



Indicators of quality in natural language composition
by Barry John Donahue

A thesis submitted in partial fulfillment of the requirements for the degree of DOCTOR OF
EDUCATION

Montana State University

© Copyright by Barry John Donahue (1982)

Abstract:

This study was designed to: (1) examine the relationships that exist between various commonly used measures of writing quality; and (2) determine to what extent experienced English teachers and prospective English teachers agree in their opinions of writing quality. The measures of writing quality chosen for comparison were Holistic scoring, Atomistic scoring, Mature Word Index, Type/Token Index, Mean T-unit Length, and Syntactic Complexity. The Holistic and Atomistic methods are subjective and thus required several human raters, while the other four methods are objective and could be scored using mechanical procedures. Four groups of raters were used in the study, corresponding to all possible combinations of subjective methods (Holistic and Atomistic) with experience levels (experienced teachers and prospective teachers). Both the Holistic and Atomistic methods provided very high reliability coefficients for all groups of raters, but there was a large range of reliabilities for the categories of the Atomistic method. The conclusions of the study were: (1) The Atomistic scoring method is more time-consuming and no more reliable or informative than Holistic scoring. (2) Many of the factors generated by Diederich do not provide reliable results between raters.

(3) The Mature Word Index and Type/Token Index are accurate measures of writing quality, while the Mean T-unit Length and Syntactic Complexity Index are not.

(4) Writers do not misuse or misplace mature words as they often do syntactic structures.

(5) Student raters judge writing as a whole in essentially the same manner as do expert raters, but are slightly less able to distinguish the various factors of quality writing.

The recommendations made in the study included preference of Holistic methods over Atomistic methods, distrust of the Mean T-unit Length and Syntactic Complexity methods, and the need to convey to prospective teachers their competence as judges of writing quality.

INDICATORS OF QUALITY IN NATURAL
LANGUAGE COMPOSITION

by

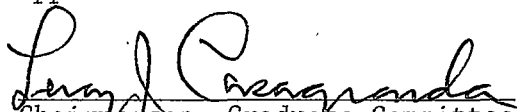
BARRY JOHN DONAHUE

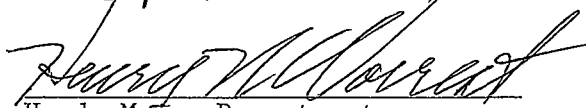
A thesis submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF EDUCATION

Approved:


Chairperson, Graduate Committee


Head, Major Department


Graduate Dean

MONTANA STATE UNIVERSITY
Bozeman, Montana

August, 1982

ACKNOWLEDGEMENT

The author would like to thank several people for their assistance in the preparation of this study. First, thanks to Lynne Jermunson, Roxanna Davant, and Merry Fahrman for helping with the initial editing and grading of the essays. Dr. Sara Jane Steen's assistance in allowing the author to use the members of her English methods classes as raters is greatly appreciated. Thanks to those raters as well as the expert raters.

The staff of the Word Processing Center at Montana State University, especially Judy Fisher and Debbie LaRue must be thanked for their very competent and professional typing and revising of the manuscript of this study.

The author would also like to thank Dr. Eric Strohmeier, Dr. Robert Thibeault, Dr. Gerald Sullivan, Dr. Douglas Herbster, and Professor Duane Hoynes for their participation on the author's Doctoral Committee. Their assistance, especially during the rush of the final weeks, has been very helpful. Finally, many thanks to Dr. Leroy Casagrande, Chairman of the Committee for his patience and guidance throughout the author's graduate work, and for his advice in the preparation of this study.

TABLE OF CONTENTS

VITA	ii
ACKNOWLEDGEMENT.	iii
LIST OF TABLES	vii
ABSTRACT	x
Chapter	
I. INTRODUCTION.	1
Statement of the Problem	3
Applications of the Results.	4
Questions Answered by the Study.	8
General Procedures	8
Limitations and Delimitations.	10
Definition of Terms.	12
Summary.	14
II. REVIEW OF LITERATURE.	16
Grading Essay Writing.	16
Holistic Methods of Evaluation	21
Atomistic Methods of Evaluation.	24
Mature Word Choice	32
Fluency.	34
Vocabulary Diversity	41
Standardized Tests	43
Summary.	45

III. METHODS	47
Essay and Rater Descriptions	47
Categories of Investigation.	51
Method of Data Collection.	52
Statistical Hypotheses	55
Analysis and Presentation of Data.	57
Calculations	61
Summary.	62
IV. RESULTS	64
Comparability of Student Rater Groups.	64
Intraclass Reliabilities	66
Comparison of Students and Experts Using Atomistic Scoring.	68
Correlations between Methods	69
Correlations of Atomistic Categories with Methods.	76
Correlations between Categories of the Atomistic Method	82
Correlations of Methods with Sum of Rankings of All Other Methods	85
Overall Correlations	87
Analysis of Variance between Expert and Student Raters	89
Summary.	92
V. DISCUSSION.	94
Summary of the Study	94

Conclusions.	96
Holistic Versus Atomistic Scoring.	97
Correlations between Methods	100
Correlations between Atomistic Categories and Methods.	103
Correlations of Methods with Sum of Rankings of All Other Methods.	107
Overall Correlations	108
Comparison of Expert and Student Raters.	108
Recommendations.	110
Suggestions for Future Research.	112
REFERENCES CITED	114
APPENDIXES	119
A. COMPUTER PROGRAMS.	119
B. RAW SCORES OF RATERS	128
C. INTERMEDIATE RESULTS FROM CALCULATION OF MATURE WORD INDEX.	147

LIST OF TABLES

1. Interpretation of the Standard Frequency Index	33
2. Comparison of Grade Point Averages for Student Groups Using Holistic and Atomistic Scoring	65
3. Reliability of Average Ratings of Holistic and Atomistic Methods and Each Category of Atomistic Method.	66
4. Average Scores for Methods Utilizing Raters.	69
5. Raw Scores for Methods Not Utilizing Raters.	70
6. Rank Ordering of Methods and Rater Groups.	71
7. Pearson Correlation Matrix of Methods and Rater Groups . . .	72
8. Spearman Rank Order Correlation Matrix of Methods and Rater Groups	74
9. Pearson Correlations between Methods and Categories of Atomistic Scoring from Students	76
10. Spearman Rank Order Correlation between Methods and Categories of Atomistic Scoring from Students.	77
11. Pearson Correlations between Methods and Categories of Atomistic Scoring from Experts	78
12. Spearman Rank Order Correlations between Methods and Categories of Atomistic Scoring from Experts	79
13. Pearson Correlations between Categories of Atomistic Scoring for Experts.	82
14. Pearson Correlations between Categories of Atomistic Scoring for Students	83
15. Pearson Correlations between Categories of Atomistic Scoring for Experts and Those for Students	84
16. Correlations between Each Method and the Sum of Rankings of All Other Methods	85

17.	Kendall Coefficients of Concordance for All Methods.	87
18.	Kendall Coefficients of Concordance for Holistic, Atomistic, Mature Word Index, and Type/Token Index Methods.	88
19.	Analysis of Variance for Holistic Rating Groups by Essays	89
20.	Analysis of Variance for Atomistic Rating Groups by Essays.	91
21.	Reliability of Ratings by the Student Group Using Holistic Scoring	129
22.	Reliability of Ratings by the Expert Group Using Holistic Scoring	130
23.	Reliability of Ratings for the Category "Ideas" by the Student Group Using Atomistic Scoring.	131
24.	Reliability of Ratings for the Category "Organization" by the Student Group Using Atomistic Scoring	132
25.	Reliability of Ratings for the Category "Wording" by the Student Group Using Atomistic Scoring.	133
26.	Reliability of Ratings for the Category "Flavor" by the Student Group Using Atomistic Scoring.	134
27.	Reliability of Ratings for the Category "Usage" by The Student Group Using Atomistic Scoring.	135
28.	Reliability of Ratings for the Category "Punctuation" by the Student Group Using Atomistic Scoring	136
29.	Reliability of Ratings for the Category "Spelling" by the Student Group Using Atomistic Scoring.	137
30.	Reliability of the Total of All Categories by the Student Group Using Atomistic Scoring.	138
31.	Reliability of Ratings for the Category "Ideas" by the Expert Group Using Atomistic Scoring	139

32.	Reliability of Ratings for the Category "Wording" by the Expert Group Using Atomistic Scoring.	140
33.	Reliability of Ratings for the Category "Organization" by the Expert Group Using Atomistic Scoring.	141
34.	Reliability of Ratings for the Category "Flavor" by the Expert Group Using Atomistic Scoring	142
35.	Reliability of Ratings for the Category "Usage" by the Expert Group Using Atomistic Scoring	143
36.	Reliability of Ratings for the Category "Punctuation" by the Expert Group Using Atomistic Scoring.	144
37.	Reliability of Ratings for the Category "Punctuation" by the Expert Group Using Atomistic Scoring.	145
38.	Reliability of the Total of All Categories by the Expert Group Using Atomistic Scoring	146
39.	Mature Words Used in the Essays.	148
40.	Contractions, Proper Nouns, and Slang Used in the Essays . .	149
41.	Topic Imposed Words Used in the Essays	149
42.	Number of Types and Tokens Used in the Essays.	150

