomic structure for error would be instructionally useful and could provide a meaningful interpretation of error causes as suggested by Shaughnessy. However, it may be exceedingly difficult to make valid and reliable decisions on error counts without better understanding both Nystrand's taxonomy and how to apply it. It also becomes difficult to assess the degree to which Nystrand's design reveals anything about a reader's personal bias set. While the taxonomy may be

instructionally valid, there is no evidence to suggest that the reader carries such a taxonomic structure to his or her bias to the writing assessment situation.

The psycholinguistic writing model assumes that the writer assembles sentences clause by clause in much the same way speakers compose what they say (Daiute, 1981). From this research, a taxonomy of sentence-level errors emerged that includes twelve types of syntactic disorders (Daiute, et al., 1981). The scheme targets problems with sentence fragments, modifiers, repetition, and parallel structures, among others, and identifies them in language that normally might be applied to speech. Though a reliable text analysis appears feasible using this taxonomy, there is no reason to believe that one kind of structure would be more likely to activate reader scoring errors than another. Further, all of the twelve types of errors hold the potential for impacting meaning in text.

While the psycholinguistic error taxonomy has a theoretical base, the errors analyzed in computer-based text analysis systems such as Writer's Workbench have a pragmatic source in that they are designed to assist writers, editors, and managers in editing limited aspects of a document (Bell Laboratories, 1982, p. 1-2; Kiefer and Smith, 1984, p. 66). Writer's Workbench System is likely the most sophisticated text analysis program currently available. It was developed by Bell Laboratories and generally requires a UNIX system main-frame computer in order to function. The premise upon which Writer's Workbench operates is that a set of standards exists against which each composition is examined. The text is analyzed to tell the reader or editor of possible errors, passive verbs that the writer has used, and vague terms that appeared in the text. The

program also generates an index value that indicates the extent to which the text is abstract in its presentation of information. The writer is also told how much he or she wrote, how long the composition is, how long the average paragraph is, and what the readability of the writing averages. On a number of these measurements, the computer system indicates to the writer how well the composition measures up to the predetermined standards and suggests to the writer how improvements might be made.

While the limitations of Writer's Workbench prevent analysis of more complex syntactic structures such as those suggested by Daiute, the strengths of the program seem to lie in its identification of clearly measurable features. For a study of scoring influences, Writer's Workbench provides a menu of some obvious features that might activate bias or incorrect score assignments to compositions. Further, some of the features measured in the program have been found to relate to judgments of writing quality. For example, longer compositions seem to be of a better quality (Grobe, 1981; Stewart and Grobe, 1979; Breland and Jones, 1984; Nold and Freedman, 1977; Thomas and Donlan, 1980). In contrast, readability has correlated both positively (Nold and Freedman, 1977; Neilsen and Piche, 1981) and negatively (Chase, 1983) in studies of student compositions.

Though one might expect more mature lexical choices to impress raters, the Chase study posed some interesting countering arguments. Unlike the compositions used in the Nold and Freedman (1977) and Neilsen and Piche (1981) studies, Chase's student compositions contained no errors, which might inadvertently trigger rater reactions. However, because it compares only two compositions of supposedly

comparable quality, the results pose methodological questions. While attempting to increase controls on the text, Chase has limited the number of opportunities from which to establish patterns of individual rater performance. Despite this possible limitation, Chase's work suggests that the readability issue is worthy of further examination.

The features of writing suspected as text-based influences on scoring errors arise from the literature on writing quality. For various reasons, researchers have been able to document with some degree of accuracy the appearance of syntactic and lexical characteristics of student writing in good and poor compositions. As a teacher develops a personal definition of writing quality, the specific criteria assimilated by the teacher can serve to assist in rating on some occasions and can interfere in others. The error taxonomies provide an opportunity for the researcher to consider restructuring the traditional ways of viewing error and thus may reveal unique bias constructs not apparent in simple error counts. Unfortunately, there is a lack of evidence for a valid choice from among those taxonomies to form the base of a research scheme. If, in future research, evidence emerges to substantiate and clarify any of those taxonomies, then inclusion in a study of rating influences might be appropriate.

The simplicity of the computer-centered analysis offered in Writer's Workbench and its lesser counterparts is seductive and could be considered a data source, but, in isolation, it is limited in meaning if one is to understand larger patterns of text features. However, the linkage of Writer's Workbench to traditional text analysis approaches does provide insight to the researcher. As a writer's tool, the program speaks to the writer in relatively traditional terms (i.e., usage, sentence structure,

grammar errors, etc.) in that the writer has been trained in such terms in traditional English courses. Consequently, it would be a sound assumption that the biases of the writer might have been formed around the structure of the traditional error terminologies. A study of influences on the quality assessments of teachers might, then, assume that the teacher's traditional training and the current strength of research in the area both would merit the structuring of a study around a traditional error taxonomy.

Validity and Reliability Threats to Holistic Scoring

The rating of writing quality is considerably more advanced than it was around the turn of the century, but even then, Thorndike (1912, p. 214) recognized the need to score writing samples on a consistent scale. He suggested to the educator that a scale produced by a colleague, M.B. Hillegas, would serve as a suitable model. However, Thorndike did not address the issues of validity or reliability and merely produced for the reader samples of writing that represented scattered points on his scale. It is interesting how Hillegas' scale and models of writing at various levels of competency resembled anchor papers from modern day holistic scoring procedures. Unfortunately, the Hillegas scale did not articulate the criteria by which each sample of writing was placed at that point on the scale.

Such devices as the one suggested in that classic early primer on education go far toward supporting scoring procedures, but they are heavily dependent, in modern forms as well, on extensive training of readers and on clear and rigid scoring criteria. These criteria must designate both the writing characteristics to be considered by the reader and the score assignments associated with each combination of characteristics (White, 1985, p. 24).

Perhaps the largest singular unknown that impacts the assignment of scores, then, is the reader. It has been shown that certain characteristics of the writing sample can lead the reader to assign a score that does not correctly reflect the quality of the writing sample in terms of the criteria (Coffman, 1971). The tendency of an instrument to be sensitive to irrelevant data in certain writing samples constitutes a major threat to test validity and reliability. To address this problem, it may be of interest for a scoring scheme, then, to include a limited array of writing characteristics (Lloyd-Jones, 1977, p. 45). However, this requires the reader to be trained to be sensitive only to those characteristics directly related to the narrow band of rating criteria. In terms of a signal detection model of scoring, this choice helps to isolate conceptually that portion of the transmission "noise"--the non-criteria related features of writing--that can be activated improperly at the time of composition scoring.

Context Threats to Rating Accuracy

Physical aspects of the writing sample have been shown to interact with other factors, thus interfering with correct assignments by triggering rater bias. Writing samples with poor quality handwriting have been shown, for example, to be rated more harshly than those produced with good penmanship (Chase, 1968; Markham, 1976; McColly, 1970). A later study by Hughes, Keeling and Tuck (1983) corroborated the findings of the earlier researchers and further indicated that the negative impact of the student's handwriting was far stronger than the power of positive rater expectations for any particular student. This powerful statement of the importance of handwriting appearance gives a hint at how difficult it is for a rater to ignore factors clearly outside of the realm of the criteria.

Sometimes, teachers might choose to use writing mechanics as a criterion for evaluation. However, handwriting quality is not even a characteristic of the text and would likely be more akin to context factors external to the composition.

Even in heavily controlled scoring situations such as those used to score the Maryland Writing Test, bias emerges as a significant factor (Ferrara, 1987). The effects of fatigue and of bias influences of preceding compositions when scoring large quantities of papers have been documented as context effect. This phenomenon involves the interplay of the environment with the scoring process. It has been demonstrated that the rater's scores are influenced significantly by scoring context (Daly and Dickson-Markman, 1982; Hales and Tokar, 1975; Hughes and Keeling, 1984; Hughes et al., 1980a, 1980b, 1983). Consequently, large-scale scoring operations often randomize order of papers to reduce the chances of two readers experiencing the same sequence. The study by Ferrara confirmed that context effect was a significant factor for the Maryland test, justifying the expensive resequencing of papers for readers. Though in Ferrara's study the impact of order on bias did not significantly alter the pass-fail rate, it suggests that context effect could be an even greater threat to validity and reliability in the classroom where the teacher has minimal training in holistic scoring. Though the teacher can randomize the order of papers, there is no second reader. Further, order effect will likely interplay with other bias factors in the classroom.

Text-Related Threats to Rating Accuracy

A related issue is the impact of writer confidence on ratings. Though scores of confident writers seem to be higher than for their less

confident counterparts (Freedman, 1983; Daly and Wilson, 1983), it is not clear if this is due to rhetorical features of the writing or skill differences. A third possibility is that the rating of the sample is inflated for confident writers. Inasmuch as a confident speaker conveys his or her confidence in the cadence and delivery of the speech, a confident writer could possibly cue the reader to his or her attitude. However, it appears as if no research in this area has been conducted to date.

Writing assessments, particularly direct qualitative assessments, seem to measure quality of writing, but they also measure or at least cue the reader about other characteristics of the writer. Consequently, researchers have worked extensively to uncover those characteristics particularly in light of the current discussions about generalized language skills (Applebee, 1984). Further, the measure of reading skills has shown to have some degree of relationship to many writing skill measures. Correlations have been used repeatedly in writing studies to reveal these relationships and to establish test validity, but not without raising questions.

Validity of Writing Assessments

Occasionally, researchers have attempted to establish the validity of direct writing assessments by performing correlations between essay scores and multiple choice measures (Godshalk et al., 1966; Breland, Conlan, and Ragosa, 1977; Breland and Gaynor, 1979). Other studies sought to validate holistically scored essay tests using academic achievement (Breland, 1977, 1983; Michael and Shaffer, 1978; Breland and Gaynor, 1979). All of these attempts produced rather lackluster

correlations, significant, but ranging from as low as r = .20 to around r = .65.

Steven Jay Gould (1981, p. 251) warned behavioral scientists of the danger of eliciting too much meaning from moderate but significant correlations. For researchers trying to find causal factors for phenomena, for bases for behavioral patterns, or for test validities, Gould's caution is worthy of consideration. Seemingly, other than face validity and sometimes construct validity, it is difficult to satisfactorily prove the validity of holistically scored writing tests. The difficulty lies in an inability to identify precisely parallel measurements that can help to substantiate that a writing assessment measures what it is intended to measure. In other words, there are few writing assessments available, and when one is a multiple choice device and another is a direct assessment, then it becomes questionable to attempt to validate one with the other. A similar problem emerges when two direct assessments of writing are correlated when the two are not based upon the same scoring criteria.

Reliability of Writing Assessments

Reliability, in contrast, is easier to establish in writing assessment. Godshalk et al. (1966) were interested in whether several readers would produce the same score. The researchers concluded that the reliability of scoring compositions was directly linked to the number of samples of a student's writing, with their results indicating a reliability of .92 where five samples from one writer were scored by five readers. However, other studies have produced less consistent findings. Blok (1985) found that where two raters were less successful in agreeing with each other (ranges from r=.415 to r=.91) any one rater remained consistent in his or her

scores. One might question Blok's latter conclusion in that intra-rater reliability was established by having a reader read and score the same composition twice. In contrast, the Maryland State Department of Education (1986) found that two raters agreed on the score of any one composition approximately 77% of the time in the 1986 administration of the test. There was a moderate correlation between the scores assigned to the two compositions for each student, but that was likely due to the differences between narrative and explanatory writing tasks.

In general, reliability in scoring direct assessments of writing is tied to the type of writing task, the number of raters, the number of samples, and the type of scoring approach (Breland, 1983). In terms of numbers, it has been shown that if one reader scores one writing sample per subject, the reliability will range around .50 (Coffman, 1966). An increase in the number of writing samples from one subject to five will raise reliability to .9 if the number of times each sample is rated is at least three (Godshalk, et al., 1966). Reliability has also been shown to be related to the number of score points in the scoring rubric through fifteen points. After that point, increases in reliability have not been indicated by the research (Breland, 1983).

Certain analytic scoring schemes, in contrast, have been shown to be more reliable than holistic scoring (Breland, 1983), except for atomistic rating approaches, which are less reliable than either holistic or limited scale analytic scoring (Moss, Cole, and Khampalikit, 1982). For holistic scoring schemes, the training of raters has been shown to be a major factor in assuring reliable and valid scores (Sweedler-Brown, 1985). Maryland, for example, has provided an extensive training and

qualification component for its state writing test (Maryland State Department of Education, 1986, 1987c; 1988). Though statistical reliability for holistic scoring is often reported in the literature, declared criteria for readers may differ appreciably from actual criteria used by individuals, or even from valid measures of writing competence. Other states and testing corporations as well have elected similar exhaustive training and scoring procedures, especially where the integrity of scores must stand up to public scrutiny (White, 1985, p. 149). While such measures may be practical for large-scale assessments, they remain beyond reach for daily classroom usage.

Summary

The classroom arises, then, as the arena where the next generation of composition assessment research should move, now that the large-scale assessment programs have etched out a frame of reference for more valid and reliable assessment of student writing. The application of signal detection theory to the rating process may assist in identifying the sources of influence on rater decisions. The studies of lexical and syntactic features of quality writing may lend insight into rater schema from which scoring criteria emerge. The documented effect of non-text features such as handwriting may help to complete the picture of the rater's thinking. One might assume that all these validity and reliability threats are at least somewhat controlled in scoring situations where raters are trained and monitored. The classroom teacher, however, would be more vulnerable to these threats while being expected to steer the student effectively toward writing competence. Consequently, a study of the types of influences on teacher ratings of student writing would help to build a foundation for

exploring interventions that might strengthen the validity and reliability of classroom writing assessment.

The difficulty presented with the task is apparent when one attempts to sort those features of writing that are unrelated to the assessment criteria from those that are seemingly related. Some features may simultaneously cue readers to two different sets of criteria. Further, the individual reader's experience and bias may magnify the effects of certain other features of writing as influences. On two separate levels, it would be important to examine how the impact of scoring influences approach universality.

The classroom English teacher in Maryland provides an especially viable subject for examining these influences. The published criteria of the state writing test can be seen to operate in harmony with the definitions of writing in some classrooms and in conflict with the definition used in others. Performance data on a cohort of Maryland teachers could assist in assembling a picture of that interaction and suggest ways to compensate for the lack of training and monitoring of scoring in the classroom.

Hirsch (1977) wrote that the largest single factor impeding effective research and instruction in writing has been assessment. The research bears out those concerns in one respect, yet has offered hope that direct writing assessment can be better understood and improved. Following Simon's (1981) suggestions for analyzing systems, atomizing the process of rating compositions may reveal the structure of assigning scores and suggest threats to validity and reliability. By studying how raters make decisions about writing quality and by documenting features that appear in the compositions they identify as good and as poor writing, some

information may emerge about what cues raters to make good and bad rating decisions.

CHAPTER 3

METHODS

Introduction

The study was designed to measure the ability of Maryland high school English teachers to rate accurately the quality of student compositions using modified holistic scoring procedures associated with the Maryland Writing Test. While the scoring criteria for the study were those included in the Maryland Writing Test scoring guide, it was of interest to determine the relationship of the teachers' rating behaviors to other characteristics of student writing as well. The impact of modified holistic scoring training on rating behavior was examined in the study, both in terms of its capacity to affect accuracy and the characteristics of student writing that may have related to scoring decisions.

The Maryland Writing Test provided an unusual opportunity to examine the rating behaviors of teachers with two levels of training in modified holistic scoring. With a clearly articulated set of generic rating criteria associated with the test and the availability of student compositions that were exemplary of the correct score assignments, it was possible to investigate some of the issues raised in the literature. Analyzing the ability of the teacher to apply the scoring criteria to compositions not only allowed a view of rater accuracy, but also allowed an opportunity to explore the influences on rater scoring decisions. To a limited extent, it was possible to separate out some composition characteristics that directly related to the Maryland Writing Test from others that did not relate. The method for the data collection required analysis of both the compositions and of the scoring decisions of raters.

Design of the Study

The study entailed the scoring of 35 student compositions by two sets of teachers who differed primarily in their level of training in modified holistic scoring of explanatory compositions. A third set of scores for the compositions consisted of certified correct scores and provided base-line data for comparison with the ratings assigned by the two groups of teachers. The compositions were analyzed for two sets of text-based characteristics--eight features that were not identified in the Maryland Writing Test scoring criteria and five that were. The scores assigned by the teachers to compositions were examined individually and by group in relationship to training and text features in an attempt to identify patterns in score assignments.

Population

All subjects in the study performed modified holistic scoring of student compositions using the Maryland Writing Test explanatory scoring criteria. However, the two groups of subjects differed in several ways. A questionnaire (Appendix A) was completed by novice and expert raters after they had scored compositions for the study. The results of the questionnaire are provided in Table 1 and confirmed that no novices had received training as Maryland Writing Test Scoring Committee members and that experts were all members of that committee. Further, information was collected regarding the inservice training, academic training, and experience with holistic scoring. The questionnaire indicated that, as a group, novices were generally less well-trained than experts and were less confident in their holistic scoring abilities.

Table 1

Summary of Results of Questionnaire on Rater Experience and Training

Response	Rater Responses (N=6) Novices	(N=8) Experts
Mean number of school-system inservices attended	3.50	4.60
Mean confidence in scoring (1=low; 2=avg.; 3=high)	2.00	2.88
Mean number of MSDE inservices attended	1.80	4.13
Percent serving as an MSDE Scoring Com. Mem.	0	100.00
Percent that attended other holistic scoring training	23.00	25.00
Percent that use holistic scoring with students	100.00	100.00
Percent that have MA degree or equivalent	50.00	85.71[a]
Percent that completed in-state English educ. program	66.67	28.57[a]

[a]N = 7 (one expert rater did not complete that portion of the questionnaire)

The fourteen study participants included four males and ten females. All six novice raters were white, five of whom were females. Expert raters consisted of two white males, one black male, three white females, and two black females.

Expert Raters

The eight teachers participating in the study who were considered to be experts were members of the Maryland Writing Test Scoring Committee, which is divided into two subcommittees of twelve members each. These teachers, after training in scoring both narrative and explanatory compositions from the 1984 and subsequent Maryland Writing Tests, convened as a subcommittee to determine by consensus the scores on a set of narrative compositions later to be used as anchor papers and training papers with contracted raters who scored the 1988 Maryland Writing Test. The eight participants in the study were teachers, whereas the remaining four were supervisors or administrators in local school systems. Their qualifying training in the scoring of explanatory compositions was limited to compositions from the 1984 Writing Test. As a consequence of their subcommittee task, this group of teachers never saw the explanatory compositions scored by their counterparts in the explanatory subcommittee and used in this study.

Novice Raters

The six novice raters participating in this study were teachers of English who taught in a Maryland high school in the 1987-88 school year. The high school that was used in the study was moderate in size, included an English department that was staffed by six veteran teachers, and was located in a suburban to rural area just outside of a major metropolitan area. The high school was selected for this study because it was possible to involve the entire English department in the study and because the English teachers had participated in a fall 1987 inservice training in modified holistic scoring. It was also important that the English department

be staffed by teachers who were certified by Maryland State Department of Education and who had achieved a 1988 passing rate in the Maryland Writing Test among ninth graders at or near the state average of 82.2%. The passing rate at the participating high school was just under 80%.

All six teachers had experienced from two to five brief local school system inservice programs on the modified holistic scoring of compositions and on the Maryland Writing Test scoring criteria and from one to five similar inservice programs conducted by Maryland State Department of Education. The most recent inservice program was conducted by the local school system in the fall of 1987 where recent revisions in the Maryland Functional Testing Program were reviewed (Maryland State Department of Education, 1987b). However, teachers were not considered to be expert raters because they had not been required to reach any set level of competency in scoring as a part of the inservice activity.

Five of the six teachers participating in the study as novice raters taught students in grades nine through twelve and were charged with the task of preparing students to take the Maryland Writing Test in January of 1988. Though the sixth teacher did not have any such students in her classes, she did use holistic scoring procedures in assessing student writing as did the other five teachers. In general, teachers of grade nine students have had the most contact with the Maryland Writing Test and modified holistic scoring. Tenth, eleventh, and twelfth grade teachers who instruct students failing the test have had the next highest contact with the criteria. The teachers participating in the study were the only ones in the English department at the high school.

To assure comparability of the teachers' scores with typical classroom writing assessment, teachers were provided with the Maryland Writing Test explanatory scoring criteria, but did not receive anchor papers to guide score assignments. A set of anchor papers is a collection of compositions drawn from the papers to be scored and which exemplify characteristics of papers that should receive each score point on the scale. While practical for large-scale holistic scoring where large numbers of compositions are scored, anchor papers are not generally useful for day-to-day classroom holistic scoring. However, the generic descriptions of papers at each score point are included in the Maryland Writing Test scoring criteria.

Other than having knowledge of the state writing criteria, Maryland teachers likely are similar in other respects to their counterparts in other states. In writing instruction outside of that associated with the Maryland Writing Test, teachers likely engage other self-identified criteria for writing quality in assessment of student work.

Rater Training

Expert Rater Training

Expert raters, as members of the Maryland Writing Test scoring committee, attended three meetings as a part of their function in assuring that the 1988 Maryland Writing Test anchor papers and training materials for contracted raters linked both with 1984 test standards and with the standards of Maryland English teachers (Maryland State Department of Education, 1988). In the first meeting, all committee members received training in the scoring of 1984 narrative and explanatory compositions. Committee members completed three qualifying rounds for each of

narrative and explanatory writing, in which they were required to score accurately at least 80% of the respective compositions.

The first rounds of scoring involved papers that were easier to rate in that papers were placed in sequence from low to high scores and were, in the judgment of assessment specialists, obvious examples of each score point. Following independent scoring, raters assembled to read each paper aloud and to discuss the correct scores. Assessment specialists facilitated the discussion and worked toward group consensus on scores. Successive scoring rounds involved more difficult tasks. The final round included a random sequence of papers judged to be more difficult to score.

The mean qualifying scores for the 24 member committee was 87.3% and 85% for the first two qualifying rounds for narrative scoring and 88.6% for the first explanatory round. All committee members qualified for narrative scoring, and all but two qualified for explanatory scoring. The final portion of the first committee meeting included a review of student compositions written to the 1988 prompts in order to allow raters to extend their evaluative skills to the current topics.

Before the scoring committee convened a second time, assessment specialists from Maryland State Department of Education, together with the project director and project coordinator from Measurement Incorporated, prescored 360 narrative and 360 explanatory compositions, all of which had been written by Maryland students for the January 1988 Maryland Writing Test. At the second scoring committee meeting, members met by domain, i.e., as an explanatory and a narrative subcommittee.

The task of each subcommittee in this second meeting was to score its respective compositions using the scoring guide or rubric and anchor papers, which had been written in 1986 to the 1984 prompts. The 1986 anchor papers had been scored using the current revision of the scoring guide. Compositions were identified for which an 80% consensus on the score assignment did not occur. Following group discussions about such papers, committee members rescored compositions until consensus on the score assignment was achieved.

At a third committee meeting, the committee selected compositions that would become part of the 1988 anchor sets for use by contracted raters at Measurement Incorporated in scoring the state test. During this meeting, the subcommittees scored additional compositions. Those papers upon which an 80% consensus agreement on the score was reached were identified as check papers. Check papers are scored periodically during the normal scoring activities of contracted raters as a means of determining if they have drifted from the rating criteria upon which they were trained.

Novice Rater Training

Novice raters attended a system-wide inservice training activity on September 1, 1987, which was attended by all English language arts teachers in the school system. Maryland State Department of Education had conducted an inservice activity in August of 1987 for supervisors of English language arts from each local school system. Supervisors received training in the revisions in the Maryland Writing Test so that they could return to their school systems and conduct similar local level inservice programs for teachers.