ABSTRACT

This study was designed to: (1) examine the relationships that exist between various commonly used measures of writing quality; and (2) determine to what extent experienced English teachers and prospective English teachers agree in their opinions of writing quality. The measures of writing quality chosen for comparison were Holistic scoring, Atomistic scoring, Mature Word Index, Type/Token Index, Mean T-unit Length, and Syntactic Complexity. The Holistic and Atomistic methods are subjective and thus required several human raters, while the other four methods are objective and could be scored using mechanical procedures. Four groups of raters were used in the study, corresponding to all possible combinations of subjective methods (Holistic and Atomistic) with experience levels (experienced teachers and prospective teachers). Both the Holistic and Atomistic methods provided very high reliability coefficients for all groups of raters, but there was a large range of reliabilities for the categories of the Atomistic method. The conclusions of the study were:

- (1) The Atomistic scoring method is more time-consuming and no more reliable or informative than Holistic scoring.
- (2) Many of the factors generated by Diederich do not provide reliable results between raters.
- (3) The Mature Word Index and Type/Token Index are accurate measures of writing quality, while the Mean T-unit Length and Syntactic Complexity Index are not.
- (4) Writers do not misuse or misplace mature words as they often do syntactic structures.
- (5) Student raters judge writing as a whole in essentially the same manner as do expert raters, but are slightly less able to distinguish the various factors of quality writing.

The recommendations made in the study included preference of Holistic methods over Atomistic methods, distrust of the Mean T-unit Length and Syntactic Complexity methods, and the need to convey to prospective teachers their competence as judges of writing quality.

CHAPTER I

INTRODUCTION

The skill of effective written communication is one of the most valuable assets which the educated person possesses. It forms the foundation upon which success in other studies may be built; it is a prerequisite to good employment and countless other tasks of social adjustment; and, it provides the means by which ideas otherwise locked tightly in one mind may be transmitted to another.

Unfortunately, competent writing is an ability which develops slowly through years of practice. As a former Chief Inspector of Primary Schools in England (cited in Maybury, 1967:19) stated:

No human skill or art can be mastered unless it is constantly practiced. A short composition once a fortnight, interspersed with formal exercises is no good at all. There must be bulk.

Furthermore, "writing is not an easy activity. It involves the total being in a process of learning a more and more complex skill"

(Carlson, 1970:viii-viii). An explanation of this complexity may be found in the dependence of the writing skill upon other, more basic skills. As Moffet and Wagner (1976:10) wrote:

Teachers habitually think of literacy as first or basic, as reflected in the misnomer "basic skill," because the two Rs occur early in the school career and lay the foundation for book learning. But we do well to remind ourselves that reading and writing actually occur last--that is, not only after the acquisition of oral speech but also after considerable nonverbal experience. The three levels of coding . . . mean that experience has to be encoded into thought before thought can be encoded into speech, and thought encoded into speech before speech can be encoded

into writing. Each is basic to the next, so that far from being basic itself literacy depends on the prior codings. It merely adds an optional, visual medium to the necessary, oral medium.

Or, simply, as Chaucer says:

The lyf so short, the craft so long to lerne.
(Parliament of Fowls, 1.1)

Because of the importance of the skill of writing and the time required to attain functional mastery of it, it is essential that teachers have precise information concerning the progress of each student toward attainment of the skill. When teachers are able to evaluate any activity with accuracy and confidence, they are better able to plan for appropriate and effective instruction; when the writing teacher obtains accurate information about a student's writing, that information can provide the basis for initial placement, independent study, remediation, and other administrative and instructional decisions.

But, while the literature contains copious quantities of suggestions and activities for writing, there is a paucity of information regarding the evaluation of writing (Lundsteen and others, 1976). As these authors pointed out (1976:52), "to evaluate something as personal and complex as writing is not a simple matter." Cooper (1975) discussed the difficulties inherent in the evaluation of writing. One problem arises from the difficulty of developing instructional objectives for writing. Cooper (1975:112) felt this was because "writing

instruction has no content, certainly not in the way that biology and algebra have content. And that is the problem with much that has been published recently on measurement of writing--writing is naively considered to be like all the other subjects in the curriculum." Written language--like oral language--is essentially a tool which requires other subjects in order to be put to work.

However, numerous authors have stressed the possibility of measuring the results of any significant educational experience. Ebel (1975:24), for example, stated:

Every important outcome of education can be measured.
 . . . To say that any important educational outcome is measurable is not to say that every important educational outcome can be measured by means of a paper and pencil test. But it is to reject the claim that some important educational outcomes are too complex or too intangible to be measured. Importance and measurability are logically inseparable.

While many educational theorists may disagree with the inclusiveness of this statement, there does, nonetheless, seem to be a considerable gap between what is currently being done in the evaluation of writing and what could--and needs--to be done (Bishop, 1978; Lundsteen, 1976).

Statement of the Problem

The problem of the study was twofold: (1) to determine the reliability of six methods of grading student essays--holistic scoring, atomistic scoring, mature word choice, syntactic complexity, mean T-unit length, and vocabulary diversity--and (2) to compare the

ratings of experienced teachers with those of pre-service teachers using holistic and atomistic methods.

Applications of the Results

The importance of the evaluation of writing becomes apparent when the uses of such evaluation are considered. Cooper and Odell (1977:ix) identified some of these uses.

Administrative

1. Predicting students' grades in English courses.
2. Placing or tracking students or exempting them from English courses.
3. Assigning public letter or number grades to particular pieces of writing and to students' work in an English course.

Instructional

4. Making an initial diagnosis of students' writing problems.
5. Guiding and focusing feedback to student writers as they progress through and English course.

Evaluation and Research

6. Measuring student's growth as writers over a specific time period.
7. Determining the effectiveness of a writing program or a writing teacher.
8. Measuring group differences in writing performance in comparison-group research.
9. Analyzing the performance of a writer chosen for a case study.
10. Describing the writing performance of individuals or groups in developmental studies, either cross-sectional or longitudinal in design.
11. Scoring writing in order to study possible correlates of writing performance.

Clearly, if teachers are to successfully accomplish these tasks, they must have confidence in the methods they use to evaluate the writing of their students. The second chapter of this study discusses only a few of the many aids available to the teacher in his search for

effective measures of ability and growth. Most of these methods have at least some degree of research support, much of which shows individual methods to have high reliability and validity. But despite the existence of these various measurement tools, many teachers are bewildered by the claims of proponents for the various methods (Green, 1963). Thus, when the need arises to select the most appropriate method in a specific situation, teachers have no basis for judgment. As a result, the natural reaction is to continue using what has been used previously.

One major reason for this situation is the dearth of comparative research among various methods. Only one study in the available literature, for example, was directly concerned with establishing the reliability between different procedures, and that study considered only closely related methods.

Before teachers are able to make wise decisions regarding evaluative methods, they will need to be aware of the strengths and weaknesses of each method as well as its correlation with other methods for various purposes. For example, two methods may measure fluency with high reliability but have uselessly low reliability as comparable measures of an overall score. A teacher who substituted one measure for the other to obtain an overall score would be grossly misled in his judgments. This study was a necessary first step in identifying

some general comparability ratings. It also provides a basis for further research in this area.

The benefits derived from any comparative research depend upon the outcome. If the methods are shown to be reliable, teachers may use either to obtain the same results and they will be confident that judgments made on the basis of both measures will be very similar (Thorndike, 1964). Research which could demonstrate such correspondence between methods would be of obvious importance in two key respects. First, teachers would be able to choose the method which is the least time consuming; if both methods give nearly identical results, the one which involves the least amount of class or evaluation time would be selected. Second, school administrators would be able to utilize the method which is most efficient. If one method involves expensive hand scoring while another, comparable method could be machine scored in seconds at little cost, a decision could be made based on economics without compromising educational considerations.

If, however, two methods produce unreliable results, teachers will know either: (1) that both methods do not measure the same thing, or (2) one method is more valid than the other. (While there are many causes of unreliability [Tuckman, 1972], it is assumed that causes such as fatigue, health, memory, etc. will not be factors in the determination of a reliability measure. Then, only elements relating to the methods themselves will be of importance.) In the

first case, further research is indicated in order to identify what each method is actually measuring. Perhaps neither method is valid, or maybe a new factor in writing skill hitherto unrecognized may be isolated. It is possible that research will eventually demonstrate that several factors which contribute to writing success must be measured by different methods. This study provided much needed information by identifying some methods which do measure different factors.

In the second case, subsequent research identifying which method is more valid, would further clarify the aspects of an effective measurement tool and allow for the elimination or improvement of a less effective tool. Again, this study provided a first look at low reliability scores which may result from differing amounts of validity. Such a low reliability score acts as a warning light to all future researchers studying the evaluation of writing; it signals that they must be very careful in their selection of a measurement instrument, for different instruments provide varying degrees of accuracy with reference to the specific trait being measured (Tuckman, 1972).

Another benefit of the study is a direct result of the comparison of preservice and expert raters. The differences and similarities in evaluation patterns between these groups may suggest some changes in teacher training. Teacher education programs should concentrate their time in areas which need practice and study to reach expert levels, while providing confidence in those areas in which students already

perform as experts do. Also, further research may show that some time-consuming grading of papers may be assigned to pre-professionals or aids, freeing teachers for other duties.

Finally, those individual factors which correlate highly with the holistic scoring plan may be identified as principle determiners of writing quality. Planning and teaching should then be directed toward these factors for more efficient instruction; if, that is, these individual factors are largely responsible for good writing, instruction should focus on them rather than on other, more superfluous, factors (Diederich, 1966).

Questions Answered by the Study

This study answered the following questions.

1. What is the rater reliability for each method of evaluation?
2. Does a significant correlation exist between any pairs of methods?
3. Does a significant correlation exist between any method and the combination of other methods?
4. Does a significant correlation exist between any method or methods and specific factors of the same or other methods?
5. Does a significant overall correlation exist between the methods?
6. Do ratings of pre-service English education majors differ significantly from those of identified experts on methods which utilize subjective ratings?

General Procedures

In order to answer these questions, six methods were selected and used to score student papers. The study was conducted from the spring of 1981 to the winter of 1982 and utilized essays of junior and senior

high school students which were scored by: (1) pre-service teachers at Montana State University, (2) expert readers from Montana secondary schools and universities, and (3) the use of four objective methods. The methods differ in many respects such as degree of objectivity, narrowness of focus, number of factors scored, stated purpose, and so forth. Each also is representative of a number of closely related measures. The categories from which methods were selected are: holistic scoring, atomistic scoring, mature word choice, fluency, and vocabulary diversity. (These categories appear throughout the literature. See especially: Lloyd-Jones, 1977; Diederich, 1974; Fowles, 1978; Finn, 1977; Hunt, 1977; and Botel and Granowsky, 1972.) It should be noted that two measures of fluency were used in the study--mean T-unit length and syntactic complexity.

A set of 18 essays formed the corpus for the study. Four groups of raters were used to score these papers. Groups A and B were composed of university professors of English Composition and current or former master secondary public school English teachers. Group A utilized the holistic scoring method and Group B utilized the atomistic method. Groups C and D were composed of pre-service English Education majors and minors. Group C used the holistic scoring method while Group D used the atomistic method. Thus, the papers were scored holistically by a group of experts and by a group of pre-service

teaching candidates. Similarly, they were scored atomistically by experts and by pre-service teaching candidates.

The rater reliability for each group of raters was obtained for each method. Because these reliabilities were high enough to justify further comparisons, correlations between the various methods and groups of raters were computed to answer the questions of significance.

Limitations and Delimitations

A basic limitation of the study was the difficulty of obtaining qualified readers to judge the essays. Because the readers had to be trained together for the holistic method, only teachers from Bozeman, Montana were included in the holistic grading group (Group A).

The selection of a topic for the essays posed another limitation. The difficulties of assembling all raters--both experts and pre-service teachers--in order to reach consensus on an appropriate topic seemed too great to warrant such an effort. Thus, readers were asked to grade a topic which may have held little interest for them and which may, in fact, have been distasteful for them to read. Similarly, the topic may have had little relevance for many readers. A good writing teacher makes assignments that have a purpose--perhaps a merely mechanical one such as checking for subject/verb agreement or perhaps one of higher level such as structural integrity--but some

purpose is usually implicit in the assignment. The obvious lack of a purpose developed by each reader could have influenced rating scores.

A number of delimitations were made in the study. First, the essay sample was derived from a single medium-sized Montana high school. While not totally representative of juniors and seniors in Montana, the sample provided an adequate range and variety of writing. Because it was the methods of evaluation that were tested, not the writing, this condition was considered to be of no consequence.

Second, the groups of expert readers were selected purposively. The persons chosen possessed the precise traits of experience, training, and ability which define the group. Also, the high correlations achieved between raters in other studies (see chapter II) suggests minimum benefit from a random sample design. That is, because all trained expert raters rate very consistently, the selected experts could be expected to typify a larger group of experts.

Third, a serious but requisite delimitation was the necessity of choosing a relatively small number of methods for inclusion in the study. Six relatively distinct methods of evaluation were identified for use. While the major types of evaluation present in the available literature are represented by these methods, there are undoubtedly many possibilities which were excluded. Some generalization to other closely related methods is surely acceptable, but the use of a

greater number of methods would have increased the power of such generalization.

Finally, the study was delimited to include only one mode of writing. Further research will need to be done to compare methods of evaluation in other modes.

Definition of Terms

Definitions for several terms used in this study are required for two reasons. First, there are the usual number of words which may not be familiar to one outside the specific area under investigation. Second, and more importantly, many terms are used by different people to mean different things; it has been necessary, therefore, to define some more common words for purposes of consistency. The following definitions are strictly adhered to throughout the study.

Analytic Scale.--A type of atomistic evaluation. It is a rating scale with three or more points for each feature of writing being rated. Some or all of these points have explicit definitions accompanying them to guide the rater.

Atomistic Evaluation Method.--A technique of evaluation in which specific characteristics within a piece of writing are identified. By combining the ratings of these characteristics, judgments about the whole composition are made. The particular characteristics chosen may or may not be dependent upon the mode of the writing being examined.

Dichotomous Scale.--An atomistic type of evaluation in which a number of statements are listed concerning the presence or absence of certain features in the writing. Responses are binary, being yes/no, present/not present, or similar options.

Essay Scale.--A type of holistic evaluation procedure consisting of a ranked set of essays to which other essays are compared. The essays to be graded are assigned the number of the essay in the scale to which they most closely correspond.

General Impression Scoring.--A type of holistic evaluation in which papers are assigned letter or number grades after a single, rapid reading. At least two raters generally rate each paper, increasing the reliability of the method.

Holistic Evaluation Method.--A technique of evaluation which considers a piece of writing as a whole which should not be divided into its various parts. Such a method examines the composition on its total merit rather than as a sum of several features or characteristics.

Interrater Reliability.--A measure of the degree to which different raters are consistent in their evaluations of some test or attribute. Also called "intraclass correlation."

Mature Words.--Words which appear infrequently in samples from immature writers, but more and more frequently as the maturity of the writer increases. Thus, they may be used to identify mature writers.

Mode.--The form, purpose, and audience of a piece of writing.

Poetry, narrative, business, drama, and expository are a few of the different types of modes.

Syntactic Complexity.--A measure of the complexity of the syntactic structure of a piece of writing. Types of embeddings, phrase modifications, etc. are given different values as based on a transformational-generative grammatical analysis of writing.

Tokens.--The total number of words in a piece of writing.

Topic Imposed Words.--Words which are Mature Words but which, because of the demand imposed upon the writer by the topic, will appear more frequently than expected. For example, "pollute" is a relatively low frequency word and would thus generally be considered as a Mature Word. If, however, a topic were assigned which required, say, a discussion of coal production, the word "pollute" would probably be assumed to be imposed by the topic and thus should not be considered as a Mature Word.

T-Unit.--As defined by Hunt (1977:92-93): "A single main clause (or independent clause, if you prefer) plus whatever other subordinate clauses or nonclauses are attached to, or embedded within, that one main clause. Put more briefly, a T-unit is a single main clause plus whatever else goes with it."

Types.--The number of different words in a piece of writing.

Summary

Writing instruction is an important responsibility of the schools. It also places severe time demands upon the teacher, both in use of class time and the time needed to evaluate papers. This study was undertaken to clarify the evaluation methods available to teachers in three ways:

1. By identifying any differences in scoring which may result from the use of different methods. This could lead to a more precise definition of the specific factors which constitute good writing.
2. By identifying methods which gave comparable results, enabling teachers to use the most temporally efficient method.
3. By establishing whether pre-service teachers rate essays in a manner comparable to the way experts do.

Specifically, the purpose of the study was to determine the comparability of grading student essays by holistic, atomistic, mature word choice, syntactic complexity, mean T-unit length, and vocabulary diversity methods. Scoring by experienced teachers also was compared to that by pre-service teachers using the holistic and atomistic methods.

The study was conducted from the spring of 1981 to the winter of 1982 and utilized student themes obtained from a medium-sized Montana high school. The limiting nature of the raters, the topic selected, the essays themselves, the restriction in the number of methods used, and the use of a single mode of writing were also discussed, and definitions of terms used in the study are given.