The revisions in the Maryland Writing Test resulted from a year-long project involving a task force that had been appointed by the State Superintendent of Schools. The task force had been charged with the responsibility of developing revisions that responded to suggestions made by an independent consultant (Educational Testing Service) that certain changes would improve the soundness of the state writing test. Those revisions included refinements in scoring rubrics, student checklists, and the addition of an indirect writing assessment component to the test. While the scoring revisions and the checklists were to be first used with the 1988 test, the multiple choice component was not to become a part of the test until 1989.

The local inservice training activity included a review of the changes as outlined in the Maryland State Department of Education brochure, "Revisions in the Maryland Functional Writing Program" (Maryland State Department of Education, 1987a). The teachers received copies of an instructional guide, "1987 Writing Supplement: Project Basic Instructional Guide" (Maryland State Department of Education, 1987d), which included extensive suggestions and resource materials for teaching writing as well as copies of anchor papers for the 1987 test, scored using the revised rubrics. The publication also included copies of the new student writing checklists and scoring rubrics.

The system-wide inservice training activity included a review of the rubrics where annotated, prescored compositions were used to exemplify how the rubrics were to be applied to the 1987 compositions. Teachers received suggestions on how to teach writing skills to students and directions on distribution of the revisions brochure to students. At no time

during the inservice activity were teachers required to score compositions to meet a qualifying level of 80% accuracy in the scoring of either narrative or explanatory compositions.

Instrumentation

Student Compositions

The primary instrument used in the study was a set of 35 explanatory compositions that were written by Maryland students during the January 1988 Maryland Writing Test administration. Compositions were photocopied, and personal identifying information was removed from the text of each student response. Such information included student name, school, names of school or community staff, names of political subdivisions, names of athletic teams, and addresses. This information was replaced with fictitious information in order to assure the continuity of text. The fictitious information was written onto the composition in a handwriting that approximated that of the original author. Further, the mechanical errors made by the original author, idiosyncratic characteristics of the writing, and actual placement of text in the original compositions were approximated in the fictitious insertions.

Certified Maryland Writing Test scores had been assigned to each composition by the Maryland Writing Test Scoring Committee, Explanatory Subcommittee, under the direction of a writing assessment specialist from the Maryland State Department of Education. These compositions were part of a group of papers scored as potential anchor papers and training papers that were used in the training of contracted raters working for Measurement Incorporated, the company scoring the 1988 Maryland Writing Test. The sequence of papers in each set was randomized to

decrease the impact of order effect on score assignments (Ferrara, 1987). Random sequences were generated from a random number table.

Questionnaire: Teacher Training and Experience

All raters were asked to answer a set of seven questions designed to collect information on experience with the Maryland Writing Test scoring criteria and professional training in modified holistic scoring. This data assisted the teachers in describing and documenting the level of experience and training with modified holistic scoring for each teacher. It also assisted in analyzing the possible causes for performance levels of outliers. A copy of the questionnaire is included in Appendix A.

Data Collection Procedures

Prior to the beginning of the study, the set of 35 student compositions to be used in the study was assembled. The certified Maryland Writing Test scores assigned to those compositions by the Maryland Writing Test Scoring Committee were provided by Maryland State Department of Education assessment specialists along with brief annotations regarding the significant characteristics of certain papers that might affect scores. The compositions were then submitted to direct quantitative and qualitative analyses to compile profiles of each composition. The analyses produced information about the appearance of features of writing that were associated with the scoring criteria and of features not associated with the criteria.

Direct Quantitative Assessment of Compositions

Research Question 2 sought to identify predictors of the score assignments made by experts and novices from among a list of eight features of writing that were not identified in the Maryland Writing Test

scoring guide. Seven of the eight features (except for handwriting ratings) were based on simple frequency counts. The features were identified and calculated as follows:

Composition length- the total number of words in the body of the composition. All words were counted including those obscured by poor handwriting.

Sentence complexity - the degree to which writing, on the sentence level, displays subordination of ideas and thoughts through the use of phrases and clauses. Sentence complexity was measured by counting the mean number of words per t-unit for a composition. A very complex sentence may contain more than one t-unit. In this study, independent clauses were considered to be t-units where they appeared as complete sentences or as parts of compound sentences. Hunt (1977, p. 92), provided the simple guidelines for t-unit identification, which he used in several of his own studies.

Punctuation errors- incorrect placement of punctuation as well as omissions. Not only were punctuation errors associated with sentence construction counted, but also errors in possessive forms of words, quotations, listings, and other similar parts of sentences were included. However, punctuation errors associated with syntactic structures such as run-on sentences and sentence fragments were not counted as punctuation errors but were considered to be syntax errors.

Spelling errors- incorrect spelling of any words appearing anywhere in the composition. Spelling errors were determined by counting the total number of words that were misspelled in each composition, even if such errors represented repeated misspellings for

any one word. Misspelled proper names were also included if such names were capable of being verified. However, most proper names and other identifying information had to be changed on the compositions for the study. Consequently, errors in the names of students, teachers, or others mentioned in the compositions were not counted if they represented fictitious names altered by the researcher, even though such fictitious names were changed to emulate the original errors where they appeared. Errors in possessive forms of words were considered to be usage errors and were not counted as spelling errors. Likewise, capitalization errors were separated out from spelling errors and are defined independently.

Capitalization errors- improperly capitalized words and omissions. Capitalization omissions appearing at the beginnings of sentences are usually tied to sentence construction problems and were not counted as capitalization errors, but were counted as syntax errors. Errors involving the capitalization of names and other terms usually do not interfere with sentence meaning, but were counted even though they represent a category of capitalization errors usually not included in other studies. However, in this study, it was of interest to determine if the appearance of such errors were related to raters' scoring decisions.

Syntax errors- errors involving sentence construction. An error was considered a sentence construction problem if it resulted in an incomplete sentence, a run-on sentence, or a mis-placed modifier.

Language usage errors- errors involving grammatical features such as verb tense agreement, subject-verb agreement, pronoun antecedents, and clause and phrase structures were considered language usage errors

for this study. Usage errors included improper use of words and omission of words.

To accomplish this direct qualitative assessment of writing, it was necessary to prepare clear, readable copies of each composition. As a consequence, the text of each composition was typed, including fictitious changes in names and locations and including all errors. Typed compositions (Appendix B) were reviewed by two readers for sentence complexity, composition length, and error frequencies. A third reader reviewed all compositions for syntax and usage errors to confirm counts in those two areas. From that information, the mean t-unit length and the frequency with which each type of error appeared per 100 words was calculated in order to allow a comparison between compositions.

Because compositions varied in length from 41 to 711 words, it was necessary to transpose all raw counts of features into standardized figures to facilitate statistical comparisons. For each composition, the number of t-units was divided into the number of words to produce a mean t-unit length. The number of words for each composition was divided by the frequency counts for errors in spelling, punctuation, capitalization, usage, and syntax. That figure was divided by 100 to produce the frequency in occurrences per 100 words. The direct quantitative assessments for each composition became part of a composition profile, which provided a summary of raw data on each composition. This profile facilitated the analysis of results by placing all data on one composition on a reference form. The results of the direct quantitative assessments of writing are included in Table 2.

Table 2

Features of Compositions

Feature	M	SD	Range	Min.	Max
No. of Words	230.66	130.32	670.	41.	711.
M t-unit length	14.39	3.33	13.88	8.5	22.38
ERRORS per 100 words					
spelling	2.57	3.97	20.12	0	20.12
punctuation	3.07	1.63	6.94	.51	7.45
capitalization	1.58	2.36	12.27	0	12.27
usage	1.46	.197	4.	0	4.
syntax	1.79	1.727	5.92	0	5.92
HANDWRITING RATINGS					
experts	2.18	.47	2.	1.	3.
novices	2.08	.49	1.67	1.17	2.83

The compositions used in the study were selected to represent all four score points, with an attempt to approach a normal distribution. This produced negatively skewed normal distributions for mean t-unit length and composition length. However, the distributions for five of the features began with a substantial number of compositions having counts at or near zero, followed by declining numbers of compositions across the range of frequencies. There were similar patterns for distributions of errors in syntax, spelling, punctuation, capitalization, and usage.

Direct Analytic Assessment of Compositions

The third research question sought predictors of scores assigned by novices and experts from among the five features identified in the Maryland Writing Test scoring guide. The five features--content, organization, attention to audience, sentence formation errors that interfere with meaning, and conventions errors that interfere with meaning--were evaluated directly through direct analytic assessment. An analytic scoring matrix (Appendix A) had been published in the fall of 1987 by the Maryland State Department of Education, using the four-point scale and language of the Maryland Writing Test scoring guide. Because of its adherence to the scoring criteria in the modified holistic scoring rubric, the analytic matrix was judged suitable for use in this study.

Typed copies of compositions, identical to those used by the three readers above, were read by three specialists for the five characteristics of writing identified in the Maryland Writing Test Explanatory Scoring Guide. Two readers were writing instruction specialists working with the Maryland State Department of Education and who had participated in the 1988 Maryland Writing Test scoring committee as observers, completing all activities required of the 24 scoring committee members. A third specialist was a Maryland English teacher who had been a member of the Maryland Writing Test scoring committee, explanatory subcommittee for several years, including 1988.

The analytic scores assigned by the three readers were guided by the analytic scoring matrix. Readers were provided with the certified Maryland Writing Test score for each composition and were asked to assign an analytic score for each of the five features. Two features--

content and organization--could be scored with 1,2,3, or 4 scores. The analytic matrix, using language from the modified holistic scoring guide, indicated a dichotomous scoring decision for attention to audience--the writer either attended to appropriate audience, or attended to the wrong audience. As a consequence, ratings of 1 or 4 were appropriate. The remaining two features--errors in sentence formation or conventions--were judged to affect meaning appreciably, minimally, or not at all. Thus, these two features were rated with scores of 1,2, or 4.

To facilitate statistical analysis, the raw scores for audience were converted to 0 and 1, and the scores for errors were converted to 1,2, and 3. Because of the high skill level of the three raters, their individual scores were averaged for each composition to produce a mean analytic rating for each of the five features. Means were then entered into stepwise regressions with certified Maryland Writing Test scores, and the scores assigned by novice and expert raters. Table 3 provides a description of the five variables.

Table 3

Analytic Scores for Five Features of Writing Identified in the MWT Scoring Guide

Feature	M	SD	Range	Min.	Max.
Content	2.629	.945	3	1	4
Organization	2.61	.913	3	1	4
Attn. to Audience	1	0	0	1	1
Errors- Sen. Formation	2.648	.518	2	1	3
Errors- Conventions	2.715	.466	2	1	3

Data Profiles

A data profile for each composition in the set was compiled to collate all of the information in the quantitative direct assessment, the qualitative analytic assessments, the certified Maryland Writing Test scores and the responses of expert and novice raters. Each data profile (Appendix B) included the following:

a. certified Maryland Writing Test scores

b. modified holistic scores assigned by each expert rater and by each novice rater

c. handwriting quality ratings assigned by each expert rater and by each novice rater

d. number of words

e. mean number of t-units per total number of words

h. error counts

1. number of times words were misspelled per 100 words
2. number of capitalization errors per 100 words.
3. number of punctuation errors per 100 words.
4. number of language usage errors per 100 words.
5. number of syntax errors per 100 words.

i. mean analytic ratings of the five features from the Maryland Writing Test scoring criteria:

1. content rating (analytic scores of 1,2,3, or 4)
2. organizational plan rating (analytic scores of 1,2,3, or 4)
3. degree to which writer addressed audience (converted analytic scores of 0 or 1)
4. degree to which sentence formation errors interfered with meaning (converted analytic scores of 1,2, or 3)
5. degree to which conventions errors interfered with meaning (converted analytic scores of 1, 2, or 3)

Collection of Teacher Responses

Each set of compositions included all 35 papers, but was placed in random order to assure that no two subjects received the same sequence of papers. Prior papers in a packet of compositions can influence the rater's expectations, thus affecting the scores (Daly & Dickson-Markham, 1982; Hales & Tokar, 1975). Though randomization of papers does not eliminate this phenomenon, it does statistically reduce its impact on score assignments (Ferrara, 1987). In this study, each teacher handled two sets of the same papers--one for modified holistic scoring and another for rating handwriting. Randomization also assisted in offsetting rater bias by assuring that any one teacher did not view the same

sequence of compositions in completing both modified holistic ratings and handwriting assessments.

Expert raters, since they are scattered across the state, were mailed instructions and scoring materials for completion. Packets included:

1. instructions for assigning modified holistic scores (Appendix A)
2. the 1988 Maryland Writing Test explanatory writing prompt (Appendix A)
3. the Maryland Writing Test explanatory scoring criteria (Appendix A)
4. one set of explanatory compositions for scoring
5. instructions for rating handwriting quality (Appendix A)
6. a brief questionnaire (Appendix A)

Expert raters were asked to complete all activities in one session as follows:

1. Subjects were to remove the set of materials from the first envelope and study the explanatory writing prompt and Maryland Writing Test explanatory scoring criteria. They were then asked to assign modified holistic scores to each composition on the basis of the criteria, writing and circling scores directly on each composition.

2. Following the modified holistic scoring of papers, subjects were asked to remove the set of compositions from the second envelope and rate handwriting quality for each paper using a scale of 1 to 3 with 3 representing the highest quality of writing. See Appendix A for instructions in completing handwriting ratings.

3. Finally, subjects were asked to complete a brief questionnaire (Appendix A) and to mail the packet back to the researcher. In order to

assure anonymity for all subjects, each was asked to identify all materials only with his or her birth month and mother's maiden name.

The novice raters completed the same sequence of activities at the participating high school. The researcher provided each teacher with a packet of materials containing two envelopes of compositions. Included in each envelope were the same instructions, scoring materials, and randomized composition packets as were used by the expert raters. The researcher read the initial set of instructions to the participants, was present during the session, and did not provide scoring assistance, but clarified procedures as requested.

Substitute teachers were scheduled in order to allow teachers to complete all tasks during the school day. Two of the teachers completed the the tasks in an early morning session, and three teachers worked in the late morning. Both work sessions took approximately one and one-half hours. The sixth teacher was absent on the day data were collected and completed the tasks independently, mailing the packet of materials to the researcher several days later.

Method of Data Analysis

Research Question 1

What is the relationship between certified modified holistic scores, scores assigned to compositions by expert raters undergoing complete scoring training several months prior to scoring, and scores assigned to compositions by novice raters who have experienced only brief inservice programs in modified holistic scoring within the last year?

Preparation of Data for Research Question 1.

In Research Question 1, the relationship of the modified holistic scores to each other was examined in three ways to determine the extent to which there were significant differences in the accuracy of score assignments, the degree to which there was a relationship in the ways scores were assigned by groups, the degree to which there was a relationship in the ways errors occurred in each group, and how accurate each group was in assigning scores. The score assignments from the expert raters were averaged for each composition, as were the score assignments from novice raters. These two procedures produced for each composition a mean score for expert raters and a mean score for novice raters. The certified modified holistic score for each composition comprised the third score for comparison with the teacher scores.

The scores assigned to each composition were analyzed to determine the level of inaccuracy in each rater's score assignment. A mean error size was calculated from the score assigned by each rater. This value represented the mean number of score points rater-assigned scores deviated from the certified Maryland Writing Test score for each composition. The means and standard deviations were reported for the two sets of error values. This procedure assisted in the comparison of novice and expert raters by group and by individual rater.

Analysis of Data for Research Question 1

Research Question 1 a asked to what extent the scores in each set of compositions correlated with the other. This level was also determined using a Pearson Product Moment Correlation. The correlation of the two sets of scores helped to determine the extent to which both sets of raters

had applied similar criteria. Further, the correlation of both sets of scores to the certified Maryland Writing Test scores assisted in determining if the criteria used by each group of raters matched the criteria that should have been used. Finally, a correlation of the certified Maryland Writing Test scores and the scores assigned by each rater was also completed as a part of this analysis in order to determine the extent to which each rater was using the Maryland Writing Test scoring criteria. A Fisher z transformation was used to determine if either of the two groups of rater scores correlated at a significantly higher level with certified scores.

Research Question 1 b asked for the extent to which the error size for the scores assigned by the two groups of raters correlate. This level was determined by performing a Pearson Product Moment Correlation with the same two sets of mean error sizes. This procedure assisted in determining the extent to which novice raters and expert raters made similar errors in score assignments.

Research Question 1 c asked for the mean difference in the number of accurate scores assigned by the two groups of raters. Consequently, a two-factor analysis of variance (ANOVA) was used to determine if the mean percent of accurate scores for novice raters was significantly different from the mean level of errors for expert raters at the extreme score points (1 and 4) and at the middle score points (2 and 3).

Research Question 1 d asked if there was a mean difference in the size of errors made by experts and novices. A two-factor analysis of variance (ANOVA) was used to determine if the mean error size for novices was significantly different from the mean error size for experts at the extreme score points (1 and 4) and in the middle score range (2 and

3). An analysis of errors was performed by rater and by group in order to develop a breakdown of error frequencies. The percents of compositions at each score point (1,2,3, and 4) for which raters assigned incorrect scores were determined by rater and by group. The analysis included all four components.

Research Question 1 e asked if there was a mean difference in the direction of errors made by experts and novices. Inaccurate scores were analyzed to determine the percent of times rater score assignments were one, two, and three points below the correct score, how many times rater score assignments were correct, and how many times raters were one score point above the correct score. The data from the error direction analysis for each rater were reduced to determine the number of times raters had scored papers higher and lower than the certified Maryland Writing Test scores.

Research Question 2

Which of the following characteristics or combination of characteristics of writing are predictors of the certified modified holistic scores, scores assigned to compositions by expert raters undergoing complete training several months prior to scoring, and scores assigned by novice raters who have experienced inservice programs in modified holistic scoring within the last year?

a. sentence complexity (as indicated by mean t-unit length)

b. number of words per composition

c. mean raters' assessments of composition handwriting quality

d. number of punctuation errors per 100 words

e. number of spelling errors per 100 words

f. number of capitalization errors per 100 words

g. language usage errors per 100 words

h. number of syntax errors per 100 words

Analysis of the Data for Research Question 2

All of the above writing characteristics, with the exception of handwriting perception ratings, were collected and verified through counts by three readers. Handwriting perceptions were assigned by both novice and expert raters, and the mean rating for each composition was calculated for the two groups. All of these characteristics of writing were entered into a stepwise multiple regression to determine those that were most successful in predicting the certified modified holistic scores, expert raters' scores, and novice raters' scores.

Research Question 3

Which of the following analytic ratings of writing features from the Maryland Writing Test scoring criteria are predictors of certified modified holistic scores, scores assigned by expert raters undergoing complete training several months prior to scoring, and scores assigned by raters undergoing inservice programs in modified holistic scoring within the last year?

a. organization

b. content

c. attention to audience

d. sentence formation errors that interfered with meaning

e. mechanical errors that interfered with meaning

Analysis of the Data for Research Question 3.

In question 3, a panel of three specialists assigned analytic scores (1 to 4) to each composition in the experimental set for each of the five features identified in the Maryland Writing Test scoring criteria. Stepwise regressions were conducted using those five scores as in 2 a through 2 h.

Research question 2 primarily involved errors, the raters' judgments of the quality of the student handwriting on each composition, and the length of the composition--characteristics not specifically identified as priorities in the Maryland Writing Test scoring criteria. Research question 3 involved the five characteristics identified in the scoring criteria. The two questions differ in that the third one sought characteristics which were expected to influence score assignments if raters had applied the Maryland Writing Test scoring criteria. Conversely, the second question sought other factors which might influence score assignments if raters had not used the state test criteria in making scoring decisions.

Summary

A set of 35 student compositions were selected from 1988 Maryland Writing Test explanatory responses, each of which had pre-assigned certified modified holistic scores. These compositions were scored by two groups of teachers, one having completed an intensive training program resulting in their being designated as certified raters and another having completed a different training program with no such qualifying requirement. The compositions were analyzed for eight characteristics that had not been a part of the Maryland Writing Test scoring criteria and five characteristics that appear in the scoring criteria. Both rater accuracy

and the relationship of rater scores to the above two sets of writing characteristics were analyzed for patterns in rater decisions.

CHAPTER 4

RESULTS

Introduction

The central purpose of the study was to examine the ability of raters to assign accurate scores to compositions. The first task was to determine how often raters could assign scores that agreed with the certified scores each composition had received in the Maryland Writing Test. However, simple accuracy counts provided a limited picture of rater behaviors. It was necessary to determine the characteristics of compositions that had been scored correctly as well as those that had not and to analyze scoring patterns in light of the appearance of those characteristics. The characteristics included two clusters of writing characteristics--one group of features that were part of the test scoring criteria and another that had been investigated previously and found to infuence raters, yet were not part of the criteria.

The results of the study are provided in this chapter with data tables and narratives. More detailed data have been provided in Appendices B through D. Results reveal information about rater behavior both by group and individually. Further, where appropriate, scores and errors are compared by score point to analyze the manner in which raters applied the scale to compositions.

Research Question 1

Research Question 1 compared the scores assigned by novice and expert raters to determine differences and relationships in scoring decisions for the two groups. Correlations of mean scores and error sizes helped to determine if raters approximated certified scores and if they

experienced difficulty with the same compositions. The analysis also included three different comparisons of novices and raters, including their accuracy and error levels as well as the direction of their errors. Consequently, five dimensions of accuracy were examined. They were:

a. the correlational relationship of the scores assigned by the two groups of raters to the certified Maryland Writing Test scores.
b. the correlational relationship between the sizes of the errors made by the two groups of raters in terms of the number of points away from the certified Maryland Writing Test scores.
c. the mean difference in the frequency with which raters assigned scores that agreed with the certified Maryland Writing Test scores.
d. the mean difference in the size of the errors made by the two groups of raters.
e. the mean difference in the direction of errors, i.e., whether raters scored compositions higher or lower than the certified Maryland Writing Test scores.

Research Question 1 a

Research Question 1 a sought to determine to what extent certified Maryland Writing Test scores correlated with scores assigned by expert raters and with the scores assigned by novices. It was also of importance to determine if the correlation between novice and expert scores was higher than the correlations of either of the two sets of scores with certified scores. A statistical comparison of the resulting coefficients for each of the three correlations determined if any of the coefficients was significantly different from the other two. Scores assigned by novice raters and expert

raters were correlated with each other and with certified Maryland Writing Test scores using a Pearson Product Moment Correlation. The results of the correlations are included in Table 4 and indicate a strong relationship between expert and novice scores.

Table 4

Results of Correlation: Expert and Novice Rater Scores and Certified Maryland Writing Test Scores

	Correlation Coefficients		
Scores	Cert. MWT	Novices	Experts
Cert. MWT	1.000	.898	.882
Novices		1.000	.906
Experts			1.000

The value needed for r to be significant was .335; therefore the correlation between each set of scores was significant. Fisher's z transformation showed that there was no significant difference between any of the correlation coefficients. This result indicated that the errors made by novices and experts were not sufficiently different to displace the relationship in the three sets of score assignments.

The relationship between novice and expert rater scores somewhat exceeded that of either of the two score sets with certified Maryland Writing Test scores. Because novice and expert rater scores both included errors, the slightly higher correlation between the two sets of rater scores

indicated a possibility of similarities in error patterns, which were capable of closer scrutiny through an examination of the relationship of the size and direction of the errors made by the two groups of raters.

Research Question 1 b

The purpose of Research Question 1 b was to examine such error patterns to determine if there was a correlational relationship in the size of errors made by the two sets of raters. The mean error sizes for novice and expert raters were used to answer this question in that these values provided a quantitative assessment of the group error for each composition. The mean error sizes for novice (.419) and expert (.375) raters were correlated using the Pearson Product Moment Correlation. The results indicated a moderate positive relationship ($r = .372$) in the errors made by the two groups of raters. However, the relationship was considerably lower than the .906 correlation observed with the two sets of rater scores.

Since the value needed for r to be significant was .335, the mean error sizes for novice raters and expert raters significantly correlated with each other. The results of the correlation indicated that the size of the errors made by novices had a small significant relationship to the size of the errors made by expert raters when compared by composition.

Research Question 1 c

Research Question 1 c compared the abilities of the two groups of raters to assign scores that agreed with the certified Maryland Writing Test scores. Scores that agreed with certified scores were termed accurate scores. Mean percentages of scoring accuracy at each score point and for the total set of compositions were compared to answer Research Question

1 c. The distribution of percents of rater accuracy by score point are included in Table 5. This data indicated that novices were more accurate in scoring compositions at the extreme ends of the scale while experts were more accurate in the middle score points.

Table 5

Percent of Accurate Scores Assigned to Compositions at Each Score Point

Score Point	Mean Percent Correct		Number of Compositions Per Score Point
	Novices	Experts	
1	70.8%	50.0%	4
2	56.9	69.8	12
3	47.2	50.0	12
4	81.0	76.8	7
Total	60.0	62.9	35

A visual inspection of the data indicated a difference in rater performance at the extreme score points when compared with performance at the middle score points. In a statistical analysis of that difference, percents of rater accuracy were compared in a two-factor analysis of variance with rater training level and score range forming the two categories. Novice and expert rater percents of accuracy were compared between compositions at the extreme score points--score points

1 and 4--with accuracies at the middle range--score points 2 and 3. The results are included in Table 6.

Table 6

ANOVA Table for the Two-Factor Analysis of Variance on Percent of Accurate Scores

Source of Variation	df	SS	MS	F	p
Rater	1	.001	.001	.009	.9238
Score	1	.369	.369	4.916	.0301
Interaction	1	.137	.137	1.826	.1812
Error	66	4.954	.075		

The result indicated no significant interaction between the rater training and score range. However, there was a significant difference between the performance of all raters at the extreme and middle score ranges, as shown in Table 6. It was useful to group papers at score points 2 and 3 together for comparison because these compositions were in the middle of the range of scores and were the only ones on which raters could have erred by scoring papers higher or lower than the certified Maryland Writing Test scores. Further, in actual assessment situations, a score of 2 could be considered as unsatisfactory progress or failing while a score of 3 could be satisfactory. Ability to discriminate 2 papers from 3 papers with a four-point scale are implicit in providing students feedback on progress. Table 7 shows that while novices scored correctly an

average of 52.1% of the papers at score points 2 and 3, experts posted an accuracy rate of 60.0% for the same compositions. At the extreme score points, novices were more accurate than experts by more than ten percent.

Table 7

Rater Training-Score Range Incidence on Percent of Accurate Scores

	Score Range					
	Extreme (1 & 4)		Middle (2 & 3)		Total (1 to 4)	
Raters	N	$\bar{X}$	N	$\bar{X}$	N	$\bar{X}$
Novices	11	77.3%	24	52.1%	35	60.0%
Experts	11	67.0	24	60.9	35	62.9
Totals	22	72.2	48	56.9	70	61.4

Though not significant, the accuracy of novices was greater only at the extreme score points (1 and 4). The accuracy of experts was higher in the middle score point range (2 and 3) where bipolar directionality of errors was possible and where the largest portion of the certified Maryland Writing Test (MWT) scores fell for the papers used in this study.

Research Question 1 d

In Research Question 1 d, the errors in score assignments made by raters were compared at the middle and extreme ends of the scoring range using an analysis of variance to determine if there was a mean difference in the size of novice and expert rater errors. Whereas Research

Question 1 c was limited to comparing the rate with which raters agreed with certified Maryland Writing Test scores, this question provided an indication of how far novice and expert scores deviated from certified scores. This was accomplished by performing an analysis of variance (ANOVA) using the mean error sizes. The mean error size reflected the mean number of points rater scores were from the certified Maryland Writing Test scores for each composition. The mean error sizes for novices and experts are compared in Table 8. For papers at score points 1 and 3 the mean error size for experts exceeds that of novices. At the other score points, novices displayed larger mean error sizes.

Table 8

Comparison of Group Mean Error Size for Novice and Expert Raters

Score Point	Novices		Experts	
	$\bar{X}$	SD	$\bar{X}$	SD
1	.292	.25	.5	.228
2	.430	.297	.312	.235
3	.442	.349	.479	.212
4	.262	.345	.232	.318
Total	.419	.327	.375	.257

When the means were compared in an analysis of variance similar to that used in Research Question 1 c, the difference at the middle and extreme score ranges proved not to be significant at the $p < .05$ level.

Likewise, the performance of novices was not significantly different than that of experts, and there was no significant interaction of rater training and score range factors. Experts, then, were not significantly more accurate than novices when the mean size of the errors made by the two groups were compared, though novices made slightly larger errors.

Research Question 1 e

While the mean error sizes of novices and experts were not significantly different, this comparison provided some information on how much difference there was in errors made by the two groups. Research Question 1 e further examined the errors made by the two groups of raters to determine whether their incorrect scores were more often higher or lower than certified Maryland Writing Test scores.

When compositions were examined for gross patterns, novice rater errors fell below the certified Maryland Writing Test score 70.24% of the time. This contrasted sharply with the errors made by expert raters, whose wrong scores were higher than the true score 50% of the time. Table 9 shows that this overall pattern was reflected at score points 2 and 3.

Table 9

Error Direction at Score Points 1 Through 4

	Novice Errors		Expert Errors	
Score Points	No.	% of Total	No.	% of Total
Score Point 1				
errors above	7	100.00	16	100.00
total	7		16	
Score Point 2				
errors above	8	25.81	13	44.83
errors below	23	74.19	16	55.17
total	31		29	
Score Point 3				
errors above	10	26.32	23	52.08
errors below	28	73.78	23	47.92
total	38		48	
Score Point 4				
errors below	8	100.00	13	100.00
total	8		13	
Overall				
errors above	25	29.76	52	50.00
errors below	59	70.24	52	50.00
total	84		104	

There were some within-group differences in error direction patterns as displayed in Appendix C. One novice rater scored two compositions two points lower than the certified score and one composition three points lower. One expert rater assigned a 4 score to a "2" paper, scoring correctly all eleven remaining "2" papers. Otherwise, all errors made by all raters in this study were one point from the certified Maryland Writing Test score. Combining the papers at score points 2 and 3 revealed that two-thirds of the novice raters and three-fourths of the expert raters repeated error patterns that were similar to that of other members of their group. Though there were differences in the primary direction of errors from rater to rater, those who placed low scores on papers that should have received a 2 generally scored 3 papers low as well.

The performance of individual raters was examined to identify error patterns for outliers. These analyses are included in Appendix C. The most accurate expert raters primarily differed from the least accurate expert rater only in an ability to score papers at score point 2 more accurately. The least accurate novice demonstrated the same degree of inaccuracy at score point 3. The remaining raters in both groups demonstrated considerable variability in error patterns.

Research Question 2

While the first research question examined several dimensions of rater accuracy, Research Question 2 involved eight features of writing that had been shown in previous studies to predict quality ratings for student compositions. It was of interest to determine the degree to which the eight features served as predictors of certified Maryland Writing Test scores and

scores assigned by novice and expert raters. The eight features were grouped together for a stepwise regression because they were not identified in the scoring criteria for the Maryland Writing Test scoring guide for explanatory writing. The results of the regression, included in Table 10, show that the length of compositions and frequency of syntax errors were the strongest predictors of scores for all three groups. However, the strength of the predictors varied for each of the score sets.

Table 10

Results of Stepwise Regressions: Eight Features of Writing Not in the MWT Scoring Guide and Modified Holistic Scores

		Scores							
		Cert. MWT			Novices			Experts	
Step	Feature	r	F	p	r	F	p	r	F	p
1	Length	.639	22.725	.0001	.709	33.348	.0001	.74	39.85	.0001
2	Syntax	.703	15.591	.0001	.825	34.212	.0001	.81	31.4	.0001
(Variables with F values not large enough to enter regression equation)										
	Spelling.	.314*	3.385		-.182*	1.059		-.201*	1.311	
	Punct.	-.3*	3.064		-.237*	1.852		-.187*	1.126	
	Cap.	-.161*	.821		-.287*	2.776		-.177*	1.007	
	Usage	-.118*	.439		-.152*	.736		-.033*	.034	
	T-length	.014*	.006		-.052*	.083		.173*	.954	
	Handwriting				-.101*	.32		-.076*	.179	

*partial r showing incremental contribution to length and syntax

From the stepwise regressions, it was evident that syntax errors contributed at a significantly lower level than length to the regression equations. The adjusted variances resulting from the regressions are

included in Table 11 and provide a measure of the extent to which each of the unique variances of each of the features contributed to the overall cumulative variance, which includes the variance of all features in that step of the regression equation. Variances were adjusted to compensate for the small number of compositions used in the study.

Table 11

Predictors of Modified Holistic Scores Resulting From Stepwise Regressions of Eight Features of Writing Not in the MWT Scoring Guide

		Adjusted Variance					
		Cert. MWT		Novices		Experts	
Step	Feature	unique	cumulative	unique	cumulative	unique	cumulative
1	Length	.39	.39	.49	.49	.53	.53
2	Syntax Er.	.07	.46	.17	.66	.11	.64

For certified Maryland Writing Test scores as well as the scores assigned by novice and expert raters, composition length is the largest single predictor of scores. While nearly 40% of the variance in the certified Maryland Writing Test scores was predicted by the number of words in each composition, the variance was respectively ten and fourteen percent less than for novice and expert rater scores. At the same time, the unique variance contributed by the frequency of syntax errors to the regression

equation for certified scores was considerably lower than for novices or experts. It was of note that no other variables appeared in group results.

Stepwise regressions for the same eight features were also conducted to determine the extent to which individual raters' scores were predicted by the features. Group rater handwriting ratings were replaced in each regression by the handwriting ratings assigned to compositions by the individual rater. The results of those regressions are included in Table 12; they show that seven of the raters--four experts and three novices--had assigned scores that were capable of being predicted through length and syntax error frequencies in that order. The table shows a large number of raters (seven) for whom two-step regressions matching the group regression did not occur. For two of the raters, a three-step regression resulted with handwriting and mean t-unit length appearing as weak, but significant features. There appeared to be no relationship between the number of errors in score assignments and the predictors.

Table 12

Results of Stepwise Regressions: Non-MWT Features Against Individual Rater Scores

		Step 1		Step 2		Step 3		Total
Rater	Er.	feature	Adj. r^2	feature	Adj. r^2	feature	Adj. r^2	Adj. r^2
ER1	9	length	.39	syntax errors	.09			.48
ER8	11	length	.44	syntax errors	.11			.55
ER3	13	length	.39	syntax errors	.09			.48
ER5	13	length	.55	syntax errors	.11			.66
NR3	14	length	.32	syntax errors	.09			.41
NR4	14	length	.44	syntax errors	.09			.53
NR5	16	length	.41	syntax errors	.12			.53
Outliers								
NR1	9	syntax er.	.46	length	.12			.58
ER6	13	syntax er.	.37	length	.10			.47
NR2	14	syntax er.	.50	length	.08			.58
ER7	14	length	.54					.54
ER2	15	length	.28	handwriting	.11			.39
ER4	16	length	.46	syntax errors	.12	t-units	.04	.62
NR6	17	syntax er.	.36	length	.12	handwriting	.05	.53

The lack of distinction between novice and expert raters is evident in the individual regressions. The outliers in both groups of raters had no commonalties at either extreme. While Novice Rater 1 (NR1) posted the lowest number of errors, the combination of composition length and syntax errors was reversed in the regression. In contrast, Expert Rater 1 (ER1), who also made only nine errors, assigned scores that were predicted by length and syntax errors only, with length entering the regression equation first, and with a moderate level of shared variance. In fact, the regression for Novice Rater 5 (NR5) was very similar to that of ER1 even though NR5 had attained the higher error count. While the results of the individual stepwise regressions support the importance of composition length as a composition predictor, the results for outliers point out the lack of consistent patterns in score assignments among raters.

Research Question 3

The third research question examined the five features of writing that were identified in the Maryland Writing Test scoring guide to determine to what extent they served as predictors of scores. One of the features, attention to audience, was not included in the regression because specialists judged that all 35 compositions should receive the same rating (1 after conversion), indicating that all writers had attempted to address the proper audience. Table 13 includes the results of the stepwise regressions and show that of the four remaining features, two--content and organization--correlated highly enough with certified Maryland Writing Test scores to serve as predictors. From the stepwise regressions, the predictors of the certified Maryland Writing Test scores

were content and organization. The single predictor for scores assigned by novice and expert raters was organization.

Table 13

Results of Group Stepwise Regressions: Four Features of Writing Included in the MWT Scoring Guide and Modified Holistic Scores

	Scores								
	Cert. MWT			Novices			Experts		
Step Feature	r	F	p	r	F	p	r	F	p
1 Content	.996	92.341	.0001						
2 Organ.	.997	10.78	.0001						
1 Organization				.915	169.987.	0001	.895	132.496	.0001
(Variables with F values not large enough to enter regression equation)									
Content				-.082*	.219		-.062*	.123	
Sen. form.	-.025*	.019		.133*	.567		-.085*	.233	
Conven.	-.023*	.017		.061*	.121		-.067*	.146	

*partial r values showing incremental contribution to variable(s) entering equation

The adjusted variances resulting from the regressions are included in Table 14. They indicate that certified Maryland Writing Test scores were predicted by analytic scores for content with organization contributing only 0.1% to the cumulative adjusted variance. Novice and expert scores, in contrast, were very strongly related to organization scores.

Table 14

Predictors of Modified Holistic Scores Resulting From Stepwise Regressions of Features Included in the MWT Scoring Guide

		Adjusted Variance			
		Cert. MWT		Novices	Experts
Step	Feature	unique	cumulative	unique cumulative	unique cumulative
1	Content	.993	.993		
2	Organization	.001	.994		
1	Organization		.833	.833 .795	.795

For both sets of raters, group data resulted in the content rating variable dropping from the equation, resulting in organization as the single predictor.

Analysis of Individual Rater Data

Stepwise regressions were also performed using the scores assigned by each individual rater as a means of identifying within-group

differences. Results from regressions are included in Table 15, showing that the single-step equations resulting from regressions with group data were reflected in the data from all but two of the raters. Organization accounted for from 54% to 82% of the variance in rater scores. This table compares with Table 12 in which the results of the stepwise regression of scores with eight features outside of the scoring criteria were displayed. In that series of regressions with individual scores, half of the raters were outliers in that their regressions produced different predictors than were produced by the regression with group scores. In the results of the individual regressions with features that were identified in the scoring criteria, two raters deviated from the group results.

Table 15

Results of Individual Stepwise Regressions: Adjusted Variance for Individual Rater Scores vs. Analytic Scores

			Adjusted Variance	
Rater	Errors	Steps	Content	Organization
NR1	9	1		.82
ER1	9	1		.77
ER8	11	1		.73
ER5	13	1		.69
ER6	13	1		.71
NR2	14	1		.68
NR4	14	1		.59
ER7	14	1		.54
ER2	15	1		.59
ER4	16	1		.59
NR5	16	1		.64
NR6	17	1		.70
Outliers				
NR3	14	1	.45	
ER3	13	1	.68	

Though content analytic ratings predicted the scores for two outliers, the variance for one (NR3) was the lowest in both sets of raters.

The scores of the novice rater posting the lowest number of errors also had the highest level of variance with organization analytic scores for both sets of raters. Further, the scores of Expert Rater 1 (ER1), who had the lowest number of errors for that group of raters also posted the highest variance for experts, again with organization analytic scores.

The results of these regressions with individual rater scores contrasted sharply with those involving the eight non-Maryland Writing Test scoring features. In Research Question 2, the error frequency for raters had appeared to have little relationship to those regression results in terms of the variance in scores. However, with this research question, a ranking of raters by error frequency showed that, in general, raters with fewer errors posted higher variance levels with organization analytic scores. The most glaring exception was Novice Rater 6 (NR6). The non-Maryland Writing Test features, even length, never approached the variance levels observed in either group or individual regressions for the analytic scores involving the four Maryland Writing Test features.

Summary of Results

The study provided a multidimensional view of rating behaviors of expert and novice raters. It indicated some similarities in rating patterns, which emerged despite considerable differences in the kinds of training that the two groups of participants had experienced.

1. Novice and expert raters both assigned scores that correlated at nearly the .90 level with the certified Maryland Writing Test scores that compositions in the study actually received. There was no significant difference between the correlations of the rating groups.

2. When raters made errors on compositions, there was a significant moderate correlation ($r = .372$) between the mean error sizes of novices and experts. Overall, errors were most often one point from the certified Maryland Writing Test scores with a total of three novice errors and one expert error falling more than one point from the certified score.

3. There was no significant difference in the accuracy level of novices (60.0%) and experts (62.9%). When the scores of all fourteen raters were pooled, compositions at the extreme score points (1 and 4) were found to be more accurately scored than those in the middle score range (2 and 3). While raters made different errors within and between groups, individually, they assigned correct scores to compositions from 54% to 74% of the time. Though expert raters were slightly more accurate, there were greater within-group differences among them than within the novice group.

4. When the mean error sizes of novices and experts were compared, there was no significant difference. As with rater accuracy, there was a slight difference in rater performance in scoring papers at the extreme ends of the rating scale and at the middle range. However, the difference was not significant at the $p < .05$ level.

5. When novice raters made errors in the scores they assigned, they were below the certified Maryland Writing Test score 70.24% of the time. Expert errors were evenly divided between those that were above the certified score and those that were below.

6. Scores assigned by novices and experts as well as certified Maryland Writing Test scores were all three predicted by the lengths of the compositions and the frequency of syntax errors. Composition length was

a stronger predictor for all three sets of scores, accountable for 39% of the variance in certified scores and around 50% of the scores assigned by both sets of raters. Syntax error frequency contributed from seven to seventeen percent of the cumulative variance in the regressions. Both features were slightly stronger predictors of expert and novice rater scores than of certified scores.

7. Six of the eight writing features suggested in the literature as possible influences on rater scoring decisions did not correlate highly enough with scores to enter regression equations for either set of raters nor for certified scores. Those features included handwriting quality, mean t-unit length as a measure of sentence complexity, and errors in spelling, punctuation, capitalization, and usage.

8. Of features appearing in the scoring criteria, content was the strongest predictor for the certified scores with 99.3% of the variance in scores accountable to content analytic ratings. Only 0.1% additional variance was contributed by organization analytic scores. However, organization was the sole predictor of novice and expert rater scores, with variance ranging around the 80% level.

9. Of the five features of writing addressed in the Maryland Writing Test scoring guide, three--attention to audience, sentence formation errors affecting meaning, and conventions errors affecting meaning--did not predict certified scores. The same three features as well as content did not predict scores assigned by either groups of raters. Attention to audience, in particular, was unsuccessful as a predictor of scores in that all compositions in the study were rated as demonstrating this quality.

CHAPTER 5

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

Summary

Purpose

The study was primarily concerned with the relationship of rater training to accuracy and to writing features. The features of student writing that were included in the study were ones indicated in the literature as possible influences on rater decisions and ones that were associated with the the scoring criteria with which raters were trained. The purpose of the study was to compare the rating performance of teachers who were trained without a qualifying component in the training design with the performance of those who were required to reach 80% accuracy in training. The approach included examining the modified holistic scores assigned by teachers for accuracy and for features of student writing that were associated with score decisions.

Prior research in writing assessment has indicated difficulty on the part of teachers to agree on scoring criteria, thus reducing both the reliability and validity of many writing quality assessments (Charney, 1984; Hirsch, 1977). Large-scale writing assessment efforts such as the Maryland Writing Test have often turned to holistic scoring as a means of achieving consistency in assessments, and have been able to maintain statistical reliability from year to year (Maryland State Department of Education, 1988). However, the applicability of holistic scoring to the classroom is a different issue. This study was designed to determine if qualifying training would improve the rating performance of the teacher,

both in terms of scoring accuracy and the likely impact on instructional feedback to students.

Design of the Study

Population

The study involved two groups of Maryland English teachers who had been trained in modified holistic scoring of student compositions. The raters differed primarily in the type of training that they had experienced. The training of both groups involved the same scoring criteria identified in the Maryland Writing Test scoring guides and included both student narrative and explanatory compositions from Maryland Writing Tests prior to 1988.

Six teachers were considered novice raters because they had undergone a brief inservice training in modified holistic scoring in August of 1987, just before the beginning of the 1987-88 school year. Though they had participated in similar training experiences in previous years, novice raters had never been required in the 1987 training or in previous training activities to reach a pre-determined level of scoring accuracy. In contrast, eight other raters had been trained in January of 1988 in holistic scoring and were required to demonstrate 80% accuracy in the scoring of student papers in both narrative and explanatory domains. The papers used in this training exercise had been written by Maryland students to 1984 Maryland Writing Test prompts. Because they had demonstrated proficiency during training, the eight teachers were designated expert raters.

The six novice raters were all members of the English department of a Maryland high school. Experts consisted of eight teachers, each of

whom taught at different Maryland schools. They were the only members of the Maryland Writing Test Scoring Committee Narrative Subcommittee who were classroom teachers. Narrative subcommittee raters were selected for the study because they had not worked with 1988 explanatory papers, thus assuring that no subjects had prior knowledge of the compositions used in the study. Both groups of raters were selected for the study since they were all classroom English teachers and trained in scoring with the primary difference being in the intensity of training experienced by participants.

Instrumentation

Central to the study was a set of 35 compositions that had been written by Maryland students in response to the explanatory prompt used in the 1988 Maryland Writing Test. For each composition, the certified Maryland Writing Test score was provided by Maryland State Department of Education as a benchmark for comparison. Compositions were photocopied from original student papers with changes in identifying information such as city and school to fictitious names. Compositions were analyzed for the two sets of writing features--those that related to the Maryland Writing Test scoring criteria and those that could be possible influences on rater decisions, but that were not part of the scoring criteria.

The features related to the scoring criteria were content, organization, attention to audience, sentence formation errors that interfered with meaning, and conventions errors that interfered with meaning. Analytic scores for the five features were assigned to the 35 compositions by three writing specialists who used a scoring matrix arising from the language and structure of the Maryland Writing Test scoring

criteria. A second group of features were not directly described in the scoring criteria, but included eight characteristics of writing that had been suggested from previous research as influencing scoring decisions of raters. Those features included t-unit length, composition length, handwriting quality, and errors in spelling, punctuation, capitalization, syntax, and usage. Handwriting quality ratings were assigned by raters, but the remainder resulted from direct counts of frequencies conducted by three readers.

Data Collection Procedures

All fourteen raters in the study were asked to assign modified holistic scores to the 35 compositions using the Maryland Writing Test explanatory scoring guide. Each rater was provided with a different randomized sequence of papers for scoring. Each was also asked to assign a handwriting rating to a second randomized set of the same compositions, based on their own definitions of quality. Also included in each rater packet was a brief questionnaire used to confirm both inservice and academic training as well as work experience with modified holistic scoring.

The packets were completed by raters in May and June of 1988 in two different ways. Five of six novice raters completed their packets under supervision in single sittings at their high school while substitute teachers covered classes. The sixth novice and all eight experts completed their packets independently at home and returned materials by mail. Experts represented different schools from across the state, and it is assumed that they had no contact during the scoring of papers.

The decision-making process involved in assigning modified holistic scores entailed collecting information from each piece of writing and evaluating the writing using a standard. The standard was a scoring rubric composed of brief descriptions of what characteristics a composition should have to receive each score. The descriptions included five features of writing that were detailed for each score point. The decisions made by raters involved applying the Maryland Writing Test scoring criteria according to the fashion in which they were trained.

Normally, holistic scoring involves use of anchor papers as standards, and raters would be trained to recognize papers that resembled the quality of the sample papers. In modified holistic scoring, both a written rubric and an anchor set of papers would provide standards for readers. Training would help raters consistently apply the standards. By emulating a classroom assessment situation, it was possible to determine if the qualifying training experienced by large-scale raters would produce more desirable results than a training experience that did not result in a set level of qualifying accuracy.

Summary of the Findings

Research Question 1

Research Question 1 examined the relationship between the three sets of scores--certified scores and scores assigned by novice and expert raters. Five findings resulted from correlations of scores and errors and from examinations of mean differences in the accuracy, errors, and error directions for the two groups of raters.