CHAPTER II

REVIEW OF LITERATURE

The teacher's ability to measure written composition has grown dramatically over the past twenty years (Bishop, 1978). This growth is evidenced by two factors. First, the number of different types of methods of evaluation has increased substantially. Second, the precision with which these methods may be used has improved as research has defined and enhanced their reliabilities. Both of these factors will be addressed in this chapter. The work of many teachers, theorists, and researchers who have developed widely divergent schemes for evaluating writing will be examined, as will the aspects of research which suggest that each method may be an effective measurement tool. A general discussion of essay grading begins the chapter.

The remainder of the chapter is organized to focus on seven methods of evaluation:

1. holistic scoring
2. atomistic scoring
3. mature word choice
4. syntactic complexity
5. T-unit length
6. type/token ratio
7. standardized "skill" tests

Grading Essay Writing

A considerable amount of disagreement appears in the literature as to the definition of "holistic." The Educational Testing Service was very specific, describing one unique method of rating essay

examinations as "holistic scoring" (Fowles, 1978). On the other hand, Cooper (1977) used the term in a generic sense to identify any of a number of methods of evaluation in which only judgments of quality are made; any method which does not involve counting the specific occurrences of a feature may thus be termed "holistic." For the purpose of efficiently cataloging the various evaluation techniques, it would seem as if the most acceptable definition would fall somewhere between these two extremes. The meaning of the word as it is employed in common usage provides a realistic definition. Thus, a "holistic" scoring method may be considered to be a method which bases its judgment of a piece of writing on the whole composition rather than on a number of separately identified parts (see The American Heritage Dictionary, 1976).

The category of "atomistic" methods, then, subsumes all of those types of evaluation which employ scoring of several distinguishable parts of a composition. The other categories pose no such problems of definition.

Several authors stress the importance of using actual writing to judge writing skills (Cooper, 1977; Lloyd-Jones, 1977; Coffman, 1971). Coffman (1971) identified three reasons for using essays as a measure of writing ability: (1) essay examinations provide a sample of actual written performance and demonstrate a student's ability to use the tools of language, (2) there is presently no alternative method which

effectively measures complex skills and knowledge, and (3) other research shows that students prepare in a different manner for different types of tests, and anticipation of an essay examination produces the greatest achievement as measured by any type of test.

Essentially, then, the preference for essay tests is based on their superior validity--actual writing is being judged rather than answers to objective questions (Cooper, 1977). Such answers do correlate with written scores, but only in the range of .59 to .71 (Goldshalk, Swineford, and Coffman, 1966). Thus, while objective testing of skills may possess some measure of concurrent validity, sampling writing was seen by these researchers as the method with the highest content validity.

Many types of objective evaluation have been developed in attempts to provide self-contained definitions of quality. This is, by applying a certain finite mechanical procedure to a set of writing samples, that set can be ordered according to how each member satisfies the criteria of the procedure. Then, by definition, the sample which receives the highest score is the best piece of writing, and so on down through the entire set. The four objective methods used in the present study can be considered as such procedures. Each can be applied by anyone familiar with the procedure, producing consistent rankings of writing samples.

There seems to be some inherent implausibility in such schemes, however. How, for instance, can an algorithmic approach of the type suggested possibly account for all the nuances of meaning generated by a creative human writer? And how can a certain set of traits--no matter how large that set is--totally define any piece of writing? How, in short, can a finite procedure properly score the infinite set of possibilities available to even the least-sophisticated writer? The answer, of course, is that it cannot. Such procedures must be seen for what they are: measures of certain very specific traits contained within the total piece of writing. This should provide the clue for the best answer to the question: what is quality writing? Quality writing is simply that which recognized experts judge to be quality writing.

This definition may seem at first sight to be question-begging, but upon further reflection it emerges as the only possible, logically defensible definition. There are three basic reasons why this is so. First, no mechanical procedure designed to measure writing can ever do so in a vacuum; that is, it needs as a reference some set of human values. Thus, a degree of subjectivity must always be at the center of any evaluation of a creative human task. No objective measure can hope to capture all the aspects of the inherently subjective task of writing. Second, writing is aimed at a human audience: Its purpose is to transmit ideas and information from person to person. The

ultimate judges of the success of writing must be the members of the audience for which it is intended. Third, those best able to judge any complicated behavior are those with a significant amount of exposure to that behavior. Thus, in the case of the evaluation of writing, those with a substantial degree of experience with writing evaluation would tend to have a broader, more reliable approach to grading; they would have a reservoir of past writing against which to make comparisons. This definition is supported by Cooper's (1977:3-4) statement:

A piece of writing communicates a whole message with a particular tone to a known audience for some purpose: information, argument, amusement, ridicule, titillation. At present, holistic evaluation by a human respondent gets us closer to what is essential in such a communication than frequency counts do.

Since holistic evaluation can be as reliable as multiple-choice testing and since it is always more valid, it should have first claim on our attention when we need scores to rank-order a group of students.

As a result of these considerations, the ratings of the expert group using the holistic method were taken as the best estimates of true writing quality in order to provide a standard against which to measure the various methods employed in this study. To the extent that another method produced results comparable to this group, that method was considered to have provided a more or less accurate representation of the quality of a piece of writing.

Holistic Methods of Evaluation

Holistic scoring provides a way of ranking written compositions. Two common methods of accomplishing such a ranking are: (1) matching a piece of writing to another piece of comparable quality from an already ordered sequence, or (2) assigning the piece a grade in the form of a letter or number based on general impressions of the paper (Cooper, 1977).

The first of these methods employs an essay scale. One such scale is that developed by the California Association of Teachers of English (Nail and others, 1960). The first step in the development of this scale consisted of creating an outline to judge the essays, some of which would ultimately form the scale. The outline consists of three main headings: content, organization, and style and mechanics. While there are subheadings which partially clarify the main headings, no specific definitions or examples of the components of the outline are given: an evaluator must decide, for example, if transitions are adequate, or to what degree all ideas are relevant to the main focus of the essay. The outline is thus seen as merely a guide which enables a judge to keep desirable qualities in mind.

The scale consists of five essays ranked from best to worst and containing proofreaders marks and marginal notes as well as critical comments relating to pertinent aspects of the outline. There is also

a summary of the typical characteristics of themes at each level of the scale.

The Association of English Teachers of Western Pennsylvania has also published an essay scale primarily to provide models for beginning teachers (Grose, Miller, and Steinberg, 1963). It presents samples of poor, average, and good themes at the seventh, eighth, and ninth grade levels. Guides in establishing the scale were: content, form (unity, coherence, and effectiveness), and mechanics. Another publication of the same association provides a similar essay scale for grades ten, eleven, and twelve (Hillard, 1963). The evaluation criteria for the model themes were: "(1) the writer must know what he is talking about and (2) he must evidence a satisfactory degree of control over his writing so that his knowledge of the subject is communicated with precision to the reader" (p. 3).

Another type of holistic scoring is that used by the Educational Testing Service to grade part of the writing sample of its Basic Skills Assessment (Fowles, 1978). This may appropriately be termed "general impression scoring," for it consists of a rating arrived at by a single rapid reading of a piece of writing. In the method, raters use a four point scale to judge the writing. In order to develop the sensitivity of the raters, a training session of 30 to 40 minutes is required. Fifteen to twenty papers typical of the group to be graded are selected as training papers. Because scoring is not

based on any set of pre-existent criteria, this training session serves to develop the raters' abilities to compare papers to each other--the only referents available. Standards evolve from the raters in the course of the training session as they grade papers and revise their personal opinions in light of comments from other raters.

Raters are typically able to read from fifty to sixty papers per hour (each paper approximately $3/4$ of a page). Each paper is read by two raters with the scores added for a total score.

As Lloyd-Jones pointed out (1977), a preference for a holistic scoring scheme is based on either of two assumptions. The first of these is that the whole is more than the sum of its parts. The second, that the parts are too many to be judged independently and may not be easily fit into a formula which will produce a result equal to the whole. Similarly, Fowles (1978:2) stated that in holistic scoring, "the discrete elements are not as important as the total expression of a student's ideas and opinions--that is, the overall quality of the response."

Highly reliable scores are obtainable by this method. A reliability of .95 has been reported for untrained holistic evaluations using five raters (Follman and Anderson, 1967). Cooper (1977) advocated training in holistic techniques to further improve reliability scores, and Coffman (1971:36) explained how such improvement occurs: "In general, when made aware of discrepancies, teachers tend

to move their own ratings in the direction of the average ratings of the group. Over a period of time, the ratings of the staff as a group tend to become more reliable." He also suggested the finer the scale used to rate essays, the higher the reliability will be. A scale of seven to fifteen units seems to be optimum. This method also has high validity, as actual writing is examined in the way it is meant to communicate--that is, as a complete unit.