1. Novice and expert raters both assigned scores that correlated at nearly the .90 level with the certified Maryland Writing Test scores that

compositions in the study actually received. There was no significant difference between the correlations of the rating groups.

2. When raters made errors on compositions, there was a significant moderate correlation ($r = .372$) between the mean error sizes of novices and experts. Overall, errors were most often one point from the certified Maryland Writing Test scores with a total of three novice errors and one expert error falling more than one point from the certified score.

3. There was no significant difference in the accuracy level of novices (60.0%) and experts (62.9%). When the scores of all fourteen raters were pooled, compositions at the extreme score points (1 and 4) were found to be more accurately scored than those in the middle score range (2 and 3). While raters made different errors within and between groups, individually, they assigned correct scores to compositions from 54% to 74% of the time. Though expert raters were slightly more accurate, there were greater within-group differences among them than within the novice group.

4. When the mean error sizes of novices and experts were compared, there was no significant difference. As with rater accuracy, there was a slight difference in rater performance in scoring papers at the extreme ends of the rating scale and at the middle range. However, the difference was not significant at the $p < .05$ level.

5. When novice raters made errors in the scores they assigned, they were below the certified Maryland Writing Test score 70.24% of the time. Expert errors were evenly divided between those that were above the certified score and those that were below.

Research Question 2

Research Question 2 examined eight writing features explored in prior research as possible predictors of quality ratings assigned to writing. Two findings resulted from stepwise regressions involving those eight features and the scores assigned by novice and expert raters as well as certified Maryland Writing Test Scores.

1. Scores assigned by novices and experts as well as certified Maryland Writing Test scores were all three predicted by the lengths of the compositions and the frequency of syntax errors. Composition length was a stronger predictor for all three sets of scores, accountable for 39% of the variance in certified scores and for around 50% of the scores assigned by both sets of raters. Syntax error frequency contributed from seven to seventeen percent of the cumulative variance in the regressions. Both features were slightly stronger predictors of expert and novice rater scores than of certified scores.

2. Six of the eight writing features suggested in the literature as possible influences on rater scoring decisions did not correlate highly enough with scores to enter regression equations for either set of raters or for certified scores. Those features included handwriting quality, mean t-unit length as a measure of sentence complexity, and errors in spelling, punctuation, capitalization, and usage.

Research Question 3

Research Question 3 examined five writing features included in the Maryland Writing Test scoring criteria. Two findings resulted from stepwise regressions involving the five features and each of the three sets of scores used in the study.

1. Of features appearing in the scoring criteria, content was the strongest predictor for the certified scores with 99.3% of the variance in scores accountable to content analytic ratings. Only 0.1% additional variance was contributed by organization analytic scores. However, organization was the sole predictor of novice and expert rater scores, with variances ranging around the 80% level.

2. Of the five features of writing addressed in the Maryland Writing Test scoring guide, three--attention to audience, sentence formation errors affecting meaning, and conventions errors affecting meaning--did not predict certified scores. The same three features as well as content did not predict scores assigned by either group of raters. Attention to audience, in particular, was unsuccessful as a predictor of scores in that all compositions in the study were rated as demonstrating this quality.

Conclusions

In large-scale educational accountability assessment programs, the primary concern is with raising rater performance to the highest levels of validity and reliability. However, in the classroom, threats to validity and reliability are numerous and virtually assure some degree of rater error. In this study, the accuracy level provided an initial statement about raters, but the subsequent analyses of error revealed some similarities and differences in the performance of the two groups of raters that helped to provide insights into the effectiveness of rater training and the likely quality of classroom assessments of student writing.

1. There was no difference in the performances of novice and expert raters. Both groups of raters made errors in score assignments around 40% of the time. Nearly all errors made by both groups of raters

were within one point of the correct score, indicating that raters had attempted to apply the criteria specified in the scoring guide. Their scores correlated significantly as did their errors, but scores correlated highly while errors correlated moderately.

Holistic scoring is an expensive assessment technique as is the qualifying training process. The expense arises both from the time required for training and the collection of training papers. The training experienced by experts in the study was intensive and required the advance collection of real papers. However, in contrast to actual large-scale procedures, these expert training procedures were generic, primarily involving papers gathered from prior tests. Such papers would be available and inexpensive to collate if similar training procedures were to be replicated for other teachers, hence reducing the materials costs, but not the time-related costs. Even then, training such as that experienced by novices would be a less expensive and less obtrusive alternative for teacher training than qualifying training.

Since the study was designed to be generalizable to writing instruction in the classroom and not to large-scale assessment, it would appear as if the added cost and time required for more intensive training such as that experienced by experts is unwarranted if the goal is accuracy. However, there were differences in error and scoring patterns that separated the performance of the two groups of raters. Novices more often underestimated the quality of student writing. At score point 2, experts assigned correct scores nearly 13% more often than novices. Further, experts were more confident in their skills.

2. Neither novices nor experts performed at what would be considered a satisfactory level of accuracy. In this study, raters scored a set of compositions that had fallen clearly under one of the four score points. They were not asked to score any compositions that were marginal, according to the judgment of assessment specialists. Raters were faced with a somewhat easier challenge than that experienced by a reader in a large-scale assessment program. In contrast, practitioners in both instruction and assessment assert that explanatory writing is both harder to produce and more perplexing to rate than narrative writing. However, it is not clear if the overall effect of the reported difficulty of explanatory writing balances the fact that papers used in the study all fairly clearly fell within each score point.

The study viewed the classroom teacher as likely to make at least some errors. Consequently, the analyses of rater performance went beyond simple calculations of the percents of times raters were correct in their score assignments. Whereas it is important that the rater in a large-scale assessment be accurate, it is generally of little interest whether the scoring errors fall into any particular patterns or not. If one assumes that the classroom teacher makes at least some errors, then it is important to know what kind of patterns occur in those errors.

As an isolated rater, the classroom teacher is without benefit of both anchor papers upon which to base standards and the averaging effect of a second reader for each composition, as occurs with large-scale assessment. Two raters trained to the 80% accuracy level generally produce defensible assessment results (Maryland State Department of Education, 1986, 1987c, 1988). If score point accuracy is the sole

measure of accuracy, one rater--regardless of the training method--scoring at the 60% accuracy level is assured of providing incorrect feedback to students on their performance 40% of the time. While there is no requirement that a teacher's assessment standards directly match state standards, the teacher is actually in danger of misguiding a student who desires to meet minimal standards for the Maryland Writing Test.

Over the course of months, a teacher might be able to compensate for some of the inaccuracy by scoring a large number of writing samples for the same student. Both increasing the number of raters reading a single composition and increasing the number of writing samples have been shown to be effective strategies for dealing with scoring problems (Breland, 1983). However, the use of different writing topics introduces new validity threats to assessment situations, requiring the teacher to generate a new interpretation of the scoring criteria with each set of papers.

Based on visual inspection of the data, it was evident that novice raters had made lower score assignments, and thus were more likely than expert raters to provide negative feedback to students. For basic writers, negative feedback, particularly when it is post-assessment feedback, has been perceived by some researchers as counter-productive in developing skills (Shaughnessy, 1977). It was also found that the compositions that were the best (score point 4) and the worst (score point 1) were most easily identified by both sets of raters, whereas papers at the middle range, were more difficult to score correctly. Unfortunately, the accurate scoring of papers at the middle score range is a critical skill since

papers at score point 2 often lead to overall failing scores, and papers at score point 3 often lead to overall passing scores in the state test.

The inability of raters to approach the 80% accuracy level points out the danger in limiting classroom assessment and feedback to modified holistic scoring. Whereas the score points on the scoring guide may be difficult for the teacher to approximate in assessment, the language is not. Annotations on student papers, based on the scoring criteria language might assist in directing student progress in the areas addressed in the scoring guide and might compensate for an inability of the teacher to weigh an overall score with precision.

3. The length of compositions and the frequency of syntax errors predicted both certified scores and rater-assigned scores. No part of the rating criteria directly favored compositions that were lengthier, but at least one part of the scoring criteria directed raters to penalize writers for sentence formation errors when such errors interfered with meaning. The similarities in the raters' performances reflected similarities in rating criteria. The regressions revealed that the scores of novices and experts both differed from certified Maryland Writing Test scores in several ways. Though length and syntax errors both were predictors of scores, the adjusted variances, however, for both features for both sets of raters were around ten percent higher than the adjusted variance for certified scores. Though composition length and frequency of syntax errors were not directly part of the scoring criteria, they were stronger predictors of rater scores than of certified scores, indicating that such features may have been somewhat more active in the raters' scoring criteria.

The importance of length as a predictor of quality ratings is not unusual. Breland and Jones (1984) found the length of compositions written for the College Board's English Composition Test (ECT) to be the strongest of several textual features in predicting scores in that test. Likewise, Stewart and Grobe (1979) and Grobe (1981) found student scores on the New Brunswick Writing Assessment Program strongly predicted by the number of words in compositions. While the Breland and Jones study involved college applicants, the New Brunswick studies used compositions written by students in grades 5,8, and 11. Both studies are comparable to the Maryland study in terms of assessment methods and age of students and contribute to the emerging pattern of research, indicating the importance of length in writing quality. Even Nold and Freedman's (1977) study of Stanford freshmen showed the importance of composition length to writing of writers who are more sophisticated than any of the other studies mentioned.

The length of the composition has the potential to bias the rater, as evidenced by the higher relationship between length and rater scores over the relationship between length and certified scores. The length of a composition is a direct measure of the extent to which a writer has fully developed a piece of writing, as was suggested by authors of several of the studies previously discussed. The link between volume of text to content and organization is evident. Consequently, the danger to the rater of being miscued as to the quality of a writing piece may be an effect that is triggered when the length is extreme in one direction or the other, not matching the rater's expectations. The study supports prior research, identifying the importance of composition length to quality, but goes further

to hint at the possibility that errors may arise from a lack of careful analysis of content.

4. Content was the strongest predictor of certified scores, but organization was the sole predictor of rater scores. The regression using features associated with scoring criteria showed a sharper difference between rater scores and certified scores. Content and organization both highly correlated with the certified scores, but the adjusted variance for content (99.3%) accounted for all but 0.1% of the cumulative adjusted variance in certified scores. However, rater scores were different enough from certified scores to result in organization emerging from the regressions with rater scores with an adjusted variance of 83.3%, indicating that raters may have placed more emphasis on organization of the compositions than on the content.

Training activities centered on all five features of the scoring guide, but some raters indicated that their scores were initially based on content and then adjusted up or down on the basis of the other four features of writing as they appeared in the writing sample. If that were true of all raters, the inability of content features to predict rater decisions may indicate specific problems in understanding part of the scoring criteria. The tendency of raters to focus on features other than content has been documented in other studies. Research on rater behavior is limited, and in no other studies had so-called correct scores been compared with rater performances. In this study, content and organization were highly correlated, and the tendency of raters' scores to be predicted by organization indicate that content may have been the intended target of readers. The net conclusion from both regressions is that the Maryland

Writing Test assesses writing in terms of content and that the teachers are approximating that emphasis, their errors accountable to other personal priorities.

5. Handwriting quality, which had been shown in previous studies to be a persistent influence on raters (Markham, 1976; Chase, 1968), was not capable of predicting scores. Neither Maryland Writing Test scores nor scores assigned by either group of raters were related to handwriting quality. In this study, raters were asked to judge handwriting quality using their own criteria, thus allowing personal bias to enter into decisions. Despite that opportunity, the scores of only one rater out of fourteen included handwriting quality as a predictor with 5% of the adjusted variance in that rater's scores accountable to handwriting quality.

Novices may have been slightly more attentive to handwriting than expert. The mean handwriting scores assigned by novices (2.18) was slightly lower than the expert mean (2.08), and the partial correlation of handwriting quality to scores was slightly higher for novices (-.101) than for experts (-.076). Though neither sets of figures were significant, this pattern coincides with the more negative view novices had of student writing. The study did not analyze the relationship of rater scoring patterns with the papers bearing the highest and lowest handwriting ratings, but such an inquiry may find that scoring errors are associated with outlying papers bearing extremely poor and extremely good handwriting, in the view of the rater.

6. Errors in spelling, punctuation, capitalization, and usage were not successful predictors of certified scores or of scores assigned by either group of raters. The scoring guide directed raters to penalize only those

writers who made excessive errors that made it difficult to understand the meaning of the writing sample. Previous research in writing assessment showed such errors to be much more important to raters, particularly with spelling errors (Baddely & Wing, 1980; Breland & Jones, 1984; Neilsen & Piche, 1981; Cooper, et al., 1984).

The inconsistency of the results of this study with prior ones may lie partly in an inability to determine from many writing assessment studies both the rating criteria as well as the training and experience of raters. The Maryland test criteria are based on the assumption that each assessed writing sample is not a final draft and that mechanical errors would be the subject of a final round of editing. Consequently, raters have approached the task with a diminished sensitivity to such errors. It is likely that many of the raters in other studies were not given comparable directions and were, in fact, allowed to consider mechanics more heavily in scoring.

Since frequency of syntax errors was second only to composition length as a predictor in that regression, it is possible that there was some relationship between modified holistic scores and sentence formation errors. The regression with features of the test criteria seemed to contradict this. However, the sensitivity of the analytic scales used in the latter regression was much less than that of the frequency counts used with syntax errors. It was possible that raters placed more emphasis on syntax errors than was merited.

A related factor is the purpose of the Maryland Writing Test, i.e., as a means of qualifying students with basic writing skills to graduate from high school. Some prior studies involved a wider scoring scale, thus

demonstrating greater sensitivity at the higher end of the scale. The Maryland Writing Test does not differentiate satisfactory writing from exemplary writing; both receive scores of 4. In college placement tests, raters may need to consider a wider range of writing characteristics to separate the most skilled students from their peers.

A study by Freedman (1979) came to similar conclusions about rater preceptions of writing quality with content being the most important characteristic of quality writing, followed by organization. That study found, as did this study, that raters paid little attention to mechanical errors. Problems with sentence structure were of little interest to raters in the Freedman study, but syntax errors were second only to length in the regression with features not related to the scoring criteria.

The Maryland Writing Test primarily measures the ability of a piece of writing to communicate meaningful information to the reader. Test responses are assumed to be second drafts, not final products. Therefore, the scoring guide directs raters to be somewhat forgiving of mechanical errors. The writing process would normally include repair of mechanical errors in the final draft (Maryland State Department of Education, 1987d). The test does not address the preparation of a final draft, but it is assumed that normal classroom programs would give students instruction and practice in needed skills. Because the public perceives that English instruction should address grammar and conventions as well as rhetorical features of writing, the Maryland Writing Test may provide a confused message to parents and students. In this study, however, teachers seemed to understand at least that aspect of the scoring criteria.

7. Sentence complexity, as measured by mean t-unit length, was not an effective predictor of certified scores or of scores assigned by either group of raters. Some studies have found sentence complexity to relate to writing quality while others have not (Hillocks, (1984). The t-unit was a difficult measure to apply to the papers used in this study in that some sentence structure problems, particularly for the worst papers resulted in what appeared to be inflated t-unit sizes. Consequently, the sentences of some writers were actually fairly complex, but were seriously flawed. As a measure of sentence complexity, i.e., the ability to subordinate thoughts, the t-unit seems to be questionable as a true complexity measure for compositions of poorer quality.

The study examined an assessment issue, i.e., how well classroom teachers could reproduce the certified Maryland Writing Test scores that had previously been assigned to compositions. However, it was more precisely an instructional issue in that the level of training experienced by teachers in the study was characteristic of what a practitioner carries into the classroom when rating student compositions. The training was of two types--one resulting in 80% rater accuracy and another with no measurable level of participant competence. Through a multi-dimensional examination of rater performance, the study was intended to help determine the level of training needed to assure rater accuracy.

The small number of participants in each group reduced the generalizability of results to other Maryland schools. However, the similarities in novice and expert rating performances pose a dilemma. One would have expected experts to perform at a higher level of accuracy than novices, given their training and backgrounds. There was no reason

to believe that the teachers who had participated as novices would have been any different than their peers across the state, but it was expected that they would have performed at a lesser level than experts. The variability in individual rater performance points out that despite the similarities in the two groups of rater results, it would be inaccurate to assume that this study represents a portrait of the average Maryland teacher. A random selection including a larger number of participants would be needed to make such generalizations.

The study involved the scoring of explanatory compositions and thus should not be generalized directly to narrative writing. While performance in rating one kind of writing might resemble the performance with another, explanatory writing includes a specific set of characteristics. Experience with rater training and instruction indicates the possibility that rater accuracy with explanatory writing would be less than with narrative writing. The predictors of rater scores and certified scores may be different for narrative writing, and the rating accuracy is assumed to be higher. However, there is no reason to believe that the lack of difference between novices and experts observed in this study might be replicated in a similar study involving narrative writing.

Because the study used Maryland teachers, it would be incorrect to expect teachers from elsewhere to perform comparably. Though experts and novices differed in the training they had experienced, both had used modified holistic scoring often in instruction, and both had experienced other inservice programs in prior years. Modified holistic scoring has become widely practiced in the state, and likely exceeds the level to which such scoring practices occur elsewhere.

Implications

Implications for Assessment

While it is always desirable for an assessment procedure to be as valid and reliable as possible, it is especially important when assessments are tied to major educational decisions such as instructional placement or graduation. The current procedures used for test development, administration, and scoring of the Maryland Writing Test include numerous expensive and time-consuming controls for validity and reliability threats. Much of the expense centers around training scoring committee members, developing training materials for raters, and training and monitoring raters.

The training experienced by expert raters was comparable to that provided to contracted raters with the following exceptions:

1. Contracted raters are always trained using sample papers from among the papers to be scored. In this study, expert raters were trained with compositions from a previous test, using the same scoring rubric.
2. Contracted raters are always trained immediately prior to the scoring of papers. Expert raters were trained several months prior to the scoring of compositions.

The scoring accuracy of expert raters in the study was significant in two ways--the low accuracy level and the lack of difference between novice and expert rater accuracy. At just over 60% accuracy, experts were not able to approach the 80% accuracy level that they had attained in training. Novices were able to achieve nearly the same level of accuracy without qualifying components in their training. However, it is not clear what most contributed to the low level of expert rater accuracy. Since the

training of experts differed from that of contracted raters in the two ways described above, it is possible that the dilution of rater accuracy was a result of both the use of generic training papers and the amount of time that lapsed between training and scoring.

Though it is unclear what causes underlie the inability of expert raters to rate compositions better than their novice counterparts, it is clear that a reduction of controls on the development, administration, and scoring of large-scale assessments is questionable. The use of a generic training procedure--as was done in this study--may open up composition scoring to undue validity and reliability threats. The time lapse between training and scoring may have contributed to the inaccuracy, but until further research separates out the effects of time lapse and training, it would be injudicious to alter current procedures.

As the rater approaches the assessment task, he or she must be focused on both the correct features of writing and on the correct criteria. A signal-detection model of writing assessment such as that described in Chapter 2 assumes that the rater carries to the scoring situation a schema composed of many criteria for judging writing quality, some of which were structured during training and some which had been assimilated through earlier experiences. This model also assumes that the composition being assessed is a contaminated signal, exposing the rater to features of writing that do not relate to the qualities being assessed as well as those that should be the sole sources of data for decision-making. Both generalized training and time lapse between training and scoring would allow for contamination of rater schema, thus contributing to rater inaccuracy.

A generalized training process does not focus the rater on the specific qualities associated with the writing topic about which students write. Though narrative criteria resemble explanatory scoring criteria, they differ enough to make generalization of one set of criteria to the other kind of writing inaccurate. Scoring criteria should include the specific characteristics the rater is to consider when assigning scores as well as the scoring scale (White, 1985, p. 24). Even two explanatory writing assignments will be judged differently because of the range of responses possible for each. A dilution of the rater's understanding of the scoring criteria may occur as the rater undergoes other experiences between training and scoring.

Implications for Instruction

Large-scale assessment procedures, by design, control accuracy threats. Rater schema is manipulated through such training procedures so that characteristics of writing which directly relate to scoring criteria become the data raters consider in making score decisions. The transference of modified holistic scoring to the classroom is a serious concern in that controls are largely absent and virtually assure that the classroom teacher is unable to focus as clearly on scoring criteria as a rater in a large-scale assessment program, thus making it difficult to establish and maintain validity and reliability of assessments in the classroom. Charney (1984) was not convinced that holistic scoring was valid and reliable for large-scale assessment let alone for the classroom. However, in Maryland, where a statewide assessment program employs modified holistic scoring, the classroom teacher needs to be able to score student writing as accurately as possible.

Some validity and reliability threats are unique to the classroom or are accentuated there. In particular, threats emerging from the scoring context are especially troublesome. While this study randomized the sequence of compositions in each packet to reduce the impact of order effect (Ferrara, 1987), most classroom teachers do not. Further, the teacher is not likely immune to other context effects such as handwriting. Though handwriting did not affect most raters in this study, repeated annoyance from individual students who write ineptly may cloud judgments of composition content. In terms of daily classroom instruction, the impact of idiosyncratic behaviors of students on writing assessments over a longer period of time is yet unexplored and may be a larger problem than indicated by this and other studies.

The study showed an overall focus of raters on the scoring criteria, and thus more on the content of compositions than on the mechanical errors made by students. The inability of the frequency of errors in punctuation, capitalization, and spelling to predict rater scores attested to the fact that teachers were approximating certified scores and thus basing their judgments around the Maryland Writing Test scoring criteria. The lack of attention of the Maryland Writing Test to mechanical errors was initially one of its most controversial points, which teachers now seem to accept in classroom instruction, and about which some critics still trouble. However, the results bear up the acceptance of Maryland teachers of the criteria and seem to separate them from subjects of prior studies who found such errors significant in score assignments (Breland & Jones, 1984; Baddely & Wing, 1980; Grobe, 1981; Cooper, et al., 1984). While the difference between those assessment tasks and the one in this study

make comparison difficult, it is at least clear that the Maryland teachers have placed mechanical errors in perspective with the scoring criteria.

The study was designed to emulate a classroom assessment situation in that no anchor papers were provided to raters. Further, the size of the set of compositions approximated what a classroom teacher might normally score. Despite the fact that novice raters had attended several inservice programs on modified holistic scoring, they felt that their training had not appreciably improved their accuracy. The expense and time involved in training were considerably higher for experts, yet superficially, the only outcome was that they were more confident and slightly more accurate. The training may have made expert errors cluster more evenly around certified scores while novices assessed compositions as worse than they really were. Perhaps the implications of that effect alone should be considered if one were designing an instructional intervention for basic writers since error analysis approaches seem to contribute to writer apprehension (Holland, 1979; Rose, 1983; Smith, 1984).

The evidence from the study does not initially support restructuring teacher inservice programs to include qualifying components for scoring accuracy. However, further examination of the literature and the data from this study raises additional issues. The study showed that composition length and frequency of syntax errors predicted both rater scores and certified scores. While composition length may relate indirectly to the degree to which a writer develops content, there remains a danger that teachers might over-invest in composition length and consider it out of context. The results of this study showed composition length to be

somewhat stronger in predicting novice scores than expert scores. This difference might also emerge in instruction.

The lack of relationship of sentence complexity to rater performance and to certified scores is worthy of further consideration. As discussed earlier, this result might have been related to the kind of writing assessed in this study. However, the use of sentence-combining strategies in instruction (Hunt, 1977) might be of limited value in preparing students for the Maryland Writing Test, yet may be critical in helping developing writers to understand how to subordinate ideas and thoughts in their writing. It seems contradictory that sentence combining strategies seem to increase composition length as well as sentence complexity (Miller & Ney, 1968; Hunt & O'Donnell, 1970), yet in this study, composition length was a predictor of scores while t-unit length was not.