Atomistic Methods of Evaluation

Several methods of evaluation exist which attempt to identify certain categories within a piece of writing and use these categories to rate the entire composition. A statement from four Indiana college departments of English lists five criteria for evaluating college freshmen in composition courses (Hunting, 1960; and cited in Judine., 1965). The following criteria and guidelines are from that statement.

	CONTENT	ORGANIZATION: Rhetorical and Logical Development	ORGANIZA- TION: Sentence Structure	DICTION	GRAMMAR, PUNCTUATION, SPELLING
Superior (A-B)	A significant central idea clearly defined, and supported with concrete, substantial, and consistently relevant detail	Theme planned so that it progresses by clearly ordered and necessary stages, and developed with originality and consistent attention to proportion and emphasis; paragraphs coherent, unified, and effectively developed; transitions between paragraphs explicit and effective	Sentences skilfully constructed (unified, coherent, forceful effectively varied)	Distinctive: fresh, precise, economical, and idiomatic	Clarity and effectiveness of expression promoted by consistent use of standard grammar, punctuation, and spelling
Average (C)	Central idea apparent but trivial, or trite, or too general; supported with concrete detail, but detail that is occasionally repetitious, irrelevant, or sketchy	Plan and method of theme apparent but not consistently fulfilled; developed with only occasional disproportion or inappropriate emphasis; paragraphs unified, coherent, usually effective in their development; transitions between paragraphs clear but abrupt, mechanical or monotonous	Sentences correctly constructed but lacking distinction	Appropriate clear and idiomatic	Clarity and effectiveness of expression weakened by occasional deviations from standard grammar, punctuation, and spelling
Unacceptable (D-F)	Central idea lacking, or confused, or unsupported with concrete and relevant detail	Plan and purpose of theme not apparent; undeveloped or developed with irrelevance, redundancy, or inconsistency; paragraphs incoherent, not unified, or undeveloped; transitions between paragraphs unclear or ineffective	Sentences not unified, incoherent, fused, incomplete, monotonous, or childish	Inappropriate: vague, unidiomatic, or sub-standard	Communication obscured by frequent deviations from standard grammar, punctuation, and spelling

Taking such guidelines a step further is the well known scale developed by Diederich (1974) which appears below.

	<u>Low</u>		<u>Middle</u>		<u>High</u>
Ideas	2	4	6	8	10
Organization	2	4	6	8	10
Wording	1	2	3	4	5
Flavor	1	2	3	4	5
Usage	1	2	3	4	5
Punctuation	1	2	3	4	5
Spelling	1	2	3	4	5
Handwriting	1	2	3	4	5
				Sum	

The scale grew out of a study conducted in 1961 (Diederich, 1966) which involved the rating of 300 papers by sixty readers. Six different areas were represented by the readers: college English teachers, social science teachers, natural science teachers, writers and editors, lawyers, and business executives. The raters were requested to place each composition in one of nine groups sequenced according to merit. The groups were to contain at least six papers from each of the two topics about which the papers were written. No other instructions or aids were given. Diederich (1966) explained the outcome:

The result was nearly chaos. Of the 300 papers, 101 received all nine grades, 111 received eight, 70 received seven, and no paper received less than five. The average agreement (correlation) among all readers was .31; among the college English teachers, .41. Readers in the other five fields agreed with the English teachers slightly better than they agreed with other readers in their own field.

This procedure has been criticized on the ground that we could have secured a higher level of agreement had we defined each topic more precisely, used only English teachers as readers, and spent some time in coming to agreements upon common standards. So we could, but then we would have found only the qualities we agreed to look for--possibly with a few surprises. We wanted each reader to go his own way so that differences in grading standards would come to light.

Through factor analysis, five clusters of evaluative criteria were identified: (1) ideas, (2) mechanics, (3) organization, (4) wording, and (5) style or "flavor." Diederich then reasoned that if each of these factors were listed and explained, future raters would be able to consider all important aspects of writing more fully and general agreement among raters could be greatly increased. It will be noted that in Diederich's scale (p. 26) four of these criteria are listed singly while the fifth, "mechanics," is further broken down into four subcategories. The scale consists of five points with low, middle and high areas identified. In his 1966 report, Diederich defined these three areas for each part of the scale. For example, the high, middle, and low areas of the "ideas" scale are given below.

1. Ideas

High. The student has given some thought to the topic and has written what he really thinks. He discusses each main point with arguments, examples, or details; he gives the reader some reason for believing it. His points are clearly related to the topic and to the main idea or impression he is trying to get across. No necessary points are overlooked and there is no padding.

Middle. The paper gives the impression that the student does not really believe what he is writing or does not fully

realize what it means. He tries to guess what the teacher wants and writes what he thinks will get by. He does not explain his points very clearly or make them come alive to the reader. He writes what he thinks will sound good, not what he believes or knows.

Low. It is either hard to tell what points the student is trying to make or else they are so silly that he would have realized that they made no sense if he had only stopped to think. He is only trying to get something down on paper. He does not explain his points; he only writes them and then goes on to something else, or he repeats them in slightly different words. He does not bother to check his facts, and much of what he writes is obviously untrue. No one believes this sort of writing--not even the student who wrote it.

"Ideas" and "Organization" are considered most important by many teachers and are thus assigned double values. It should be noted that the Diederich scale "is both qualitative and quantitative; that is, the scale provides for assessing both the quality of ideas and style and the quantitative amount of 'correctness' in such things as grammar, punctuation, and spelling. It is rare to find both factors in a scale." (Lundsteen, 1976:53)

Another analytic scale (in Judine, 1965:159-160) is reproduced on the following page. This scale was developed in a school district in Cleveland Heights, Ohio. Student writers, other student writers, and teachers all use the form for self, peer, and student evaluations.

Lloyd-Jones (1977) found analytic scales such as Diederich's too general and attempted to increase the precision of such scales by insisting one scale be developed for each mode of writing. He termed the result "Primary Trait Scoring." A specific mode of writing is

PURPOSE	
A. Content-50%	
Convincing	Unconvincing
persuasive, sincere, enthusiastic, certain	
Organized	Jumbled
logical, planned, orderly, systematic	
Thoughtful	Superficial
reflective, perceptive, probing, inquiring	
Broad	Limited
comprehensive, complete, extensive range of data, inclusiveness	
Specific	Vague
concrete, definite, detailed, exact	
B. Style-30%	
Fluent	Restricted
expressive, colorful,	
Cultivated	Awkward
varied, mature, descriptive, smooth appropriate	
Strong	Weak
effective, striking, forceful, idioms, fresh, stimulating	
C. Conventions-20%	
Correct Writing Form	Incorrect Form
paragraphing, heading, punctuation, spelling	
Conventional Grammar	Substandard
sentence structure, agreement, references, etc.	

chosen and the characteristics required for successful communication in that mode are identified. Other researchers (Cooper, 1977; Lundsteen, 1976) have also stated the need to develop separate methods to evaluate each different mode. Support for this position includes research showing an increasing variation between ability in various modes of writing in elementary grades as age increased (Veal and Tillman, 1971). The expository (explanation) mode showed the greatest increase in quality through grade levels, while the argumentative mode showed the least increase.

Moslemi (1975) identified creative writing as a unique mode and used a five-point scale to rate four traits: (1) originality, (2) idea production, (3) language usage, and (4) uniqueness of style. Three judges from varied specialties--sociology, English as a foreign language, and English literature--were used in her study. Despite the diversity of background, after a short training period, an inter-rater reliability of .95 was obtained. Other researchers have also found high correlations between judges on rating scales. Folman and Anderson (1967) reported reliability scores for five raters to be .94 for the California Essay Scale, and .93 for the Diederich scale.

Fowles (1978) suggested that the use of analytic scales requires only one rater per paper because of the high reliability factor. The criteria upon which a scale is based make it easy to judge the correctness of response. As a result, no experience is necessary for

raters using an analytic scale. She also pointed out, however, that only certain traits are judged, and that raters must be careful to check details exactly.

Closely related to the analytic scale is the dichotomous scale. Cooper (1977) presented the following a scale for evaluating writing done in a dramatic mode.

	YES	NO	
LANGUAGE	_____	_____	1. Conversation sounds realistic.
	_____	_____	2. Characters' talk fits the situation.
	_____	_____	3. There are stage directions.
SHAPE	_____	_____	4. Stage directions are clear.
	_____	_____	5. Opening lines are interesting.
	_____	_____	6. There is a definite beginning.
CHARACTERIZATION	_____	_____	7. There is a definite ending.
	_____	_____	8. The ending is interesting.
	_____	_____	9. The characters seem real.
MECHANICS	_____	_____	10. The characters are consistent.
	_____	_____	11. The form is consistent.
	_____	_____	12. Spelling rules are observed.
RESPONSE	_____	_____	13. Punctuation rules are observed.
	_____	_____	14. The work is entertaining.
	_____	_____	15. The work made me think about something in a way I hadn't previously considered.
Totals:	_____	_____	

Cooper (1977:9) doubted, however, "whether dichotomous scales would yield reliable scores on individuals, but for making gross distinctions between the quality of batches of essays, they seem quite

promising, though apparently requiring no less time to use than an analytic scale for the same purpose."