The study raises questions about the level of inaccuracy in classroom assessments as it applies to improving student writing. Clearly, a single modified holistic score, without annotation, tells a student little of his or her writing quality. If those scores can be no more than 60% accurate, then it is not likely that such feedback can be of much value. However, analytic annotations, describing the extent to which the student performed in terms of the five features of writing identified in the scoring guide, might be of greater value to the student. Several participants in the study asked if annotations were needed along with holistic scores, leading one to believe that annotations are a normal part of their classroom activities. Further, the specialists who assisted with analytic scoring found the scoring matrix cumbersome in that they were forced to limit their assessments of composition quality to a simple categorized score for each

of the five features. It would seem that English teachers are more comfortable in communicating their assessments in annotations than in a score. Perhaps, though, the value of the holistic score as a complement to annotations is in balancing the importance of the five criteria in the final assessment.

Limiting assessment feedback to single holistic scores, then, cannot provide a significant contribution to a student's skill development, but coupled with annotations, the effect of the potentially inaccurate score could be buffered with documentation. While Hillocks (1984) found that working with scales proved to be an effective instructional strategy, such work included both teacher and student scoring. Such practices familiarize Maryland students with the modified holistic scoring technique, with the scoring criteria, and with some of the more critical aspects of quality writing. This supports the premise that teachers need to diversify feedback to students on writing performance and to be cautious about the accuracy of modified holistic scoring.

Implications for Future Research

The current study did not separate out the effects of time lapse and training on expert rater performance. As a part of a larger inquiry, or as an independent effort, it would be of value to determine if a generic training procedure is as effective as a specific one and if training must occur immediately prior to scoring papers. Both issues have implications for staff development and assessment programs in terms of cost and effectiveness. While assessment programs have a small margin for error, instructional feedback systems can tolerate slightly less accurate assessment procedures. In Maryland, where writing competence is a graduation

requirement, classroom writing assessment must be as accurate as possible.

The design of this study included both an examination of error patterns and of predictors of rater score assignments. It was based on an assumption that each rater's scoring criteria were composed of various elements from the rater's schema. Some of those elements were directly related to the scoring criteria on which raters were trained, and others were unrelated to the criteria, but had been indicated in the literature as possible influences on rater decisions. A signal detection model of rater decision-making would suppose that the perfect rater detects only those features of writing that relate directly to the criteria. The criteria, however, consist of elements introduced by training and elements particular to the individual rater's schema. Training would assist raters in applying only those introduced elements from the criteria to student writing, thus detecting only those "signals" from the writing related to the criteria.

This study produced evidence that both groups of raters attempted to apply the criteria. Composition length and frequency of syntax errors were slightly stronger in predicting the scores assigned by both groups of raters than they were able to predict certified scores. This indicated possible influence by scoring criteria other than those in the scoring guide. In assessing rater performance, certain individual raters assigned scores that were predicted by other criteria outside of the scoring guide.

Such data might be of use in training programs as feedback to raters, allowing them to identify possible deterrents to achieving qualifying accuracy levels. However, further research would be necessary in identifying additional features for inclusion in regressions. Further, it is

possible that alternative error taxonomies might better match the schema of some raters. Consequently, research would need to follow to refine the method used in this study. Once a reliable methodology would be developed, then the effect of performance feedback to raters could be the subject of further inquiry.

In the real world, no rater would ever be able to disengage personal rating criteria fully. The assessment task varies from narrative to explanatory writing and from topic to topic. However, the signal detection model proposes to researchers a decomposable hierarchy of the task of scoring writing that might lead to a better understanding of how the rater integrates personal experience, the incoming data, and the scoring criteria. A research effort would need to focus on the causes of both score errors and correct judgments. The evidence from this study supports the observations of practitioners that the act of assessing writing is as personal as the act of creating it. The Flower and Hayes (1980b) model of writing assumes that there is some orderly sequence of events that occurs when the writer creates, but that the process is highly individualized. The contribution of the Flower and Hayes model to writing instruction could be paralleled in assessment with a signal detection model, but the goal would be to improve both assessment techniques and training. Again, the approach would need to facilitate the individualized nature of rating the quality of writing.

Simon (1981) challenged those studying information processing to study the elements of processes as a means of understanding the whole. With writing assessment becoming more exacting than in the past, it may be time to go further than calculating rater accuracy. By analyzing rater

decisions and identifying predictors of those decisions, improvements may emerge in rater training and in the instruments and procedures of writing assessment. The goals of future research should be to examine alternative writing tasks that might better match the cognitive processes of both writers and raters, to improve assessment techniques, possibly providing a structured procedure for data collection and judgment that better matches rater cognitive processes, and to redesign training procedures. A signal detection model of assessment or some comparable cognitive model could provide a framework upon which to plan research.

Recommendations

1. English teachers should not limit writing assessment to modified holistic scoring. Teachers desiring to use holistic scoring should consider supplementing scores with written annotations, including the criteria addressed in the scoring guide. Since teachers in this study averaged around 60% accuracy in rating, the scores they assign will likely waiver adequately from the standards enough make scores alone of little value to students.

2. Staff development planners should be cautious about upgrading modified holistic scoring training procedures to include more intensive rating instruction. The more expensive and time-consuming training experienced by expert raters was not significantly more effective than the short inservice program provided to novice raters. The reason for the similarity in rater performances may be due in part to the length of time between training and scoring, or it may relate to the use of generic training materials. Future research might differentiate the effects of these two variables.

3. Staff development planners should begin exploring alternative training models that provide raters feedback on their rating performances. A primary characteristic of novice rater training was a lack of explicit discussion of scores assigned by raters. Expert rater training centered on consensus discussions where raters' disagreements with rating decisions were processed, thus facilitating the individual's assimilation of the rating criteria. The statistical analyses used in this study could be adapted to provide explicit feedback for raters on the characteristics and trends underlying rater decisions.

4. Teachers should not use modified holistic writing scores as a basis for assigning letter grades to student work. This study produced no evidence that holistic ratings alone were accurate. Further, teachers were least accurate at the middle score range (Score points 2 and 3). Since scores of 2 often lead to failure of the state test, and scores of 3 often lead to passage of the test, accuracy at that portion of the scale is especially critical. A teacher attempting to convert holistic scores to letter grades would likely convey incorrect information to students about their writing.

5. Researchers in large-scale assessment programs may wish to explore an information-processing model of holistic scoring such as the signal-detection model extracted from the ideas of Simon (1981) and Weaver (1949). The results of this study and other prior research suggest that alternative training and scoring procedures might better match the cognitive activity patterns of the rater.

Appendix A

Materials Used in Data Collection

1. Instructions
2. The Explanatory Prompt
3. Maryland Writing Test Explanatory Scoring Criteria
4. Instructions: Handwriting Quality Ratings
5. Questionnaire
6. Matrix for Analyzing Explanatory Writing

Instructions

Enclosed you will find two packets labeled "Set A" and "Set B." Before beginning work, please place your mother's maiden name and your birth month in the spaces provided on each packet envelope. Please complete all activities in Set A before proceeding to Set B. In Set A you will find the following:

* **an Explanatory Prompt**
* **the Maryland Writing Test Explanatory Rubric**
* **a set of student compositions**

Please use these materials to complete the following activities:

1. Study the prompt and scoring criteria (rubric).
2. When ready, score the student compositions with a 1,2,3, or 4, based on the scoring criteria.
3. Write and circle all scores directly on compositions.
4. When you have completed scoring compositions, place all materials back in the envelope marked Set B.
5. Read the instructions in the envelope and complete activities as requested.
6. When you have completed all activities, please place all materials back in the envelope and return both envelopes to the researcher.

The Explanatory Prompt

Suppose your principal wants one new activity for your school, It may be a new club, a new class, or a new recreational activity. Write a business letter to your principal explaining your ideas about one new activity for your school.

Before you begin writing, think about one new school activity. Think about why you would like to have the activity and who will take part in it. Consider a plan for this new activity. Think about how this new activity will benefit the school.

Now, write a business letter to the principal explaining your ideas about one new activity for your school.

Maryland Writing Test Explanatory Scoring Criteria

1 = These responses provide sufficient evidence that the writer saw the prompt and attempted to respond to it. The responses lack development and have insufficient information to complete the writing task. The little information included may be confusing or not relevant to the topic, resulting in a lack of clarity.
- o These responses lack sufficient details to explain the topic, and the details present are often vague.
- o An organizational plan, if established, is not maintained.
- o The writer may not address the intended audience
- o Errors in sentence formation interfere with meaning and confuse the reader
- o Errors in spelling, usage, punctuation, and capitalization interfere with meaning and confuse the reader.

2 = These responses contain little development and have a minimal amount of information to complete the writing task. The information included does not clearly explain the topic, and irrelevant information interferes with clarity.
- o These responses may have details, but the details may be too general or may not adequately explain the topic.
- o An organizational plan is established and minimally maintained.
- o The writer addresses the intended audience.
- o Errors in sentence formation, if present, may interfere with meaning.
- o Errors in spelling, usage, punctuation, and capitalization, if present, may interfere with meaning.

3 = These responses are adequately developed and have enough information to complete the writing task. The information is presented clearly, and irrelevant information does not interfere with clarity.
- o These responses contain some specific details that adequately explain the topic, although some details may not contribute to the development of the explanation.
- o An organizational plan is established and generally maintained.
- o The writer addresses the intended audience.
- o Errors in sentence formation, if present, do not interfere with meaning.
- o Errors in spelling, usage, punctuation, and capitalization, if present, do not interfere with meaning.

4 = These responses are well developed and have enough or more than enough information to complete the writing task. The information is presented clearly through specific details.
- o These responses contain specific details that more than adequately explain the topic.
- o The organizational plan is established and consistently maintained, with minor, if any, lapses.
- o The writer addresses the intended audience.
- o Errors in sentence formation, if present, do not interfere with meaning.
- o Errors in spelling, usage, punctuation, and capitalization, if present, do not interfere with meaning.

Some "4" papers may go well beyond these characteristics.

INSTRUCTIONS
HANDWRITING QUALITY RATINGS

Enclosed you will find the same student compositions you have just scored. This time, the compositions have been placed in a different random order, and you are to assign handwriting quality ratings to each as follows:

1= compositions which have, in your judgment, the worst handwriting of the entire set of compositions

2= compositions which have, in your judgment, an average quality of handwriting in comparison with other papers in the set

3= compositions which have, in your judgment, the best handwriting of the entire set of compositions

1. Please place your ratings ***(1,2, or 3)*** directly on each composition, and circle ratings so they can be easily located.

2. When you have completed the handwriting ratings, please complete the questionnaire attached to the back of the packet of compositions.

3. When you have completed both handwriting ratings and the questionnaire, please place all materials back in the envelope marked "Set B" and return both envelopes.

Questionnaire

Your mother's maiden name_____ Your birth month_____

INSTRUCTIONS: Please complete each of the following questions by checking the choices that most correctly describe your background and experience. After each question, a blank is provided for you to provide an explanation as requested.

1. How many school system-sponsored inservice programs on the modified holistic scoring of writing or in the scoring of the Maryland Writing Test have you attended?

_____1 _____2 _____3 _____4 _____5 or more

To what extent did those inservice programs improve your confidence in your ability to perform modified holistic scoring?

_____very much _____moderately _____little

2. How many MSDE-sponsored inservice programs on the modified holistic scoring of writing or in the scoring of the Maryland Writing Test have you attended?

_____1 _____2 _____3 _____4 _____5 or more

3. Have you ever served on the Maryland Writing Test Scoring Committee?

_____yes _____no. If yes, please give the years_____

4. Have you undergone any other inservice or academic training in holistic scoring of compositions? _____yes _____no

If yes, please give the years._____

5. Do you use holistic scoring in evaluating student compositions as a regular part of your current job? _____yes _____no. If yes, please describe how you use holistic scoring and how often you use it:

_____

6. Have you had any experience in teaching students writing skills in preparation for the Maryland Writing Test? Please include appropriate assistance instruction for students preparing to take the writing test again after a previous failure. _____yes _____no. If yes, please describe

_____

7. What academic training do you have?

Degrees Institutions majors minors

_____

_____

_____

Matrix for Analyzing Explanatory Writing

STUDENT ACTIVITY SHEET #18

Explanatory Domain Grid

Criterion	1	2	3	4
Content Details	Vague or insufficient	Minimal or too general	Specific and sufficient	Specific and sufficient or more than sufficient
Organizational Plan	If established, not maintained	Established and minimally maintained	Established and generally maintained	Established and consistently maintained
Audience	May not address intended audience	Addresses intended audience	Addresses intended audience	Addresses intended audience
Sentence Formation	Errors interfere with meaning and confuse the reader	Errors may interfere with meaning	Errors do not interfere with meaning	Errors do not interfere with meaning
Conventions	Errors interfere with meaning and confuse the reader	Errors may interfere with meaning	Errors do not interfere with meaning	Errors do not interfere with meaning

Appendix B

Compositions Used in the Study

1. Compositions
2. Composition Profiles

Score 4 #1

037015

Linda Bridges
January 6, 1974

Mr. Whaley, Principal
492 Mountain Drive
Clavens, Md. 29423

Dear Mr. Whaley:

I have come to a conclusion about your proposal. I feel you are definitely correct to wand a new school activity. Although you may feel this activity should be educational, such as a club or a new class, I feel a new recreational activity is needed. A girls basketball team has never been aviable to Clavens High Students or to any other students in the County. Why hasn't this matter ever appeared? I do believe such an activity would be gratefully appreciated by many students. As you already know, the boys team starts their season in late fall. I feel girls need some~~thing~~ kind of activity during this period. Basketball would be an excellent choice. Girls are not allowed to play on the boys team and there is not another team to play on until the spring sports come. The time between November and February is very long. Girls need to exercise to be in good condition. I feel girls are qualified to, they have as much coordination as boys do. Basketball is a very fun sport that should not only be played in Physical Education Class.

All girls who are physical capable can play the sport. All girls may participate, age is not a rescrition.-Although problems may occur in the girls grades. You know as well as I do that because of sports activities students grades drop. To stop this unwanted problem we should have some restrictions. Girls who get D's and F's would not be qualified to participate in playing on a team. The coaches of each team will also carry a certain amount of players. Those who are not as good as others will be "cut from the team."

I have made a fairly simple plan that will need some adjustments. But I do believe it proves that the new sports team is wanted and can be fit into our system of activities. I think two teams will be perfect to start the season. As the program continues, many other girls may want to play, so another team could be added. The teams will consist of 9th and 10th graders on the Junior Varsity and 11th and 12th on the Varsity team. Then in later years a freshmen team could be added. There may be some lower class girls who are equally talented as the upper classmen. If they are playing as well as the Varsity players, the coach could move them up. Practices could be placed at normal practice times, right after school. 3 to 5:30 practices could be changed to 5:30 to 7 practices. The girls team could make a deal with the boys team, such as alternating times. Girls could practice 3 out of 5 days at the early time and the boys could use the other 2. I do think something could be worked out. There may be one

problem, money! We do need the money for basketballs, uniforms and other equipment. I know having another team will be hard to fit into our budget, but students will be very grateful. The girls team won't mind practice uniforms like the boys teams. They could wear normal shorts and shirts. If possible they would be purchased at a later date. By having a new Basketball team many other High Schools in Harris County could participate. As the years go on, the teams will progress throughout the state. If so, we could turn our leagues into an all Conference division. Where each school will play others who are about equal in the amount of students. They could be divided into C,B,A,AA.

The benefits from having a Girls basketball team are incredible. By being the first school in the state to start a girls team will bring alot of Publicity. From the very minute it begins, our school could grow a wonderful name to itself. The girls would be getting a new team to play on and the county will be involved in more. More of our school activities. The grades of students will hopefully begin to go up instead of decreasing.

Thank you for your time and please reply to my idea. I think I have come up with a very promising plan for the new girls basketball team.

Sincerely,
Linda Bridges

Score 4 #2

090704

Wendy Newton
2849 Apple
Baton Rouge, Louisiana

64932

January 7, 1988

Mr. William Baylor, Principal
West River High School
6481 West River Road
Baton Rouge, Louisiana 64932

DDear Sir:

I have been informed that you are searching for an idea for a new activity. After much thought, I have come up with what seems a most suitable proposal: a spring fair.

A spring fair would be appropriate for many reasons. First, our ~~school~~ school has yet to plan their annual spring activity, and this fair seems perfect. It would ~~be~~ also be very easy to plan and carry out with the help of the students here at West River High School. Besides being a fun activity for the students, it would also involve the community. The people in our community have always wanted a part in the school system, but lately, we haven't been giving them a chance. They could be volunteers to help Set up the fair, or just come and support it by participating. Finally, I have consulted several groups, clubs and teams (including the class counsuls) and most of them think this idea would be very promising if chosen.

A spring fair would be a very fun activity to have without a lot of elaborate planning. We could hold the fair on a Saturday ~~in~~ afternoon in April. We could use the schools' parking lot for bakesales, games, booths, and contests. For example, there could be Kissing booths, raffles, refreshment stands, garage sale items and bobbing for apples. Also, school memorabelia, such as sweat shirts, key chains and buttons, could be sold. In the fields, a few small roller coasters and a ferris wheel could be set up. A three-legged race, a water balloon toss, a batting range and many other games such as these could be played. This fair could be a big success without a lot of planning.

A spring fair would benefit out school greatly. It would unite the school in setting it up and it would give them a lot of school spirit. It would be an easy fundraiser and would bring in large amounts of money. Also, much needed community involvement would be achieved. In addition, this would make a fun and exciting spring activity for the students and faculty. Finally, a spring fair would give our school a lot of positive publication in our county.

I believe that a spring fair is the best idea for a new activity at West River High School. I have given it much thought and hope you realize its potential. Thank you for your time.

Score 3 #3 083731

438 Butler Ct
Chicago, Ill. 12912
January 6, 1988

Suburban High School
2465 Madison Hwy.
Chicago, Ill. 19825

Dear Sir,

I think a football team would be a great recreational activity for the school. It will benefit the school finacially. If the team gives the school half the money they earn you can put the money twoards books or whatever else. We can earn money in different ways. We can sell pizza at carnivals. We can get the kids to go around their neighborhoods selling candy bars and collecting money. Parents of the players can also make donations. I think the biggest money raiser is a food stand at the game if we have this football team. The other half of the money that doesnt go to the school goes to the team for equitment and other purposes. If your thinking who will take part in it, i've already asked kids all around the school about it and they think it's great. The team can play at the middle school stadium for a few years till the team has enough money to maybe build a field of its own. I see a lot of talent out of the students here which could mean scolarship possibilities. If we start off good we could add another trophy case to the school for football. Since this sport is a boys sport we could start a cheerleading squad for the ~~gr~~ girls. I've already talked to some girls and they are very interested in doing that. So I think with the fund raiser and a lot of support this could be a great activity for the school.

Sincerely,
David Wayne

Score 3 #4 079329

47287 Manson Road
Benton, MD 26843
January 7, 1988

Harvey Bailey, Principal
Jones High
64721 Benton Pike
Benton, MD 26843

Dear Sir,

My rather and I have been involved in the sport of radio controlled electric race cars for over two years. This sport always opens new doors to new ideas which leads to success. An ideal way of getting everyone that wants to be involved in R/C cars is to have a club meet about once a week to talk about everyones progress and problems.

The only concern I would is the cost, but saving little by little will get you into this exciting, rapidly growing sport I have a feeling that there are many people in this school that have radio controlled cars, but they just play around with them in the street and forget about then in a week or two. This meeting would encourage these people and many, many others to join and have fun racing in competition. There are many local race tracks in Benton County for racing offroad dirt 1/10 scale and carpet 1/12 scale.

This sport is very productive and rewarding. The only thing that prevents a person from having fun and learning how to build is what they don't want to do to keep the race car in top shape Benefits of this sport of racing is that you meet many new people, have a solid grip on a hobby, encounter and conquer problems, to keep all of your equipment neat together, and have fun and safe racing. This would also help people to learn about mechanical functions of the car, and how the ~~electricl~~ electrical system works, and the most important of all, to keep up their grades without this hobby affecting then

Score 3 #5 036428

645 Green Blvd
Corden, M.D.
1/7/88

Principal of Center High school
6417 Wellsey R.D.
Corden, M.D.26431

Dear Mr. Williams

I am writing to explain to you my new Idea fora school activity. ~~Ithin~~

I think the school ~~should~~ have an end of the year ~~activity~~ trip for those students who malde honor role for 2 quders or those students who make student of the month. I think if you make good grade and are a good student you should be rewarded for your work it would almost be like a paycheck for the work you did during the year. Our parens get paid for doing wok so why don't we.

I think the studens will benefit from this because it will give them a goal too work for. I will also give them a real sense of ~~achevment~~ pride. I asked of just comming to school and passing with lower grades, the students will try for higher grades knowing that they will be rewarded the end of the year. I think this try will benefit ~~the~~ the school because it will make a better working enviomet for the school serious about their work. I think the trip should be to an amusmet park like kings Dominon Hershey Park. The trip should be an all day, overnigh trip. The trip could also be to be to a more exaxie place like a foreign country or ~~s or~~ anoker state

Finally, as you can see the trip would ~~b~~ benefit the whole school both the teacher and the studens

Score 3 #6 027623

Dear Sir,

My idea for an activity is very clear that we should have a recreation center in the area. I think this will help our school and our area in which we live. There should be different age limits such as: Mondays from 3:30 to 5:00 will be 11 and under, Friday's at five o'clock ~~will~~ until 9:00 will be 18 and over nite. These arrangements will keep our youths out of trouble and off the corner trying to sell drugs. I think our shool and the area will appreciate the recreation respect what it's there for.

I will personally help those who are interested and those who are willing to behave themselves while having fun at the same time. I will take part in keeping things in order inside the recreation center. We will need all the parent support we can get becaus it won't be easy trying to make this an success. The first things we are going to need are: pool table, ping, pong table, card table, checkers table and some very good video games. We can have a party to raise the money because parties, and disco's make money in this area. I have the idea that we can do it if we sincerly want this project to work.

Thank You For Reading the Report of

Score 2 #7 075475

6451 Polanty Street
Fike's Mills, Md. 81212
January 7, 1988

Mrs. Louise Daily principal of Fike's School
12135 Fike's Mills Road
Fike's Mills, Md 81212

Dear Mrs Daily

!Hi! My name is Sharon Pelter, I attend Fike's Mills School. I'm writing you about a new school activity. It will a dancing activity it just for fun. This activity will be plan in very special way like for 5-8 grade on Wednesday after school can start practice. I will love to take part of it just for fun. When we can have cowboys and cowgirls or break dancing lots of other thing.

To do a special thing in the spring or summer just plan for 5-8 grade students. The dancing will cost 50c to see on Wednesday starting March 28, 1988.

This activity will bring lots of children out of doing drugs or killing eachother. I will like to say this dance or talent show will be the memorable thing at school. To say to all the children of our school it will be a great deal of pleasure to see some of your out practice for the show.

The benefit of the show will raise money for the school or need or homeless people. Think about this activity because I will like to see some of our committees helping in this show. !Please! thank you for all of the consideration you are giving

Yours truly,
Sharon Pelter

Score 2 #8 057146

6435 Malenka Dr.
Swansdon, M.D 28121
January 7, 1988

Cornerstone High School
Principal
2225 Maple Dr.
Swansdon, MD 28121
1.7.88

Dear Principal:

I am writing to you because I have made one exceptional idea about a new activity that would make recreational activities better for the school. I would really like the idea if they put a pool table in the Gym for our recreational activity because it will be lots of fun for everyone. Teacher's could take part in the activity as well, as the student's. But furthermore if it is possible that you go along to what I'm explaining, and do it, I know you'll be thinking on how you will do it. "Simple, all you have to do is make easy little plans. Like for example You could get the Gym teachers to give half and half of money in order to get the pool table, and the requirement's that come with it. But not just yet! First you have to find an old room or recreational room that know one uses in the gym or near it. I personally hope if you decide on doing it that it'll succeed, and make recreational activity better for the school.