Mature Word Choice

Some words in the lexicon occur more frequently than others. The importance of this fact has been recognized for hundreds of years. As Lorge (1944) pointed out, Talmudist scholars have used word counts in their studies of the Torah since at least 900 A.D. For them, the significance of the appearance of a rare word was a subject of considerable interpretation. The first large list of word frequencies compiled in the United States is The Teacher's Word Book of 30,000 Words (Thorndike and Lorge, 1944). This book is a listing of four separate word counts which represent a total sample of approximately 18 million words. The most current word list is the Word Frequency Book compiled by Carroll, Davies, and Richman (1971). The authors extracted over five million words of running text from more than 1000 publications. The texts used included textbooks, workbooks, kits, novels, poetry, general non-fiction, encyclopedias, and magazines. The project was undertaken to provide a lexical basis for the American Heritage School Dictionary. From the word list thus obtained, an index of the frequency of occurrence was generated for each word. This is called the Standard Frequency Index (SFI) and is defined as

$$\text{SFI} = 10(\log_{10} \delta + 10)$$

where δ is the ratio of the number of tokens of a word type to the total number of tokens as that number increases indefinitely. A sample word and its probability of occurrence is given in Table 1 for several levels of SFI.

Table 1
Interpretation of the Standard Frequency Index

SFI	Probability of the Word's Occurrence in a Theoretical Indefinitely Large Sample		Example of a Word with Designated SFI
90	1 in every	10 words	the (88.7)*
80	1 in every	100 words	is (80.7)
70	1 in every	1,000 words	go
60	1 in every	10,000 words	cattle
50	1 in every	100,000 words	quit
40	1 in every	1,000,000 words	fixes
30	1 in every	10,000,000 words	adheres
20	1 in every	100,000,000 words	cleats
10	1 in every	1,000,000,000 words	votive (12.7)

*Where no word has the designated SFI, the SFI of the closest word appears in parentheses.

Finn (1977) utilized the SFI as an index of mature word choice. The index was applied to 101 themes written by students in grades 4, 8, and 11 which provided a data base of approximately 15,000 words. He discussed two themes and showed that one contains a greater number of mature words than does the other. His analysis demonstrates an attempt to use a word frequency count as the basis for an objective measure of maturity in word selection.

Fluency

Measures of fluency are designed to provide objective data concerning aspects of syntactic structure; the ways in which a writer puts words together can provide an indication of the degree of control which that writer has over the structural forms of language. Researchers who attempt to define this control try to objectively measure one or more of these structural forms. As Endicott (1973:5) stated: "That people tend to perceive and process language in terms of units of some kind seems obvious, but what these units are and how they are perceived are questions that have not been resolved."

One such unit which has received much use is the T-unit. Hunt (1977:92-93) defined the T-unit as "a single main clause (or independent clause, if you prefer) plus whatever other subordinate clauses or nonclauses are attached to, or embedded within, that one main clause. Put more briefly, a T-unit is a single main clause plus whatever else goes with it."

Since its development, many researchers have employed the T-unit as a measure of fluency (Hunt, 1977; Gebhard, 1978; Dixon, 1971; Belandger, 1978; Fox, 1972). Hunt showed in his 1965 study that mean T-unit length tends to increase as students get older. Cooper (1975) found that an increase of .25 to .50 words per T-unit per year has been shown to be a normal growth.

With the T-unit as an example, several researchers have extended the investigation of syntactic structure in both breadth and depth. For example, other measures of syntactic structures were proposed by Christensen (1968). He claimed the developmental studies of Hunt and others are leading teachers in the wrong direction. Hunt's studies suggest that the more complex a piece of writing, the more mature the writer. Christensen suggested this is not necessarily the case and proposed that it is not sheer complexity that ought to be taught, but rather proper use of structures. In a study of non-professional, semi-professional, and professional writers, Christensen investigated structures which he termed "free modifiers" and "base clauses." A free modifier is a structure which modifies constructions rather than individual words (such a modifier is "bound"). The total number of words in free modifiers as well as their position within a T-unit were found to be significant indexes of writing quality. A base clause of a T-unit is what is left when the free modifiers are removed. The mean length of base clause was also found to be significant. Nemanich (1972) indicated that there is a significant increase in the use of the passive voice between students in grade 6 and adult professional writers.

Following the lead of these researchers who have focused on one or two indicators, many researchers have combined several syntactic units into a single measure of syntactic complexity. Endicott (1973)

used psycholinguistic terms to develop a model of syntactic complexity. He defined a complexity ratio which depends upon certain syntactic operations and transformations.

Botel and Granowsky (1972) developed a formula for determining syntactic complexity in order to measure the syntactic component of writing. Their primary concern was to provide a new method of judging readability. Various structures are assigned values on a scale of 0 to 3 and the sum of these values is then divided by the number of sentences to provide the complexity score. The scoring guidelines follow.

Summary of Complexity Counts

0-Count Structures

Sentence Patterns - two or three lexical items

1. Subject-Verb-(Adverbial): He ran. He ran home.
2. Subject-Verb-Object: I hit the ball.
3. Subject-be-Complement-(noun, adjective, adverb): He is good.
4. Subject-Verb-Infinitive: She wanted to play.

Simple Transformations

1. interrogative (including tag-end questions): Who did it?
2. exclamatory: What a game!
3. imperative: Go to the store.

Coordinate Clauses joined by "and": He came and he went.

Non-Sentence Expressions: Oh, Well, Yes, And then

1-Count Structures

Sentence Patterns-four lexical items

1. Subject-Verb-Indirect Object-Object: I gave her the ball.

2. Subject-Verb-Object-Complement: We named her president.

Noun Modifiers

1. adjectives: big, smart
2. possessives: man's, Mary's
3. predeterminers: some of, none of.... twenty of
4. participles (in the natural adjective position): crying boy, scalded cat.
5. prepositional phrases: The boy on the bench...

Other Modifiers

1. adverbials (including prepositional phrases when they do not immediately follow the verb in the SVAdv. pattern.)
2. modals: should, would, must, ought to, dare to, etc.
3. negatives: no, not, never, neither, nor, -n't
4. set expressions: once upon a time, many years ago, etc.
5. gerunds (when used as a subject): Running is fun.
6. infinitives (when they do not immediately follow the verb in a SVInf. pattern): I wanted her to play.

Coordinates

1. coordinate clauses (joined by but, for, so, or, yet): I will do it or you will do it.
2. deletion in coordinate clauses: John and Mary, swim or fish. (a 1-Count is given for each lexical addition)
3. paired coordinate "both . . . and": Both Bob did it and Bill did it.

2-Count Structures

Passives: I was hit by the ball. I was hit.

Paired conjunctions (neither...nor, either...or): Either Bob will go or I will.

Dependent Clauses (adjective, adverb, noun): I went before you did.

Comparatives (as...as, same...as, -er than...., more...than) He is bigger than you.

Participles (ed or ing forms not used in the usual adjective position): Running, John fell. The cat, scalded, yowled.

Infinitives as Subjects: To sleep is important.

Appositives (when set off by commas): John, my friend, is here.

Conjunctive Adverbs (however, thus, nevertheless, etc.): Thus, the day ended.

3-Count Structures

Clauses used as Subjects: What he does is his concern.

Absolutes: The performance over, Mr. Smith lit his pipe.

Golub and Kidder (1974) stated that while both Hunt's T-unit measure and Botel and Granowsky's syntactic complexity formula do indeed provide relevant data, both are time consuming and tedious. A measurement tool is needed which can be easily used and which will define specific structures that can be taught to increase writing maturity. The authors reported a study in which sixty-three structures were subjected to multivariate analysis and ten variables which correlated highly with teacher ratings were assigned weights through canonical correlation analysis. This research led to the development of the following tabulation sheet which allows calculation of a Syntactic Density Score (SDS) (Golub, 1973).