Sincerely yours
Andy Watkins
Andy Watkins

Score 2 #9 093522

William Steward Foster
Po Box 689 Harbor St.
Waterston, MD 26593
Lewis Roper High School
P.O.Box 4125
Wardsy MD 26219 ~~Dear principal~~

Dear Mr. Mason

I think that we Should have a New activity in School. We Should get 5 or 10 minte off all oul class and make a 8th pare For Drive So when you get 16 year old it will make it more easy for you. Because some people like me wont this lisue but are afraid to go and try to get them. All we got to do is make it like a wer day class do work and homework ever day but not on weaken But 15 on up should be the only one could take the class
The one who play around put tham out of the class and can come back intil ness year So that why I am writing to tall you

But for the other kids you could let them pick thir class. See you could make up activity and let them pick what they wont to take. But not over 27 people in a class. And so evey tran we swich activity than every body could have a chanse.

Score 1 #10 048486

6455 Wayne Dr.
Pontiac, Mich.
64321

Mr. Williams
Suburban High School
6465 Philip St.
Harten, Mich.
43615

Dear Mr. Williams

I would like to have a cooking club at our school we would cook foods like cookies, cakes, and Home-made ice-cream. So we can go on a Cruise to Italy and England. We, would go to Italy to go skiing in the Swiss Mountian and go to England to meat all the rock-stars ~~Ou~~ Our cooking club would a good benefit for our school.

Plus how about a club for the disabled we could help them with them with there hom work in our school. Help them get around the school so they can get to class on time. This would help the school alot

How about a club for the Smart people of ~~s~~ Suburban High School They can help out with the stuppied Kids of our school. By helpping them to do there, home-work

Lets ~~had~~ have a club ~~cold~~ called a Key-Club for all the singers in this school of our So they can go to other schools like Taylor middle

Your truely
Thomas Case

Thomas Case

051391

Score 3 #11

January 6, 1988
64551
Waynesville, MD

Always High School
Mr. Sanders, Principal
41011 Typeset Road
Mt. Burningham, MD 27404

I am writing to you, to express my ideas about a new activity that should be introduced to Always High School. This activity would be a computer learning program. The program would be educational yet at the same time be fun. This new program would take place every Monday, Tuesday and Thursday. It would last an hour - 2:45 until 3:45 - so that the participants would have time to catch the activity that will depart from the school at approximately 4:00 P.M. The program would teach the students the fundamentals of computer operations and capabilities. I would also have different levels for those who excell and for those who learn at a slower pace. Furthermore students interested in computers would get the chance to learn more about computer and its functions. This program would benefit the school by the enrichment of the students capacity to learn and apply what he/she has learn. In turn giving the students a positive towards learning.

Sincerely,
Stanley Thompson

Score 2 #12 019546

645 Sheedy Lane
Mortimer, Maryland 28516
January 7, 1988

Dr. Martin E. Welmen, Principal
Lentup Middle School
1620 Maple Drive
Mortimer, Maryland 28516

Dear Dr. Welmen, Principal:

We of class 9-17 would like to ask you a favor. The girls of our class have gathered some ideas about getting a new recreation center for the girls of this school. We would like to include, new equipment, better activities for girls. Such as a girls swim team, socker team, baseket ball team and so forth. We are asking you this because we feel that the girls of the school don't get enough equal right. We also need more activities that ~~any~~ are sutible for them.

We are not trying to say give us everything. But what we're trying to say is we want more equal rights. Some school would say no to equal rights. But we hope you won't let us down, because we need all of the suport we can get.

As you look over this letter, keep in mind that ~~our~~ we are as equal as the boy of this school. We have worked very hard to get our freedom and rights, because we do deserve them.

Sinserally Yours,
Barbara Spiehlman

Score 3 #13 059329

Edwin Washington
6429 Suburban

Avenue

Bentley, Maryland 21
January 7, 1988

Mr. Wilson Smith
Paterno Senior High
6420 Lakeview Drive
Bentley, Maryland 21681

Dear Sir:

~~I'm~~ I'm writing to you too inform you about a new school activity. We should have some type of fast food restaurant after school for Paterno student's. I think that we should have this activity because we can't go into the other food stores after we get out of school. Another reason why we should have this activity because other food store complain about us Paterno student's when we inter their stores. I think that the students and you Mr. Smith should take part in this activity.

I think that we should first get everybody together and discuss this thing about the restaurant business. We could also find out with your help, Mr. Smith to see if our plan can go into an effect. We must decide where could we have the food store built. After we get together with, Mr. Smith and discuss everything and find the perfect place to build it, maybe we could have it built by the year 1989.

I think this new activity would benefit our school in ways like for example: all the money we make for the food stre could buy computers, have better bathrooms, and buy new uniforms for all school sports. We could also ~~al~~ help the students in our school who can't afford trip. With this of business we could also help pay the student class dues who can't afford it and hopefully give teacher's extra money on their pay checks because "they always say that they don't make enough money".

Sincerely yours,
Edwin

Score 2 #14 058829

455 Marble Court
Los Angeles CA

21465

January 7, 1988

Mr. R. Able
M. Willard Henshaw High School
943 Marble Court
Los Angeles CA 21465

Dear Sir: A new class I thought of is a metals class for the students. The metals class is for the students to educate them in the field of metal I think this class is a good idea because the students of this school who what to learn this trade can. If we can get this into effect it will have students who would want to take it. I for one would take it because I am xxxxxxxx interested in the subject. We have the equitment for this and the materials. I think I could get a teacher that will volunteer for this job after school. This will also teach and inspire students to go into the field of metals. Thank you for your time I hope you will approve the metals class. ~~Thank~~

Sincerly,
Jeff Stands
Jeff Stands

Score 3 #15 027843

RT. 6 Box 893B
Weynant, Md. 26459
Thomas Wesley

Tyson Morton High School
Farley, Md. 64544
Mr. Ramsey

January 7, 1988

Dear Mr. Ramsey,
One new activity I think that would be good is a study time I think that since were all in high school, we need to study more. The reason why this activity would be good is because, every one needs to graduate. This is a activity everyone can take part in. If we shorten the time of classes and use this activity, I think this could help teach the students to learn more. This would be a great Idea for senoir because they might be planning to go te college someday, after they graduate.
If every one studys more then every one would be able to graduate. Although you can't force anyone to study, I still think that this activity would be good for Tyson Morton High. By given us the free time to study, I think that every student will concentrate in school alot betyter. I also think that this activity will give us enough time to see out friends. I think that what this Idea will bring to students will make them smarter and alot happier, because they can be with their friends.

The first step in making this plan a success, is time. The next step is placing people into certain groups, like the 9th graders in one group in a another and so on. The reason why I said all that is because sometimes students feel uncomfortable with people they don't know. This time to study will also help teachers-as things they have to do, like going over papers checking them etc. I think that a good time to start this activity is next summer when students start off for vacation. This new plan can help me out alot in my worst subjects. I can understand, that if this activity cannot work out, but please try and give this plan a chance and if this activity doen't work out we can go back the old fashion way, just plan school.

Sincerly yours,
Thomas Wesley

~~Sincerly yours,~~
~~Thomas~~

083466

Score 1 #16

Dear Principal I wolud like to start a cooking class. I got some openions from other people. We think this will helped students keep away from drugs so we would like to meet with you one Tuesday the neith to talk it over.

Yours truly
William Stams

Score 2 #17 051096

To Mrs. Flynn Principal of Taylor Mountain High. I feel as though we should have a chess club because every one does not know how to play ~~chess~~ It Would be Exciting and challenging learning how to Play Chess . Chess is a very complex and fascinating game it would really get the students thinking and at the same time it would be fun it would really motivate a person I suggest one. In a a chess club

Tammy Manors
8th grade student

Score 2 #18 001531

Corbett Maynard
4610 N. Sheraton Rd.
January 7, 1987

Mr. Saylor
Dayton Senior High

Dear Mr. Saylor

I have asked student in our school if they wanted to do some type of an activity, every decided they wanted to have a party once every month. I think that sounds great. I hope you would let us use the gym, every must pay $200 to the door, and any type of snack. We would need a d j for all of our music, and a magician to do a show for us, I hope that isn't to much trouble

Sincerly yours,
Corbett Maynard

Score 1 #19 008364

Artesian high school
4960 Scottsdale Road
Lincoln Corners Maryland
64625

To the principal
artesian high school

Dear principal I want to have a new activity becues all of the ohter activirty or tack own so I that we shad have a new activity it shad be like overe 3 year we shad have a new activity.

The new activity shab be sike and I think that we sade have it orve year and a nother activity we sade have is polo we sade have gerallout door sotps and over sopet we sade have is football.

A new activity will be good for us we dont have that ment activity as it is so I think that we sade have allest 20 activity and ouver we sade get a new one and this is how i felly.

and one outher thing will meat over outher day and if we do I want say it will be evend for me and it be on weeking it will be bater for me and i want think for out help and I want end my letter with a sac

Sencoue
your.
Wayne Daily

Score 4 #20 012934

6625 Sunrise Road
Layton, Md 64742
December 2, 1987

Mr. Leslie Wharton
7273 North Taylor Avenue
Layton, Md. 64742

Dear Sir:

Some of my friends and I have been thinking of a new club for only the Christmas season. Since we are nearing Christmas we would like # to help those who are less fortunate. Suppose, after school about 3:00 pm, on Wednesdays, Thursdays, and Fridays, in the cafeteria, anyone interested could meet. We thought ~~mab~~ maybe making some arts and crafts and selling them we could raise money. Many of the people already interested are very good in art, knitting, sewing, and cooking. We could also collect any donations people are willing to give. Our goal is $1,500.00 and we plan to reach it. Anyone can join, and it doesn't matter what grade you are in.

Hopefully, when we raise our money we would like to give it to at least three different charities. We though the Homeless children, the Aids Foundation of Maryland, and Resthaven Rest Home. When we distribute the money we would like permission for a school bus and give the money in person. We were planning on doing that on Saturday, December 21, 1987, so it will not interfere with our school work. We could leave at approximately 12:00 noon and arrive back at the school at 4:00 pm. While out distributing the money, we would like to stay and talk to some of the people.

We would really enjoy helping people this holiday season. Our "Christmas Club" will benefit the homeless and needy and will let us feel good about ourselves. Giving is what Christmas is all about, isn't it? Since the school hasn't ever done anything like this before I think many people would like to help out. Thank you for your time and hopefully many more of us can be thankful this Christmas.

Sincerly,
Deceice Tabbs
Deceice Tabbs

Score 2 #21 013884

4345 Forestview Street
Overlord, Maryland
January 7, 1988

Mr. Thompson
Fenster High School
6000 Creekton Street

Dear Mr. Thompson

One Activity for the School would be a club for the School because it would help the School. like Say if we wanted to go on a trip they could walk around the School collecting money. Or Say if the kids wanted to have a party and they were to invite Some 9# grades like 961, 962 and others classes. They would have to come to our classes and collect the money from each class and use the money to buy decorations and food and drinks. but they Should be off help when a Student gets in trouble the teachers Sends the Student down to the club and have a talk to the Student and maybe the Student will Straighten up you'll never know antil you found out. But the club Should be concerning the School really becaus they can rase money and help people in the School that would be nice is we had Something like that in Fenster Senior High School.

Sincerly Donald Willson

Score 4 #22 013176

4350 Williams Road
Creek Ford, Maryland 2764
January 7, 1987

Mr. Bill Jones
Principal, Pigeon Mountain High
1148 Maple Road
Creek Ford, Maryland 21217

Dear Mr. Jones:
I am writing to you because I have an idea about a new activity that would broaden the school's xxxxxxxx recreational activities.

The idea that I have is that I think we should add a ski team to our pension list of new activities. I really think that a ski team would benefit our school greatly. I have spoken to many students and most of them expressed a wide interest in this sport.

One benefit of skiing is a health factor. It has been proven to be a very good form of exercise. Skiing is something like an aerobic workout. It is good for the whole body, arm, legs and especially your heart.

A second factor I think you should consider is that a lot of our school's students get together and travel to a ski lodge for the weekend. I think it would be much more beneficial if the students travel as a school. This way would be a whole lot safer for the students, it would ease parents minds, and I believe that a ski team would encourage ~~good~~ school pride and spirit.

A third reason I think Pigeon Mountain should have a ski team is that it would also promote good sportsmanship. I think it would give the students a sense of achievement. Many other schools have a ski team and out school should compete too. The other schools just post up a sign up sheet, rent a bus, and get parents to chaperone. That is no harder to do than any other sport. and finally, in my opinion skiing helps teach equality. Its not a sport just for men. It is not a sport just for women. Anyone can learn how to ski and anyone can take part in it.

I hope that you would please consider a Pigeon Mountain ski team. I want to thank you for your very valuable time.

Sincerely
Susan Amory

Score 4 #23 070761

Gayle Morrison
Box 312 Whaley Dr.
Whaley, MD 97213
December 2, 1987

Mr. David Hunter, Principal
Henry Dobson High School
4132 Turbine Rd.
Turbine, MD. 81215

Dear Sir,

I would like to offer an idea to you for a new school activity. I think Peer counselors would be a good club to have in the school.

Peer counselors could help other teenagers who have problems. The counselors could help students who have peer pressure. Students who have problems with peer pressure could have someone tohelp them handle the pressure. The counselors could also help them in areas such as smoking, drinking, family, relationships, and many other problems.

My friend Melanie has talked about how she would like to have teenagers just like her help her relate her problems with other teenagers who have or had the same problems. I would like to have someone support me while I get through a problem. I would also sign up to be a counselor if we had the club. I could help others while I also help myself understand what everyone goes through. This club would reward people by knowing they helped someone and could help others too.

If you wanted the club you could start by organizing a meeting where students could sign up to be counselors. The students could start by studing about pshycology at the meetings. The counselors could make posters and flyers which introduce the club. They could hang the posters on the walls of the freshman, sophmore, junior, and senior halls. The flyers could be put on cars in the school parking lot and on lockers. Soon students would know about the counselors, what they do, and that they are there for them when they need help. If a student wanted to ~~see o~~ talk to someone ~~counselor~~, he could go to the office and ask to see a counselor. The office could have a list of the counselors and call one down out of class. If this did not work, the president of the club could set up a time for students to see the counselors that day.

I think this club would benefit the school and the process of growing up for the students. They would have an understanding of the facts of life and about people..The club would also help the school by having the students work without having to worry about their problems.

I hope you will seriously consider my idea. The activity could be a great benefit to our community and school.

Sincerely,

063162

Score 3

#24

Eric Riordon
Middleton Senior High

School

January 7, 1988

Mr. Bob L. Wayson
Principal
Middleton Senior High School
Middleton, Maryland 64962

Dear Mr. Wayson:

I have a new sport it is called Water Polo. But if we are going to have this sport we will need a pool about 50 feet wide and 100 feet long. There are many uses of a pool. But water polo is one of the best sports. All we need are 1 ball, 2 goaly nets, and head protection. This sport is one of the most physical sports. Water polo will keep you in shape.

This is the way the is played. You have ten people on each side of the team. All that you have to do is: you swim with the ball by your head down to the goal or as nere ~~(it)~~ to it. Then you pick the ball up and try to through it in to the goal. There are many move you can make like going under water, over the water, and throughing the ball to a team mate. All of these you are swimming. So you will be in the best of shape ever. Thank you for letting me talk to you and I hope that you will consider it.

Sincerely,
Eric Riordon

Score 4 #25 04909

647 Folger Court
Delton, Maryland 81201
August 3, 1987

F.M.W. High School
6465 Paint Bill Road
Grandiose, Md. 24100

Dear Sir:

As chairman of the Senior Band I would like to suggest an activity for our senior band members. Because the band has never had a chance to show off our talents or learn about music from other cultures; I suggest a European tour. The trip would be this summer during the whole month of August and the benefits would be remarkable. The strength of the music department would grow and so would the number of students enrolling in band. This trip would also help the students with music history and give our school much puplicity as a great institution of learning.

My plan for the tour is not complicated, but requires much money. This school year we would have fund raisers to get enough money, for about 100 students plus chaperone airfare. What we could not raise the Rockfish County Board of Education would supply. We would have to call the AJK tour company and order the #5 tour of musical Europe. The price is $1059.29 for adults and $639.50 for children; spending money is not included. The tour is very simple and aranges concerts for our students in such places as Hiedelburg, Venice, Vadusy, Rome, London, Paris, etc.

First we would all meet at BWI Airport on August 1 and take Lufthansa airlines to England. After the tour there we will fly to France and tour there. Next we will take a bus to Switzerland, up through West Germany, and down on to Austria. Then we will fly over the Alps to Italy, take a ship to Spain, and a bus to Portugal. Finally on August 29 we will again fly to BWI to go home.

I do hope you will aprove our plan for this educational activity and will help us to carry it through. The trip will be very benefical and enjoyable for all.

Sincerely;
Joanne Jessup,
Joanne Jessup

Score 3 #26 071781

3271 Savage Road
Westmoreland, Maryland

24681

January 7, 1988

Area High School
4432 Maple Lane
Westmoreland, Maryland 24681

Dear Mrs. Freeberger

As a student of Area High School I feel as though it's my responsability to see that my school is the best. I know Area has many different kinds of sports and we appreciate it. I have a great idea lets talk to some of the faculty members about a T.V program. It would be fun, enjoyable and also educational at the same time. For those who want to attend or already attend Area and also interested in electrical things this would be great we could have a dance to raise some of the money also when games come up we could sell popcorn, chips, punch...ect. We will have enough to get some supplies and a T.V manager by 1989. This idea would be very nifty because in the morning we listen to annoucements ~~of a~~ on a intercome but with a T.V we could see the person, they could tell today's weather, activities for before and after school, have bafflers and quizes for students and prizes if they get them right, tell yesterday's scores for a game that had been played and who scored the most even video tape it for thoes who wasn't there to see it live, I also would like to have a special for the class with the best attendence like a movie or ice cream sandwich for free at lunch time. We could have a program for the principal to tell whats going on at Area give out awards to students who has done some serious improvement. We could stay in our rooms to see programs in the auturuim instead of everybody squeezing in and loud talking the people upstage only invite maybe a few classes. I hope you consider my idea for Area. I put alot of though and ideas too. Thank You

Score 1 #27 057758

5654 Freemman Street
Basinet, Maryland 62146
November 19, 1987

Mr. Mallory
Eastman Senior high school
29065 Fourth Avenue
Basinet, Maryland 62145

Dear Mr. Mallory,

I am writing this letter to talk to you about a new activity for Eastman high school the activity that I am talking about is having more sport in the school. The sport that I am taking about is baseball the sport is very fun to play. However, many student like to play baseball and would like to take part in the activity. The activity will bring more interest to the student in school. I hope that you will thing about the new idea that I am writing you about the new activity in Eastman high school.

Sincerely Yours
William Spock

Score 4 #28 001039

Rt. 51 Box 295
Sunbury, MD 64631
January 7, 1988

Mrs. Lewis
West Hampshire High
Sunbury, MD 64631

Dear Mrs. Lewis:

I have a rather interesting proposition for you. I believe that we should have an activity and study hall period everyday.

I believe that the seven class periods, not including lunch could be shortened by seven minutes. Then we could have forty-nine minutes for an activity and studyhall period at the end of the day, considered, eighth period. Those students that have to make up missed work could do so during this time and others that need the time for a studyhall could use it as that.

This could also be benefitial to the teachers because then with this period they would not have to stay afterschool with their students when they need extra help or when making up work. This could also be used to serve afterschool detentions. Teachers could simply take away their free period.

I also believe this would be benefitial on the students part. Students would be able to get their homework done because they would have friends there if any problems would occur and it would be considered school time. Therefore students would probably be learning more and there would then be more students passing their classes.

I hope you will take into consideration my proposition of an activity and study hall period, for our school. Thank you for your time and cooperation.

Sincerely
Sally Barnes
Sally Barnes

Score 2 #29 03936

Kim Benton
4567 Pittston Rd.
Lake Hebron M.d. 74172

Harris Rogers, Principal
6465 Bay View Lane
Ocean High School
29234 Mt. Ples. M.d.

January 7, 1987

Dear Mr. Rogers

Iam one of the students here at Ocean High School who would like to see more activities for the students. Iam in a drama club, that is trying to pull together a few ideas for a new activity. There are 12 people in my drama club who all have different ideas. We would like to here a few ideas from you. So far we've come up with the Ocean High School Hopscotch team but there are different rules. The rules are: first you have to bob for apples, and how ever many you get is the number of hops you take, but for some reason I dont think the students will like the idea. The next idea is the Ocean High School pig-out team. The rules for the pig-out are first you get together people from all the schools around the area and have a few of the teachers and parents bake as many pies as possible, then whom ever can eat the most in a minute wins, but the students didn't take to that very well. The next idea was the Ocean High School Swim team. Which got a fairly good response from students. Some of the students are trying to come up with some good ideas that not only will agree with the parents and the School Board, but the students have to enjoy the sport as well. We've found that with the clubs there really wasn't much response from the students. In sports we've found a few ideas. The all teachers baseball team, but the only problem there was the teachers didn't agree. I think a talent show monthly would be fun, but no one else liked the idea so it got voted out. Some of the students would like a new class, such as a class where you don't have to bring books or paper, or pen, or notebook but the parents disagree with that, so it also got voted out. My teacher thinks a class that every two days out of the week a student will be teacher, that got a ten percent rating from the kids. We have been get new ideas everyday from all the students in the building, but some how we have to find a new activity everyone will like. Another idea was a recreational wing to be built onto the school, and sale drinks for 25c put in a few pool tables, and a few video games. We could raise the money through fund drivers. But so enough we discovered how much money that would cost, so that got voted out. Our teacher sent out ballets to everyone yesterday and today we'll find out what the new activity will be. The teacher said that it was the Ocean High School Swim team.

Sincerly yours,

Score 3 #30 00562

6799 Crabline Ct.
Thorou, MD 71248
January 7, 1988

Mark Picadillo
9495 Crabline Ct.
Thorou, MD 71248

Dear Sir,

BayLine High School needs a ski team. This team would be the only one in Chesapeake County. People would have to try out for it like any other team. There would be practices every other day at Ski Mountain. The ones who make the team would have to have there own equiptment like ski's, poles, boots and proper ski clothing. "The Rockfish ski team could travel to ski slopes around the area and compete in matches. It would benefit the school because if we had a good team we would bring our school recognition and would be famous. Another reason to have a ski team is to show other schools that our school is modern and original & then maybe some other schools will get ski teams. For this team all we would need is transportation and uniforms. We could get money for those from a fund raiser or from the schools budget. "The Rockfish Ski Team" would be a great asessment to the school & to the kids.

Score 2 #31 001877

Dear Dr. Mary Taylor
I would like to start a basketball club for boy from the age 12-18. It would twice a week it would the boy how to play basketball The class should have 15 boy. We need to have 4 teachers for this class.

Score 2 #32 018832

Louie Chamberlain
1603 White Oak Court
Clemson, Maryland 21:
January 7, 1988

Mr. B. Bailor
3745 Twilight Road
Phillips Landing, Maryland 42436

Dear Sir:
I feel as though we should have a special art class. The reason for this is because I think that we have very talented students who can draw, at our school. Art, is fun and recreational for the students of our school. We can get the most talented students with good grades and put in one or two special art classes. This special activity may also benefit our school. We can go on special educational trips, and even participate in contests with other schools. If we win the school contest we can go on to the county wide contest. That means we have a chance to be awarded for our talented students, and that could give the school a good name.

Sincerely Yours;
Louie Chamberlain

Score 3 #33 039912

654 Grant Street
Clydesdale, Maryland

64653

January 7, 1988

Ms. Gayle Lytle, principal
Hilltop Senior High School
6000 Poplar Lane
Towswain, Maryland 24623

Dear Ms. Lytle:

Greetings! My name is Linda Wilson and I am a sophmore at Hilltop High. I am writing in reference to your idea for a new activity for our school.

I think I have come up with a nice idea. My idea is a Latin Club. That's right. A Latin Club! I think the students could benefit from the language because it would help students interested in their already required languages: English. spainish, and French
The students would be speaking one of the first languages used for business and education. Latin was also the foundation for the Romance Languages, they are: Spainish, French, Italian, Romanian, and Portugese. It would be like learning five languages in one! Another thing I think the students could benefit from Latin is when they go to college! Some colleges find it impressive that a student have Latin down as a language they learned in high school.

I think the reason I'm so interested in Latin is because it's really helped me in my study of other languages, and it's a really fun language because while speaking it you can pick upon alot of words you already know Like I've already said it's really a good thing to have on a college application.

A plan for the club, I think, should be better off left for you to decide! That doesn't mean I don't have any ideas! I think it should be offered as another foreign language elective for those who want it. Then again, I think that it should be kept a club run by students or the students pick a qualified teacher! Students should also be able to plan our own sensible field trips, not anything crazy or wild like a amusement park, but something educational and fun!

The school would also benefit from this because it will draw more students who don't already attend our school, to this school because they like the available courses. It would also give students who already attend a chance to learn a different and unique language! Thank you for your time!

Sincerly,
Linda Wilson

Score 2 #34 038685

485 Wentworth Drive
Hartford MD 92104

463 wellford Road
Hartford MD, 92104

Dear Mr. Fenster,

I think we should have activity week. It is where Monday thru thursday we sind up for activities to do. Then on friday we can pick to go to the bowling alley, Roller Skating Swimming or just stay in school and do activity. It would be alot of fun and I will try to get the whole 9th graders to do it. The cost of this would be for the bowling alley $3.35 swimming $4.75 Roller Skating $5.25. The things that we could do in this activity is we can write all kinds of games on a piece of paper. Like checkers, monlopy, Board Walk, Parcheeses, cards etc. Then we let all the students pick what they want to do. After everybody is done we see how many games we get or if you don't want to do that then we can have more dances. Everybody likes to go to our dances. They can't go unless we have them. I think we should have a dance like every two weeks. People sit home on fridays and wish they had something to do. If you have more dances then they would have something to do. I know I gave you two ideals but I wanted to tell you about both of them.

Sincerely yours,
Mandy
Creighton

Score 3 #35 069590

Rt. 12 Box 45B
Hygrop, MD 64791
January 7, 1988

Mr. Tom Fisher
Rt. 6 Box 912
Westover, MD 64618

Dear Mr. Tom Fisher

There is a problem in our school which has been brought to my attention Too many of our soccer players aren't staying in good physical health. Something needs to be done about this growing problem. Several suggestions have been made, but one such solution just might be the answer. The solution of which I am speaking about is Indoor Soccer

At first, alot of students asked "Why?" My first reason is (as I stated before). so that the regular season soccer players can stay physically active. My second reason is. Indoor Soccer is a common interest of many students. Not only to soccer fans, but other people as well. It would be an easy task setting up the gym for an activity Such as this one. I'm sure that the equipment needed could be easily found within. the schools perimeter. As you can see, this activity would not only suit the needs of the soccer players, but other students too. Every body likes to take a little time off from schoolwork.

I am hoping you will take this letter into consideration, ~~I have~~ for I have shown you an uncostly way to keep students entertained in school. Nobody can tell where this activity might go from here, but we are anxious to find out. I ask you to make a descision, and, at the same time thank you ~~four~~for your cooperation and ~~time~~ support.

Sincerely,

COMPOSITION PROFILE

Number 1 **Serial No.** 37015 **Certified MWT Score** 4

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	4	4	4	4	4	4			4.00	0	0.000
Expert Raters	4	4	4	4	4	4	4	4	4.00	0	0.000

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
11.85	711	0.70	2.39	2.11	0.844	1.406

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
4	3.67	1	3	3

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	2	3	2	3	3	3			2.67
Experts	2	3	3	1	3	3	3	2	2.50

COMPOSITION PROFILE

Number 2 **Serial No.** 90704 **Certified MWT Score** 4

Raters	1	2	3	4	5	6	7	8

Modified Holistic Scores

									Mean Score	Error Size Total	Error Size Mean
Novice Raters	4	4	4	4	4	4			4.00	0	0.000
Expert Raters	4	4	4	4	4	4	4	4	4.00	0	0.000

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
13.30	399	1.00	1.25	0.50	0.754	0.251

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
4	4	1	3	3

Handwriting Ratings

Raters	1	2	3	4	5	6	7	8	Mean
Novice Raters	2	1	1	1	2	1			1.33
Experts	2	3	2	3	3	2	2	2	2.38

COMPOSITION PROFILE

Number 3 **Serial No.** 83731 **Certified MWT Score** 3

	Raters								Mean Score	Error Size Total	Error Size Mean
	1	2	3	4	5	6	7	8			
Modified Holistic Scores											
Novice Raters	2	4	3	3	3	2			2.83	3	0.500
Expert Raters	3	2	3	3	3	4	2	2	2.75	4	0.500

Non-MWT Features

T-Unit	**Len.**	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
14.56	262	1.53	4.20	0.38	1.145	0.000

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
3	2.67	1	3	3

Handwriting Ratings

	Raters								Mean
	1	2	3	4	5	6	7	8	
Novice Raters	2	3	3	1	3	3			2.50
Experts	2	3	2	2	2	3	2	2	2.25

COMPOSITION PROFILE

Number 4 Serial No. 79329 Certified MWT Score 3

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	3	2	3	3	2	2			2.50	3	0.500
Expert Raters	4	3	3	4	3	4	4	3	3.50	4	0.500

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
20.92	272	1.10	2.21	0.00	2.574	1.103

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
3	3	1	3	3

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	2	1	2	2	2	2			1.83
Experts	2	3	2	2	2	2	2	2	2.12

COMPOSITION PROFILE

Number 5 **Serial No.** 36428 **Certified MWT Score** 3

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	3	3	2	2	4	3			2.83	3	0.500
Expert Raters	3	2	3	2	3	3	3	2	2.62	3	0.375

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
17.77	231	6.93	3.03	2.16	3.030	3.463

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
3	3	1	3	3

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	1	1	1	2	1	1			1.17
Experts	1	1	1	1	1	1	1	1	1.00

COMPOSITION PROFILE

Number 6 **Serial No.** 27623 **Certified MWT Score** 3

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Mean
Novice Raters	3	3	1	2	2	2			2.17	5	0.833
Expert Raters	3	3	3	2	3	3	2	2	2.62	3	0.375

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
18.42	221	1.81	2.72	1.81	0.452	2.262

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
3	2.67	1	3	3

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	3	3	2	3	3	3			2.83
Experts	3	3	2	2	3	2	3	3	2.62

COMPOSITION PROFILE

Number 7 **Serial No.** 75475 **Certified MWT Score** 2

Modified Holistic Scores

	Raters								Mean Score	Error Size Total	Error Size Mean
	1	2	3	4	5	6	7	8			
Novice Raters	2	2	2	2	2	2			2.00	0	0.000
Expert Raters	2	3	2	2	2	2	2	2	2.12	1	0.125

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
13.06	209	1.44	2.39	0.96	1.914	3.828

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
2	2	1	2	2.33

Handwriting Ratings

	Raters								Mean
	1	2	3	4	5	6	7	8	
Novice Raters	2	2	2	2	3	3			2.33
Experts	3	3	2	2	2	3	2	2	2.38

COMPOSITION PROFILE

Number 8 Serial No. 57146 Certified MWT Score 2

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	2	2	2	2	2	2			2.00	0	0.000
Expert Raters	2	2	2	3	2	2	2	2	2.12	1	0.125

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell	Punct.	Capt.	Usage	Syntax
22.38	179	0.00	6.70	2.23	1.117	2.793

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
2	2	1	2.33	2.33

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	1	2	1	2	3	2			1.83
Experts	2	3	2	2	1	2	2	2	2.00

COMPOSITION PROFILE

Number 9 Serial No. 93522 Certified MWT Score 2

Raters

1	2	3	4	5	6	7	8

Modified Holistic Scores

								Mean Score	Error Size Total	Error Size Mean
Novice Raters										
1	1	1	2	1	2			1.33	4	0.667
Expert Raters										
2	3	2	2	2	3	2	2	2.25	2	0.250

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
17.50	175	10.86	4.57	4.57	4.000	2.286

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
2	2	1	1.33	1.33

Handwriting Ratings

Raters

1	2	3	4	5	6	7	8	Mean
Novice Raters								
2	1	1	2	3	2			1.83
Experts								
2	3	2	2	2	2	2	2	2.12

COMPOSITION PROFILE

Number 10 **Serial No.** 48486 **Certified MWT Score** 1

Modified Holistic Scores

	Raters								Mean Score	Error Size Total	Error Size Mean
	1	2	3	4	5	6	7	8			
Novice Raters	1	1	1	1	2	1			1.17	1	0.167
Expert Raters	1	1	2	2	2	2	2	2	1.75	6	0.750

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
13.42	161	4.35	7.45	2.48	2.484	4.348

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
1	1	1	1.67	2.67

Handwriting Ratings

	Raters								Mean
	1	2	3	4	5	6	7	8	
Novice Raters	2	2	1	3	1	2			1.83
Experts	1	2	1	2	1	1	1	2	1.38

COMPOSITION PROFILE

Number 11 **Serial No.** 51391 **Certified MWT Score** 3

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	3	3	3	3	3	3			3.00	0	0.000
Expert Raters	3	2	3	3	3	4	3	3	3.00	2	0.250

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
18.22	164	0.61	3.66	0.00	1.829	1.829

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
3	3	1	3	3

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	3	3	3	1	3	1			2.33
Experts	3	3	3	3	3	2	3	2	2.75

COMPOSITION PROFILE

Number 12 **Serial No.** 19546 **Certified MWT Score** 2

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	2	1	2	2	2	3			2.00	2	0.333
Expert Raters	2	2	2	3	2	2	2	2	2.12	1	0.125

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
16.36	180	2.22	4.44	1.11	2.778	0.556

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
2	2	1	2	2

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	2	1	2	2	2	2			1.83
Experts	2	2	2	2	3	2	2	2	2.12

COMPOSITION PROFILE

Number 13 Serial No. 59329 Certified MWT Score 3

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	3	3	2	2	3	2			**2.50**	3	**0.500**
Expert Raters	3	2	3	4	4	3	3	3	3.12	3	0.375

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
20.75	249	0.80	3.21	0.00	0.803	2.008

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
2.67	3	1	2.67	2.67

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	2	1	2	2	2	2			1.83
Experts	2	1	1	1	1	1	2	1	1.25

COMPOSITION PROFILE

Number 14 Serial No. 58829 Certified MWT Score 2

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	2	3	2	2	2	2			2.17	1	0.167
Expert Raters	2	2	2	2	2	2	2	2	2.00	0	0.000

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
14.44	130	0.77	3.85	0.00	0.769	0.000

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
2	2	1	2.67	2.67

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	1	2	2	2	2	1			1.67
Experts	1	3	3	2	1	1	3	2	2.00

COMPOSITION PROFILE

Number 15 **Serial No.** 27843 **Certified MWT Score** 3

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	3	3	3	4	3	3			3.17	1	0.167
Expert Raters	3	4	2	4	4	4	3	3	3.38	5	0.625

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
16.25	325	4.62	3.39	0.62	1.846	0.307

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
3	3.33	1	2.67	2.67

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	2	2	2	2	3	3			2.33
Experts	2	3	2	2	3	2	2	2	2.25

COMPOSITION PROFILE

Number 16 **Serial No.** 83466 **Certified MWT Score** 1

Modified Holistic Scores

	Raters								Mean Score	Error Size Total	Error Size Mean
	1	2	3	4	5	6	7	8			
Novice Raters	1	1	1	2	1	1			1.17	1	0.167
Expert Raters	1	1	1	1	1	2	2	1	1.25	2	0.250

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
10.25	41	0.98	4.88	0.00	2.439	2.439

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
1	1	1	2.67	2.67

Handwriting Ratings

	Raters								Mean
	1	2	3	4	5	6	7	8	
Novice Raters	1	1	1	2	2	3			0.20
Experts	2	1	2	1	1	2	1	2	1.50

COMPOSITION PROFILE

Number 17 **Serial No.** 51096 **Certified MWT Score** 2

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	1	1	2	2	1	1			1.33	4	0.667
Expert Raters	1	2	1	1	1	2	2	1	1.38	5	0.625

Non-MWT Features

T-Unit	**Len.**	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
8.50	68	0.00	1.47	5.88	0.000	5.882

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
2	2	1	2.67	2.67

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	1	1	1	2	1	1			1.33
Experts	1	2	1	1	1	2	2	1	1.38

COMPOSITION PROFILE

Number 18 Serial No. 1531 Certified MWT Score 2

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	1	1	1	2	1	1			1.17	5	0.833
Expert Raters	2	2	1	1	2	2	2	1	1.62	3	0.375

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
9.00	81	0.00	4.94	2.47	3.704	4.938

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
2	2	1	2.33	3

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	2	3	2	2	3	3			2.50
Experts	2	3	3	2	2	3	2	2	2.38

COMPOSITION PROFILE

Number 19 Serial No. 8364 Certified MWT Score 1

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Mean
Novice Raters	1	1	1	2	1	1			1.17	1	0.167
Expert Raters	1	1	1	1	2	2	2	1	1.38	3	0.375

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
11.27	169	20.12	3.55	1.78	2.367	5.917

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
1	1	1	1	1

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	1	1	1	1	1	2			1.17
Experts	1	2	1	2	2	2	2	1	1.62

COMPOSITION PROFILE

Number 20 Serial No. 12934 Certified MWT Score 4

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	4	4	1	3	3	4			3.17	5	0.833
Expert Raters	4	4	4	4	4	4	3	4	3.88	1	0.125

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
13.05	287	0.34	1.05	0.34	0.000	1.045

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
4	4	1	3	3

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	2	2	3	3	3	3			2.67
Experts	3	3	2	3	3	2	3	2	2.62

COMPOSITION PROFILE

Number 21 Serial No. 13884 Certified MWT Score 2

Modified Holistic Scores

	Raters								Mean Score	Error Size Total	Error Size Mean
	1	2	3	4	5	6	7	8			
Novice Raters	1	2	1	2	1	1			1.33	4	0.667
Expert Raters	2	2	1	2	2	2	2	2	1.88	1	0.125

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
16.30	163	2.45	0.61	12.27	3.681	3.681

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
2	2	1	2	2.33

Handwriting Ratings

	Raters								Mean
	1	2	3	4	5	6	7	8	
Novice Raters	2	2	3	1	3	3			2.33
Experts	3	3	3	3	2	2	3	2	2.62

COMPOSITION PROFILE

Number 22 **Serial No.** 13176 **Certified MWT Score** 4

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	4	4	4	4	4	3			3.83	1	0.167
Expert Raters	4	3	3	4	4	4	4	4	3.75	2	0.250

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
11.63	314	0.64	0.96	0.00	0.000	0.319

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
4	4	1	3	3

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	2	2	3	3	3	3			2.67
Experts	2	3	3	2	3	2	3	2	2.50

COMPOSITION PROFILE

Number 23 **Serial No.** 70761 **Certified MWT Score** 4

Modified Holistic Scores

	Raters								Mean Score	Error Size Total	Error Size Mean
	1	2	3	4	5	6	7	8			
Novice Raters	4	4	4	4	4	4			4.00	0	0.000
Expert Raters	4	4	4	4	4	4	4	4	4.00	0	0.000

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
15.60	390	0.77	0.51	0.26	0.000	0.000

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
4	4	1	3	3

Handwriting Ratings

	Raters								Mean
	1	2	3	4	5	6	7	8	
Novice Raters	2	2	3	3	3	3			2.67
Experts	3	3	3	2	3	2	3	3	2.75

COMPOSITION PROFILE

Number 24 Serial No. 63162 Certified MWT Score 3

Raters

1	2	3	4	5	6	7	8

Modified Holistic Scores

								Mean Score	Error Size Total	Error Size Mean
Novice Raters										
2	2	2	2	2	2			2.00	6	1.000
Expert Raters										
2	2	2	3	2	2	2	2	2.12	7	0.875

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
12.40	186	1.08	1.08	1.08	1.613	2.689

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
3	2.67	1	3	3

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	2	1	2	2	3	2			2.00
Experts	3	3	2	2	2	2	2	2	2.25

COMPOSITION PROFILE

Number 25 Serial No. 41909 Certified MWT Score 4

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	4	4	2	3	3	4			3.33	4	0.667
Expert Raters	4	3	3	3	4	4	4	4	3.62	3	0.375

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
14.57	306	1.31	2.61	0.98	0.000	0.000

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
4	4	1	3	3

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	2	1	2	1	3	2			1.83
Experts	3	3	3	3	3	2	3	2	2.75

COMPOSITION PROFILE

Number 26 **Serial No.** 71781 **Certified MWT Score** 3

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	3	3	4	2	2	2			2.66	4	0.667
Expert Raters	3	3	2	3	3	3	4	3	3.00	2	0.250

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
16.61	299	3.01	3.68	0.00	0.669	2.676

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
3.33	3	1	3	3

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	2	2	1	1	3	3			2.00
Experts	3	3	2	2	2	1	3	2	2.25

COMPOSITION PROFILE

Number 27 Serial No. 57758 Certified MWT Score 1

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	1	2	2	2	2	1			1.66	4	0.667
Expert Raters	1	2	1	1	2	2	2	2	1.62	5	0.625

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
13.86	97	2.06	1.03	4.12	3.093	2.062

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
1	1.33	1	2.67	2.67

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	2	1	2	2	3	3			2.17
Experts	2	2	3	2	1	1	2	2	1.88

COMPOSITION PROFILE

Number 28 **Serial No.** 1039 **Certified MWT Score** 4

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	4	4	4	4	4	3			3.83	1	0.167
Expert Raters	3	3	3	3	3	4	3	3	3.13	7	0.875

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
13.81	221	2.71	3.62	0.00	0.452	0.000

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
4	4	1	3	3

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	3	2	3	3	3	3			2.83
Experts	3	3	3	3	3	3	3	3	3.00

COMPOSITION PROFILE

Number 29 **Serial No.** 36936 **Certified MWT Score** 2

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	2	2	2	3	2	1			2.00	2	0.333
Expert Raters	1	1	1	2	3	2	4	2	2.00	6	0.750

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
12.94	466	0.43	1.72	0.64	0.644	0.644

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
2	2	1	3	3

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	3	3	3	2	3	3			2.83
Experts	3	3	2	2	3	2	3	3	2.62

COMPOSITION PROFILE

Number 30 Serial No. 562 Certified MWT Score 3

Modified Holistic Scores

	Raters								Mean Score	Error Size Total	Error Size Mean
	1	2	3	4	5	6	7	8			
Novice Raters	3	3	3	3	3	3			3.00	0	0.000
Expert Raters	2	3	2	2	3	3	3	3	2.62	3	0.375

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
14.17	170	2.94	2.94	1.18	0.588	0.000

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
3	3	1	3	3

Handwriting Ratings

	Raters								Mean
	1	2	3	4	5	6	7	8	
Novice Raters	2	2	2	1	2	2			1.83
Experts	2	3	2	2	2	2	3	2	2.25

COMPOSITION PROFILE

Number 31 Serial No. 1877 Certified MWT Score 2

Modified Holistic Scores

	Raters								Mean Score	Error Size Total	Error Size Mean
	1	2	3	4	5	6	7	8			
Novice Raters	1	1	1	2	1	1			1.17	5	0.833
Expert Raters	1	2	1	1	1	2	2	2	1.50	4	0.500

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
8.60	43	0.00	4.65	0.00	9.302	6.977

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
2	2	1	2.67	2.67

Handwriting Ratings

	Raters								Mean
	1	2	3	4	5	6	7	8	
Novice Raters	2	1	1	1	2	2			1.50
Experts	1	2	2	2	2	2	2	2	1.88

COMPOSITION PROFILE

Number 32 **Serial No.** 18832 **Certified MWT Score** 2

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	3	3	2	2	2	2			2.33	2	0.333
Expert Raters	2	3	2	2	2	3	2	2	2.25	2	0.250

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
13.44	121	0.00	3.31	0.00	0.826	0.000

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
2	2	1	2.67	2.67

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	3	2	3	2	3	3			2.67
Experts	3	2	2	2	3	2	3	2	2.38

COMPOSITION PROFILE

Number 33 **Serial No.** 39912 **Certified MWT Score** 3

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	3	4	2	4	4	4			3.50	5	0.833
Expert Raters	4	4	3	4	4	4	4	4	3.88	7	0.875

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
13.15	342	0.88	2.05	0.29	0.826	0.000

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
3	3	1	3	3

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	2	1	2	3	2	2			2.00
Experts	2	3	2	1	3	2	2	2	2.12

COMPOSITION PROFILE

Number 34 Serial No. 8685 Certified MWT Score 2

Raters

1	2	3	4	5	6	7	8

Modified Holistic Scores

								Mean Score	Error Size Total	Error Size Mean
Novice Raters										
2	3	2	2	2	1			2.00	2	0.333
Expert Raters										
3	2	2	3	3	3	2	2	2.50	4	0.500

Non-MWT Features

T-Unit Len.		Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
12.35	210	1.43	3.80	2.86	1.905	1.905

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
2	2	1	2.67	2.67

Handwriting Ratings

				Raters				
1	2	3	4	5	6	7	8	Mean
Novice Raters								
2	1	1	3	3				3.00
Experts								
2.2	2	2	3	3	1	2	2	2.12

COMPOSITION PROFILE

Number 35 **Serial No.** 69590 **Certified MWT Score** 3

Modified Holistic Scores

	Raters 1	2	3	4	5	6	7	8	Mean Score	Error Size Total	Error Size Mean
Novice Raters	4	4	2	2	2	2			2.67	6	1.000
Expert Raters	3	3	3	3	4	3	4	4	3.38	3	0.375

Non-MWT Features

T-Unit	Len.	Errors Per 100 Words: Spell.	Punct.	Capt.	Usage	Syntax
12.83	231	1.30	3.46	2.16	0.866	0.433

MWT Features

Content	Organization	Audience	Sen. Form. Errors	Conven. Errors
3	3	1	3	3

Handwriting Ratings

	Raters 1	2	3	4	5	6	7	8	Mean
Novice Raters	2	1	2	2	3	1			1.83
Experts	2	3	3	3	2	2	2	2	2.38

Appendix C

Data Tables for Individual Rater Errors

Table 16

Individual Novice Rater Errors: Number of Points Above and Below Certified Maryland Writing Test Scores

	Number of Score Points From Certified MWT Scores									
	-3		-2		-1		0		+1	
Rater	n	%	n	%	n	%	n	%	n	%
Score Point 1										
1							4	100		
2							3	75.00	1	25.00
3							3	75.00	1	25.00
4							1	25.00	3	75.00
5							2	50.00	2	50.00
6							4	100.00		
Total							17	70.83	7	29.17
Score Point 2										
1					5	41.67	6	50.00	1	8.33
2					3	25.00	4	33.33	5	41.67
3					4	33.33	8	66.67		
4							11	91.6	1	8.33
5					5	41.67	7	58.33		
6					6	50.00	5	41.67	1	8.33
Total					23	31.94	41	56.94	8	11.11

	Number of Score Points From Certified MWT Scores									
	-3		-2		-1		0		+1	
Rater	n	%	n	%	n	%	n	%	n	%
Score Point 3										
1					2	16.67	9	75.00	1	8.33
2					2	16.67	7	58.33	3	25.00
3			1	8.33	5	41.67	5	41.67	1	8.33
4					6	50.00	4	33.33	2	16.67
5					5	41.67	5	41.67	2	6.67
6					7	58.33	4	33.33	1	8.33
Total			1	1.39	27	37.50	34	47.22	10	13.89
Score Point 4										
1							7	100		
2							7	100		
3	1	14.29	1	14.29			5	71.43		
4					2	28.57	5	71.43		
5					2	28.57	5	71.43		
6					2	28.57	5	71.14		
Total	1	2.38	1	2.38	6	14.29	34	80.95		