SYNTACTIC DENSITY SCORE

Number	Description	Loading	Frequency	L	X	F
	Total number of words					
	Total Number of T-units					
1.	Words/T-Unit	.95	X	_____	_____	_____
2.	Subordinate clauses/T-unit	.90	X	_____	_____	_____
3.	Main clause word length (mean)	.20	X	_____	_____	_____
4.	Subordinate clause word length (mean)	.50	X	_____	_____	_____
5.	Number of Modals (will, shall can, may, must, would. . . .)	.65	X	_____	_____	_____
6.	Number of Be and Have forms in the auxiliary	.40	X	_____	_____	_____
7.	Number of Prepositional Phrases	.75	X	_____	_____	_____
8.	Number of Possessive Nouns and	.70	X	_____	_____	_____
9.	Number of Adverbs of Time (when, then, once, while. . . .)	.60	X	_____	_____	_____
10.	Number of gerunds, participles, and absolute phrases (unbound modifiers)	.85	X	_____	_____	_____
	TOTAL			_____	_____	_____
	SDS: S.D. Score (Total/No. of T-units)			_____	_____	_____
	Grade Level Conversion			_____	_____	_____

Grade Level Conversion Table:

SDS	.5	1.3	2.1	2.9	3.7	4.5	5.3	6.1	6.9	7.7	8.5	9.3	10.1	10.9
Grade														
Level	1	2	3	4	5	6	7	8	9	10	11	12	13	14

The authors have programmed the SDS on computer with results very closely correlated to hand tabulation. Thus the goal of ease of use together with teachably defined structures is attained.

Belanger (1978) pointed out that the SDS is dependent upon the length of the writing sample, since variables 1 to 4 would be constant whatever the number of T-units, while variables 5 to 10 are gross scores and would vary with the length of the writing sample. To add these two groups of scores is to add variables of two different

types.' As a solution, Belanger suggested dividing variables by 10 rather than by the number of T-units.

Gebhard (1978) sought to measure the use of syntactic structures among groups of college freshmen and professional writers, and to determine how writing from these two groups compared. All measures of fluency used--sentence length, clause length, and T-unit length--indicated significant differences between freshmen and professional writers. Sentence combining transformation length--a measure of syntactic complexity--also gave significant results. The rest of the study provides uncertain results. Gebhard (1978:230) concluded:

Unfortunately, perhaps, the results of this study point in no specific direction for instruction in the improvement of syntax. In other words, outside of a few structures, such as coordinate conjunction sentence beginnings and extensive use of prepositional phrases, it is not possible on the basis of the results realized here to say, "It is clear that professional writers use these syntactical forms as do highly rated freshmen. Go, therefore and teach the use of these devices."

Such a simplistic solution to the problem of composition teaching is not at hand. If this study testifies to anything, it seems to this researcher, it testifies to the organic and holistic nature of the written communication act. The better freshman has internalized the dialect of written English to a greater extent than his less able classmate.

O'Donnell (1976) provided a useful history of attempts to measure syntactic complexity over the past forty years.

Vocabulary Diversity

Most authors recognize good vocabulary as an essential ingredient of quality writing. Vocabulary choice is a significant part of Diederich's (1974) category of "Wording," while Page (1966) gave "aptness of word choice" as an example of an intrinsic variable which is of importance to a grader of writing. Other researchers include word choice under categories of "diction," "usage," "fluency," and so on.

A number of approaches may be taken to obtain a quantitative measure of vocabulary diversity. The easiest may be to simply total the number of words written; i.e., find the number of tokens. Similarly, one may determine the number of different words appearing in a writing sample; i.e., find the number of types. Unfortunately, both of these techniques produce results dependent upon the length of the writing sample--a long paper consisting mostly of common poorly selected words would be rated better than a short paper with a distinct, carefully chosen vocabulary. A simple attempt to correct this problem might seem to be a ratio of types to tokens. Carroll (1964) demonstrated that this measure too is a function of length. "Sometimes the type-token ratio (the number of different words divided by the number of total words) is used as a measure of the diversity of richness of vocabulary in a sample, but it should be noted that this ratio will tend to decrease as sample size increases, other things being equal,

because fewer and fewer of the words will not have occurred in the samples already counted" (Carrol, 1964:54).

To offset the effect of sample length, Carroll went on to suggest a different measure. "A measure of vocabulary diversity that is approximately independent of sample size is the number of different words divided by the square root of twice the number of words in the sample." Another technique was presented by Herdan (1964). He defines

$$K = \frac{\sum f_x X^2}{[\sum f_x X]^2} - \frac{1}{N}$$

where f_x is the number of vocabulary items with frequency X . $1/K$ is then interpreted as an index of diversity. Herdan also stated that word usage cannot be viewed theoretically as following a random distribution, but that the use of any word influences to some degree the choice of subsequent words.

In the Word Frequency Book, (Carroll, Davies, and Richman, 1971) an "Index of Diversity" is defined as

$$D = m/s$$

where m is the mean and s is the standard deviation of the logarithmic distribution of probabilities of word types. D is independent of sample length. This formula is closely related to the formula given above by Herdan.

Standardized Tests

While standardized tests were not used in the study, a discussion of such tests is nonetheless useful for two reasons. First, many of the characteristics--good and bad--of standardized tests can be identified in some of the methods used in the study. Second, many of the more recently developed methods of evaluating essays--holistic scoring and primary trait scoring as examples--were created in reaction against standardized tests (Applebee, 1981). It is useful to have an awareness of this historical development when such methods are studied.

Many experts are critical of standardized tests. Moffett and Wagner (1976) stated that standardized tests are designed to eliminate the bias inherent in teacher-made tests. In order to do this, the score a student receives on such a test must be compared to the scores of other students. There are two principle ways of making this comparison. First, most standardized tests are norm-referenced; that is, some "normal" population is used as a basis against which other scores are referenced. The authors (1976:431-432) questioned the efficacy of such a comparison, stating that a student's

only reason for comparing his performance with other's would be to know where he stands in the eyes of adults manipulating his destiny. For your own diagnosing and counseling purpose, comparison among students has no value.

The second method of comparing students is by criterion-referencing; that is, students are measured against an established standard rather than against a normal population of other students. Criterion-referenced tests are usually designed to be politically safe--they utilize only minimal standards in order that most students will pass. Moffet and Wagner (1976:432) further state:

Obviously there is a kind of score group in the minds of the test-makers, only it is not a particular population actually run through a particular test but rather a general notion of what most students have done, and can do, based on common school experience. . . .

. . . .In short, criterion-referencing differs not so much from norm-referencing as might appear at first blush, because both set low standards based on moving large masses a short way. This low center of gravity owes to the misguided practice of treating all students at once in the same way, of standardizing.

Thus, Moffett and Wagner concluded that standardized tests are not the proper instruments to measure a student against himself.

Some of the same criticisms were made by Applebee (1981). He stated that standardized tests of vocabulary, usage, and reading comprehension are easy to administer and score, and are highly correlated with writing ability. However, he also identified two major deficiencies of such tests. First, the higher level skills of development, organization, and structure are not measured by standardized tests; second, teachers should be teaching usage exercises rather than writing if usage were actually a direct measure of writing ability.

In an assessment of the accountability system used in the Michigan schools (House, Rivers, and Stufflebeam, 1974:668), the investigators found that standardized tests were not reliable indicators of learning.

Contrary to public opinion, standardized achievements tests are not good measures of what is taught in school. In addition, many other factors outside school influence them. Even on highly reliable tests, individual gain scores can and do regularly fluctuate wildly for no apparent reason by as much as a full grade-equivalent unit.

Summary

Many evaluation strategies are available in the literature. The preference for using actual writing samples as opposed to objective tests was discussed, and a number of evaluation strategies were grouped into seven categories: holistic scoring, atomistic scoring, mature word choice measurement, syntactic complexity, T-unit length, type/token ratio, and standardized tests.

Several methods of holistic and atomistic scoring were discussed along with research supporting the reliability and usefulness of each. Of particular interest were the ETS system of holistic scoring, the analytic scale of Diederich, and the concept of developing unique rating scales for each mode of writing advanced by Lloyd-Jones and others. Examples of selected charts and grading scales were presented.

Similarly, several objective types of evaluation were discussed. Mature word choice measurement can provide an index of the number of

less frequently used words as a sign of mature writing; a measure of syntactic complexity gives an indication of the degree of sophistication of syntactic structure; T-unit length provides a measure of a specific type of syntactic structure that has been the basis of much research; and a type/token ratio can indicate the degree to which a writer varies his word selection. Finally, standardized tests tend to measure things other than writing skills.

CHAPTER III

METHODS

This chapter describes the procedures which were used in order to determine the reliability of six methods of grading student essays. The specific methods used were representative of the categories of holistic scoring, atomistic scoring, mature word choice, syntactic complexity, T-unit length, and type/token ratio which were described in Chapter II. The Holistic and Atomistic methods are given considerable attention because of the need to obtain several subjective rater opinions. The other four methods utilize objective scoring procedures and may be scored by anyone familiar with the various instruments or formulae.

Also in this chapter the general questions investigated are transformed into specific null hypotheses and alternative hypotheses. The results of the study were determined from these hypotheses and the outcomes of the various statistical procedures described.

Essay and Rater Descriptions

The essays used in the study were obtained from two English Composition classes at the high school in Belgrade, Montana. Both classes contained a mix of juniors and seniors. The students were given one fifty-minute period to write extemporaneously about the following paragraph.