Number of Score Points From Certified MWT Scores

	-3		-2		-1		0		+1	
Rater	n	%	n	%	n	%	n	%	n	%
Score Points 1-4										
1					7	20.0	26	74.28	2	5.72
2					5	14.29	21	60.0	9	25.71
3	1	2.86	2	5.72	9	25.71	21	60.0	2	5.71
4					8	25.71	21	60.0	5	17.14
5					12	34.29	19	54.28	4	11.43
6					15	42.86	18	51.43	2	5.71
Total	1	.48	2	.95	56	26.67	126	60.00	25	11.90

Table 17

Individual Expert Rater Errors: Number of Points Above and Below Certified Maryland Writing Test Scores

	Number of Score Points From Certified MWT Scores							
	-1		0		+1		+2	
Rater	n	%	n	%	n	%	n	%
Score Point 1								
1			4	100				
2			3	75.00	1	25.00		
3			3	75.00	1	25.00		
4			3	75.00	1	25.00		
5			1	25.00	3	75.00		
6					4	100.00		
7					4	100.00		
8			2	50.00	2	50.00		
Total			16	50.00	16	50.00		

	Number of Score Points From Certified MWT Scores							
	-1		0		+1		+2	
Rater	n	%	n	%	n	%	n	%
Score Point 2								
1	3	25.00	8	66.67	1	8.33		
2	1	8.33	8	66.67	3	25.00		
3	5	41.67	7	58.33				
4	3	25.00	6	50.00	3	25.00		
5	2	16.67	8	66.67	2	16.67		
6			9	75.00	3	25.00		
7			11	91.67			1	8.33
8	2	16.67	10	83.33				
Total	16	16.67	67	69.79	12	12.50	1	1.04
Score Point 3								
1	2	16.67	8	66.67	2	16.67		
2	5	41.67	5	41.67	2	16.67		
3	4	33.33	8	66.67				
4	3	25.00	5	41.67	4	33.33		
5	1	8.33	7	58.33	4	33.33		
6	1	8.33	6	50.00	5	41.67		
7	3	25.00	5	41.67	4	33.33		
8	4	33.00	6	50.00	2	16.67		
Total	23	23.96	50	52.08	23	23.96		

	Number of Score Points From Certified MWT Scores							
	-1		0		+1		+2	
Rater	n	%	n	%	n	%	n	%
Score Point 4								
1	1	14.29	6	85.71				
2	3	42.86	4	57.14				
3	3	42.86	4	57.14				
4	2	28.57	5	71.43				
5	1	14.29	6	85.71				
6			7	100.00				
7	2	28.57	5	71.43				
8	1	14.29	6	85.71				
Total	13	23.21	43	76.79				
Score Points 1 - 4								
1	6	17.14	26	74.29	3	8.57		
2	9	25.71	20	57.15	6	17.14		
3	12	34.29	22	62.85	1	2.86		
4	8	22.86	19	54.28	8	22.86		
5	4	11.43	22	62.85	9	25.71		
6	1	2.86	22	62.85	12	34.29		
7	5	14.29	21	60.00	8	22.86	1	2.86
8	7	20.00	24	68.57	4	11.43		
Total	52	18.57	176	62.86	51	18.21	1	0.36

Appendix D

Distributions of Scores

Figure 1

Distribution of Certified Maryland Writing Test Scores by Score Point

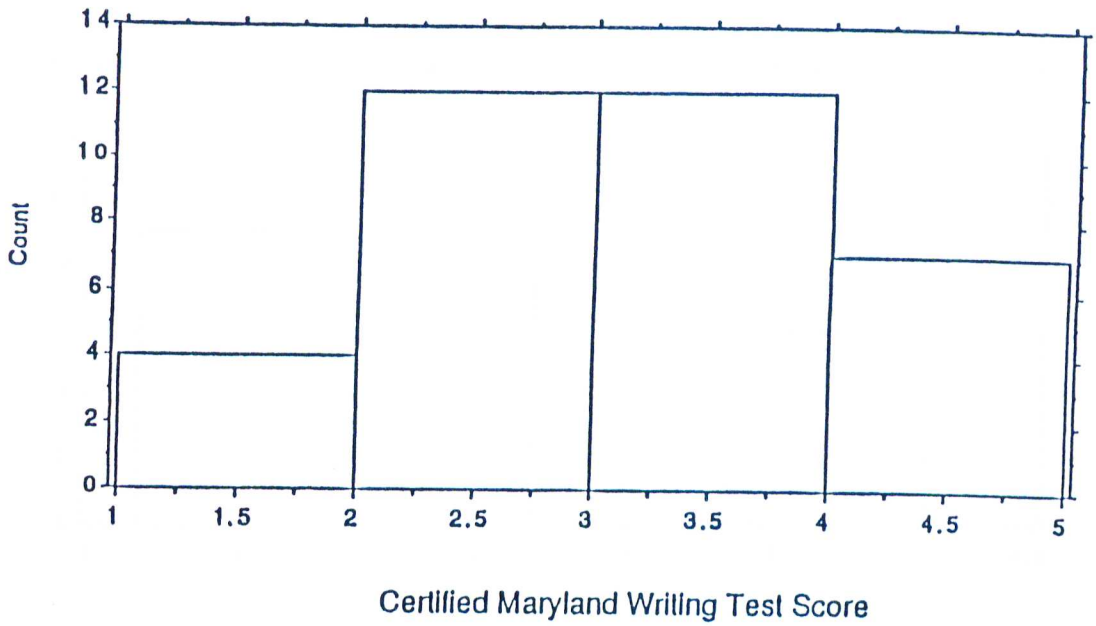


Figure 2

Distribution of Novice Rater Scores by Score Point

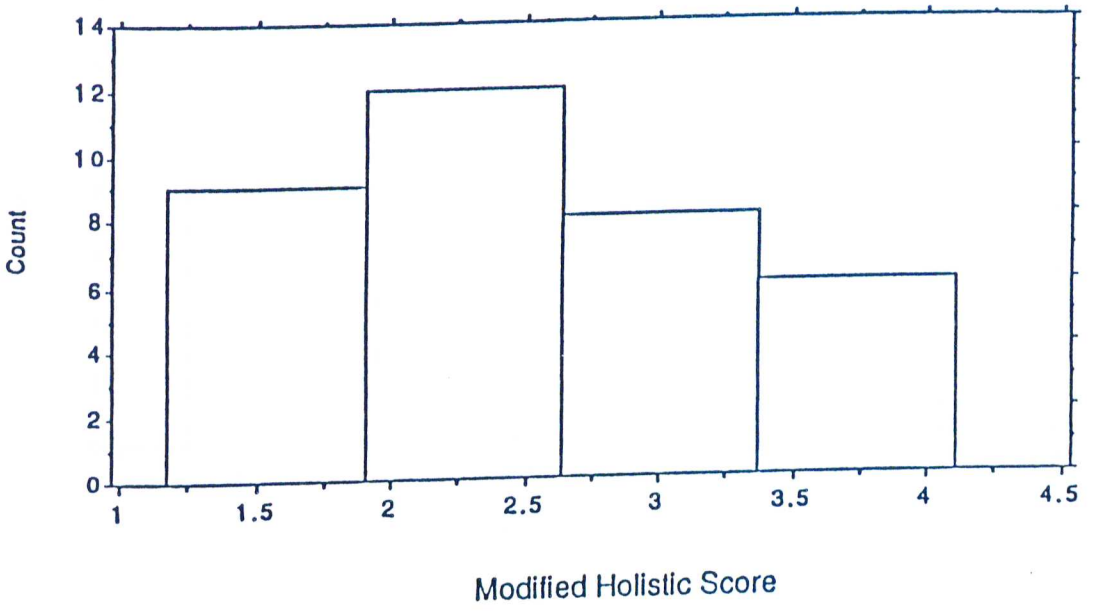
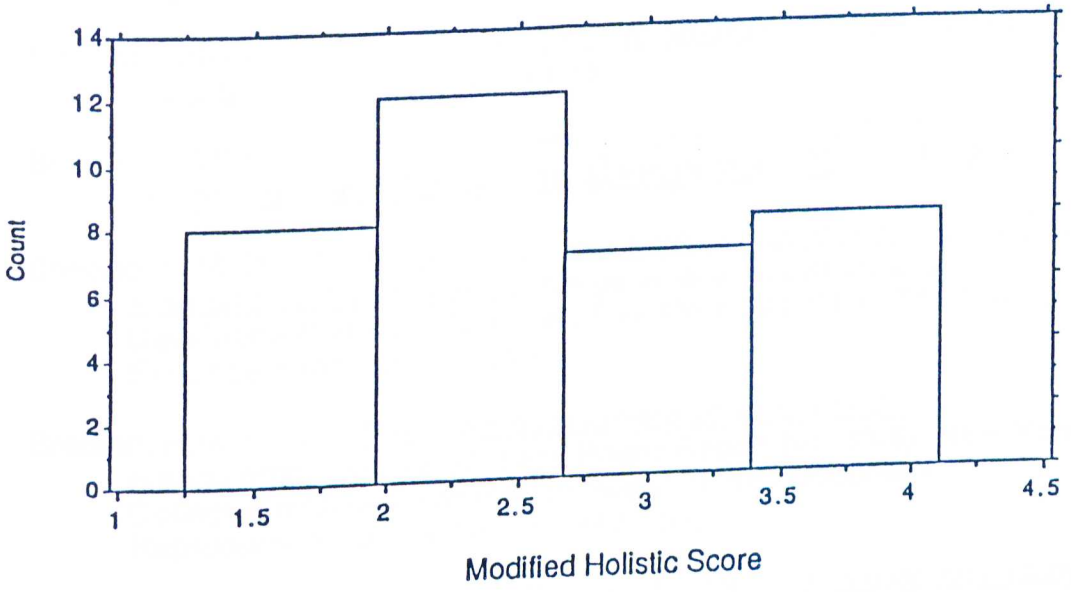


Figure 3

Distribution of Expert Rater Scores by Score Point



REFERENCES

Applebee, A.N. (1984). Writing and reasoning. Review of Educational Research, 54, 577-596.

Baddely, A.D. and Wing, A.M. (1980). Spelling errors in handwriting: A corpus and a distributional analysis. In U. Frith (Ed.), Cognitive processes in spelling (pp. 251-286). New York: Academic Press.

Bartholomae, D. (1980). The study of error. College Composition and Communication, 31, 253-269.

Bell Laboratories (1982). UNIX Writers Workbench Software Reference Manual. Piscataway, N J : Author.

Blok, H. (1985). Estimating the reliability, validity, and invalidity of essay ratings. Journal of Educational Measurement, 22(1), 41-52.

Breland. H.M. (1977). A study of college English placement and the Test of Standard Written English (College Board Research and Development Report RDR-76-77-4). Princeton, N J : The College Entrance Examination Board.

Breland, H.M. (1983). The direct assessment of writing skill: A measurement review (College Board Report No. 83-6). New York: College Entrance Examination Board. (ERIC Document Reproduction Service No. ED 242 756)

Breland, H.M., Conlan, G.C., and Ragosa, D. (1977). A preliminary study of the Test of Standard Written English. Princeton, N J : Educational Testing Service.

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

Breland, H.W. and Jones, R.J. (1984). Perceptions of writing skills. Written Communication, 1, 101-119.

Brossell, G. (1983). Rhetorical specification in essay examination topics. College English, 45, 165-173.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. Research in the Teaching of English, 20. 65-81.

Chase, C.I. (1983). Essay test scores and reading difficulty. Journal of Educational Measurement, 20, 293-297.

Clemson, E. (1978). A study of the basic skills assessment direct and indirect measures of writing ability. (ERIC Document Reproduction Service No. 204 409)

Coffman, W.E. (1971). On the reliability of rating essay examinations in English. Research in the Teaching of English, 5, 24-36.

Combs, W. (1976). Further effects of sentence-combining practice on writing ability. Research in the Teaching of English, 10, 137-149.

Cooper, C.R. (1977). Holistic evaluation of writing. In C.R. Cooper and L. Odell (Eds.), Evaluating writing: Describing, measuring, judging (pp. 3-32). Urbana, Ill.: National Council of Teachers of English.

Cooper, C.R., Cherry, R., Copley, B., Fleischer, S., Pollard, R., and Startisky, M. (1984). Studying the writing abilities of a university freshman class: Strategies from a case study. In R.Beach and L.S. Bridwell (Eds.), New directions in composition research (pp. 19-52). New York: Guilford Press.

Crowhurst, M. (1980). Syntactic complexity and teachers' quality ratings of narrations and arguments. Research in the Teaching of English, 14, 223-231.

Culpepper, M.M. and Ramsdell, R. (1982). A comparison of a multiple choice and an essay test of writing skills. Research in the Teaching of English, 16, 295-297.

Daiute, C.A. (1981). Psycholinguistic foundations of the writing process. Research in the Teaching of English, 15, 5-22.

Daiute, C.A., Allen, R.L., Jandreau, S.M., Chametzky, R.A., and Bever, T.G. (1981). Psycholinguistic studies of the writing process in adolescents (Report to the National Institute of Education, Grant No. NIE G-80-0041).

Daly, J.A. and Dickson-Markman, F. (1982). Contrast effects in evaluating essays. Journal of Educational Measurement, 23, 33-41.

Faigley, L.L. (1979). The influence of generative rhetoric on the syntactic maturity and writing effectiveness of college freshmen. Research in the Teaching of English, 13, 197-206.

Ferrara, S. (1987, April). Effects of essay order on rater's score assignments in a large-scale writing assessment. Paper presented at the annual meeting of the American Educational Research Association, Washington, D C.

Flower, L.S. and Hayes, J.R. (1980a). The cognition of discovery: Defining a rhetorical problem. College Composition and Communication, 30, 161-164.

Flower, L.S. and Hayes, J.R. (1980b). The dynamics of composing: Making plans and juggling constraints. In L.W. Gregg and E.R. Steinberg (Eds.), Cognitive processes in writing (pp. 31-50). Hillsdale, N J : Lawrence Erlbaum.

Freedman, S.W. (1979a). How characteristics of student essays influence teacher evaluations. Journal of Educational Psychology, 71(3), 328-338.

Freedman, S.W. (1979b). Why do teachers give the grades they do? College Composition and Communication, 30, 161-164.

Freedman, S.W. (1984). The registers of student and professional expository writing: Influences on teachers' responses. In R.Beach and L. Bridwell (Eds.), New directions in composition research (334-347). New York: Guilford Press.

Freedman, S.W. and Calfee, R.C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, and S.A. Walmsley (Eds.), Research on writing: Principles and methods (75-98). New York: Longman.

Godshalk, F.I., Swinford, F., and Coffman, W.E. (1966). The measurement of writing ability. Princeton, N.J.: Educational Testing Service

Gorrell, D. (1983). Toward determining a minimal competency entrance examination for freshman composition. Research in the Teaching of English, 17, 263-274.

Gould, S.J. (1981). The mismeasure of man. New York: W.W. Norton and Company.

Greenberg, K.L. (1981). Competency testing: What role should teachers of composition play? College Composition and Communication, 33, 366-376.

Grobe, C.H. (1981). Syntactic maturity, mechanics, and vocabulary as predictors of quality ratings. Research in the Teaching of English, 15, 75-86.

Grobe, S.F. and Grobe, C.H. (1977). Reading skills as a correlate of writing ability in college freshmen. Reading World, 16, 50-54.

Hales, L.W. and Tokar, E. (1975). The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay question. Journal of Educational Measurement, 12, 115-117.

Hayes, J.R. and Flower, L.S. (1980). Identifying the organization of writing processes. In L.W. Gregg and E.R. Steinberg (Eds.), Cognitive processes in writing (pp. 3-30). Hillsdale, N J : Lawrence Erlbaum.

Heller, M. (1979). The reading-writing connection: An analysis of the written language of university freshmen at two reading levels. Dissertation Abstracts International, 40, 4452A.

Hiebert, E.H., Englert, C.S., and Brennan, S. (1983). Awareness of text structure in recognition and production of expository discourse. Journal of Reading Behavior, 15, 63-79.

Hillocks, Jr., G. (1984). What works in teaching composition: A meta-analysis of experimental treatment studies. In N.L. Stein (Ed.), Literacy in American schools: Learning to read and write (pp. 137-174). Chicago: University of Chicago.

Hillocks, Jr., G. (1986). Research on written composition: New directions for teaching. Urbana, Ill., ERI Clearinghouse on Reading and Communication Skills and National Conference on Research in English.

Hirsch, Jr. E.D. (1977). The philosophy of composition. Chicago: University of Chicago Press.

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

Hughes, D.C. Keeling, B. and Tuck, B.F. (1983) Effects of achievement expectations and handwriting quality on scoring essays. Journal of Educational Measurement, 20, 65-70.

Hunt, K.W. (1965). Grammatical structures written at three grade levels: Research report no. 3. Urbana, IL: National Council of Teachers of English.

Hunt, K.W. and O'Donnell, R. (1970). An elementary school curriculum to develop better writing skills. Tallahassee: Florida State University.

Huntley, R.M., Schmeiser, C., and Stiggins, R. (1979). The assessment of rhetorical proficiency: The role of objective tests and writing samples. Washington, D.C.: Government Printing Office. (ERIC Document Reproduction Service No. ED 173 419)

Kiefer, K. and Smith, C.R. (1984). Improving students' revising and editing: The Writer's Workbench system. In W. Wresch (Ed.), The computer in composition instruction (pp. 65-82). Urbana, IL : National Council of Teachers of English.

Lloyd-Jones, R. (1977). Primary trait scoring. In C. Cooper and L. Odell (Eds.), Evaluating writing: Describing, measuring, judging (pp. 33-68). Urbana, IL: National Council of Teachers of English.

Lindsay, P.H. (1970). Multichannel processing in perception. In D.I. Mostofsky (Ed.), Attention: Contemporary theory and analysis (pp. 149-183). New York: Merideth.

Markham, L.R. (1976). Influences of handwriting quality on teacher evaluation on written work. American Educational Research Journal, 13, 277-283.

Maryland State Department of Education. (1986). Technical report: Maryland functional writing test-II spring 1986 administration (Report W-86), Baltimore, MD: Author.

Maryland State Department of Education. (1987a). Pursuing perfect prompt. School, 34(3), 4.

Maryland State Department of Education. (1987b). Revisions in the Maryland Functional Writing Program, Baltimore, MD Author.

Maryland State Department of Education. (1987c). Technical report: Maryland functional writing test-II spring 1987 administration (Report W-87), Baltimore, MD : Author.

Maryland State Department of Education. (1987d). 1987 Maryland Writing Supplement: Project Basic Instructional Guide, Baltimore, MD : Author.

Maryland State Department of Education. (1988). Technical report: Maryland functional writing test-II winter 1988 administration (Report W-88), Baltimore, MD : Author.

McColly, W. (1970). What does educational research say about the judging of writing ability? Journal of Educational Research, 64, 147-156.

McCready, M.A., and Melton, V.S. (1981). Feasibility of assessing writing using multiple assessment techniques: Research report. Ruston: Louisiana Technical University. (ERIC Document Reproduction Service No. ED 182 465)

McNicol, D. (1972). A primer of signal detection theory London: George Allen and Unwin Ltd.

Michael, W.B. and Shaffer, P. (1978). The comparative validity of the California State University and Colleges English Placement Test (CSUC-EPT) in the prediction of fall semester grade-point average and average English course grades of first semester entering freshmen. Educational and Psychological Measurement, 38, 985-1001.

Miller, B. and Ney, J. (1968). The effect of systematic oral exercises on the writing of fourth-grade students. Research in the Teaching of English, 2, 44-61.

Mellon, J.C. (1969). Transformational sentence-combining: A method for enhancing the development of syntactic fluency in English composition (NCTE Research Report No. 10. Urbana, IL : National Council of Teachers of English.

Moss, P.A., Cole, N.S., and Khampalikit, C.A. (1982). A comparison of procedures to assess written language skills in grades 10, 7, and 4. Journal of Educational Measurement, 19, 37-47.

Neilsen, L. and Piche, G.L. (1981). The influence of headed nominal complexity and lexical choice on teachers' evaluation of writing. Research in the Teaching of English, 15, 65-74.

Nold, E.W. and Freedman, S.W. (1977). An analysis of readers' responses to essays. Research in the Teaching of English, 11, 164-174.

Nystrand, M. (1982). An analysis of errors in written communication. In M. Nystrand (Ed.), What writers know: The language, process, and structure of written discourse (pp. 57-74). New York: Academic Press.

Odell, L. (1981). Defining and assessing competence in writing. In C. R. Cooper (Ed.), The Nature and Measurement of Competency in English (pp. 65-94). Urbana, IL : National Council of Teachers of English.

O'Hare, F. (1973). Sentence-combining: Research report no. 15. Champaign, IL : National Council of Teachers of English.

Pedersen, E.L. (1977). Improving syntactic and semantic fluency in writing of language arts students through extended practice in sentence-combining. Unpublished doctoral dissertation, University of Minnesota.

Potter, R.R. (1967). Sentence structure and prose quality. Research in the Teaching of English, 1, 17-28.

Reesink, G.P. Hollemann-van der Sleen, S.B., Stevens, K., and Kohnstumm, G.A. (1971). Development of syntax among school children and adults: A replication-investigation. (From Psychological Abstracts, 1971, 47, Abstract No. 10536.)

Ritchie, D. (1986). Shannon and Weaver: Unravelling the paradox of information. Communication Research, 13, 278-298.

Rose, M. (1983). Writer's block: The cognitive dimension, Carbondale, IL : Southern Illinois University Press.

San Jose, C. (1972). Grammatical structures in four modes of writing at fourth grade level (Doctoral Dissertation, Syracuse University, 1972). Dissertation Abstracts International, 33. (University Microfilms No. 73-9763.)

Schmeling, H.H. (1970). A study of the relationship between certain syntactic features and overall quality of college freshman writing (Doctoral dissertation, George Peabody College for Teachers, 1969). Dissertation Abstracts International, 30/11A, 4970. (University Microfilm No. 70-9763)

Shanahan, T. (1984). Nature of the reading-writing relation: An explanatory multivariate analysis. Journal of Educational Psychology, 466-477.

Shanahan, T. and Lomax, R.G. (1986). An analysis and comparison of theoretical models of the reading-writing relationship. Journal of Educational Psychology, 78, 116-123.

Shannon, C. (1949). The mathematical theory of communication. In C. Shannon and W. Weaver (Eds.) The mathematical theory of communication. Urbana, Ill., University of Illinois Press.

Shaughnessy, M.P. (1977). Errors and expectations: A guide for the teacher of basic writing. New York: Oxford University Press.

Simon, H. (1981). The sciences of the artificial. Cambridge, MA : MIT Press.

Smith, L., Winters, L., Quellmalz, E., and Baker, E. (1980). Characteristics of student writing competence: An investigation of alternative scoring systems. Los Angeles: UCLA Center for the Study of Evaluation.

Smith, M. (1984). Reducing writing apprehension. Urbana, IL :National Council of Teachers of English.

Stewart, M.F. and Grobe, C.H. (1979). Syntactic maturity, mechanics of writing and teachers' quality ratings. Research in the Teaching of English, 13, 207-215.

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

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

Thomas, D. and Donlan, D. (March, 1980). Correlations between holistic and quantitative methods of evaluating student writing, grades 4-12. Paper presented at the combined Annual Meeting of the Conference on English Education and the Secondary School English Conference, Omaha, Ne. (ERIC Document Reproduction Service No. 211 976)

Thorndike, E.L. (1912). Education: A first book. New York: Macmillan.

Weaver, W. (1949). Recent contributions to the mathematical theory of communication. In C. Shannon and W. Weaver (Eds.), The Mathematical Theory of Communication. Urbana,IL : University of Illinois Press.

White, E.M. (1985). Teaching and assessing writing. San Francisco: Jossey-Bass.

Williams, J.M. (1981). The phenomenology of error. College Composition and Communication, 31, 152-168.