Imagine that a large company near you has been found to be seriously polluting a local river. Some people have been

talking about closing the company down until something can be done about the pollution. If the company is closed down, many people will be out of work. Write your feelings about whether to shut down the company. Be sure to indicate why you feel the way you do.

This paragraph was chosen because of its use in previous research (Finn, 1977:71) and also because it seemed to provide both a legitimate point of focus and direction and a sense of open-endedness which would allow students to give divergent responses.

Twenty-nine students completed the assignment. One paper was written in outline form and was removed from the study. It was felt that because the other papers were all written in a normal prose format, the obvious difference in form of this paper might contribute to scoring differences.

From the remaining 28 papers, 10 were removed to be used as training papers for the groups using the holistic scoring method. Eighteen papers were then left for actual rating purposes. The training papers were selected to represent the same approximate range of quality as the remaining papers as suggested by ETS (Fowles, 1978). To accomplish this, this researcher and two other experienced teachers rated the 28 papers on a five point scale. The average score for each paper was determined, and the 10 training papers were selected to represent a sample from the entire body of papers.

The papers were then entered into a computer file and typed (by line printer) copies were made for use by the various raters. As

Gebhard (1978) suggested, the typing of papers eliminates the "halo" effects of handwriting. Initially, spelling errors were to have been corrected as well, but the difficulty of distinguishing spelling errors from usage errors as well as the desire to keep the samples as much like actual writing as possible led to a precise transcription of the papers.

Four groups of raters were used in the study. Groups A and B were each composed of 10 expert readers of English composition. These readers were identified as experts on the basis of their education and experience. All have degrees in English or education with an emphasis in English, and all have taught in the public schools. These groups were composed of university professors of English or education, and master secondary school English teachers. Groups C and D consisted of English education majors and minors at Montana State University enrolled in an undergraduate English methods course entitled "English and the Teaching of Composition" during spring quarter 1981 and winter quarter 1982, respectively. A major goal of this course was to train students in the evaluation of writing. Most of these students were seniors preparing to teach within a year. Group C consisted of 14 students, while Group D consisted of 10 students.

One group of expert readers (A) and one group of pre-service teachers (C) performed a holistic evaluation. The other group of

experts (B) and the other group of pre-service teachers (D) performed an atomistic evaluation.

It was necessary to control one major contaminating variable--the rater groups must be comparable for each type of rater (i.e., expert or student) in order for the effects of the method to be examined. The two expert rater groups were matched by experience and educational level. The two students groups were matched on the basis of class, cumulative grade point average, and experience. All students had completed equivalent lower division pre-requisite courses and performed their grading procedures during the final week of the quarter when enrolled in the methods course. In addition, the same professor taught the methods course in the same fashion both quarters. A t-test was used to determine the significance of the difference between the mean grade-point averages of the two groups.

In the proposal for this study, the comparability of the student groups was to have been determined by the use of a standardized test of English skill as suggested by Follman and Anderson (1967). This would have required the use of a second class period from the English methods course, however, and the instructor of the course was not willing to give up this additional time. Therefore, the method described above was selected as the only feasible way to establish comparability.

Contaminating variables left uncontrolled include many factors of the raters: sex, age, job goals, personality, etc. Also, the choice of topic and mode of the writing sample have been selected without regard to the feelings of the raters. This is discussed more fully in the "Limitations and Delimitations" section of Chapter I.

Categories of Investigation

The categories under investigation in the study were the six methods of rating essays. These are: (1) Holistic scoring, which requires readers to rate essays based on a single, rapid reading; (2) Atomistic scoring, which provides raters a list of factors of which to rate essays; (3) Mature word choice score, an objective system of determining the maturity level of word usage; (4) Syntactic complexity score, an objective measure of the level of sophistication of syntactic usage; (5) An objective method of determining mean T-unit length; and (6) An index of the ratio of types to tokens. These methods were selected because each provides a different way of measuring writing. Undoubtedly, many other methods could have been selected in addition to or in place of these methods; the methods used, however, represented those most frequently cited in the available literature by testing services, writing evaluation guides and researchers.

Method of Data Collection

The holistic evaluation required readers to assign a score of 1 (highest) to 5 (lowest) to each paper. There is much variation in the literature concerning the number of points in the rating scale. ETS (Fowles, 1978) used a scale of 1 to 4, while the National Assessment of Educational Progress (Writing Mechanics, 1975) used a scale of 1 to 8. Several authors have suggested a three-point scale (Thomas, 1966; Hillard, 1963; Grose, Miller, and Steinberg, 1963), and many others a five-point scale (Ward, 1956; Blackie, 1965; Green and others, 1960). The five-point scale was chosen on the basis of the literature as well as its compatibility with the traditional grading system (i.e., A, B, C, D, F).

Both groups using the holistic scoring method required a short training session which followed the procedure suggested by Fowles (1978). This session utilized the ten training papers previously removed. These were graded by the researcher and two assistants, as explained above, using the same five-point scale which was used by the raters. A paper with a score of 3 was the first training paper used and the remainder of the papers were placed in random order after it. At the training session, copies of each paper were given one at a time to the readers. Each reader read and scored the paper without discussion. When all readers had scored a paper, the scores were marked on the blackboard and the paper was discussed.

The training sessions for both groups of holistic raters were run in the same manner. The researcher compiled a list of instructions which he read to the groups. In both sessions, when six of the training papers had been scored and discussed, the scores clustered at two points, allowing the groups to move into the actual rating phase of the session as recommended by Fowles.

Raters were instructed that there was to be no discussion during the rating period and were also asked to use each score in the 1 to 5 range at least once among the 18 papers. This was to insure that all raters would utilize the full range of the scale.

The groups using the atomistic scoring method used the following scoring sheet for each paper and the descriptions of each factor as found in Diederich (1974). It should be noted that the rating form used is the Diederich (1974) analytic scale without the "Handwriting" factor.

	<u>Low</u>		<u>Middle</u>		<u>High</u>
Ideas	2	4	6	8	10
Organization	2	4	6	8	10
Wording	1	2	3	4	5
Flavor	1	2	3	4	5
Usage	1	2	3	4	5
Punctuation	1	2	3	4	5
Spelling	1	2	3	4	5

The papers given to each rater in each of the four groups were randomly ordered. Thus, every rater went through the papers in a different sequence.

The mature word choice scores were obtained through a procedure based on that described by Finn (1977). First, a computer was used to count the number of occurrences of each word in all themes as well as to calculate the total number of types and of tokens. The same information was then produced for each individual theme. For consistency, different graphic forms of the same word were combined; that is, obvious misspellings - e.g., "polution" - were corrected before the lists were compiled.

Also, spaced words were combined (water ways), connected words were separated (alot), and homophones were corrected (their/there/they're). In most cases, if a word exists and was neither an obvious spelling error nor a homophone error, it was left--even if used improperly semantically--as it appeared. The word frequency list of Carroll, Davies, and Richman (1971) was used to determine the Standard Frequency Index (SFI) of all words. This is a measure of the frequency with which a word would be expected to occur. (Refer to Chapter II, table 1 for the probability and a sample word for various levels of SFI). Mature Words were considered to be those with an SFI less than 50.

The list of the number of occurrences of all words was used to determine which words were "Topic Imposed Words," that is, those which may be of low frequency but which were demanded by the topic and thus should not be counted in the mature word category (see Finn, 1971).

Those Mature Words which appeared five or more times in the sample themes were considered Topic Imposed and were eliminated from the count of Mature Words. Also eliminated from the count of Mature Words were proper nouns, slang, contractions and numerals. For each theme, the number of words or "tokens" and the total number of Mature Words (after eliminating the above listed categories) was counted. The Mature Word Index (MWI) for a paper is then the adjusted frequency of Mature Words divided by the number of tokens.

The Syntactic Complexity score was obtained by using the Syntactic Complexity Formula developed by Botel and Granowsky (1972), which is explained in Chapter II.

The Mean T-Unit Length for each paper was obtained by first determining the number of T-units in the paper and then dividing this number by the number of tokens.

The Type/Token Index was determined by Carroll's (1964) formula

$$\frac{T}{\sqrt{2N}}$$

where T is the number of types and N is the number of tokens.

Statistical Hypotheses

The specific null hypotheses which were tested and the alternative which was selected for each in the event of rejection of the null are listed below.

1. Null--No significant correlation exists between scoring method A and scoring method B (where A and B are replaced by all possible distinct pairs of the six scoring methods).

Alternative--A significant positive correlation exists between scoring method A and scoring method B (where A and B are replaced by all possible distinct pairs of the six scoring methods).

2. Null--No significant correlation exists between factor A and scoring method B (where A is replaced by each factor of the modified Diederich analytic scale in all possible combinations with B which is replaced by each of the methods).

Alternative--A significant positive correlation exists between factor A and scoring method B (where A is replaced by each factor of the modified Diederich analytic scale in all possible combinations with B which is replaced by each of the methods).

3. Null--No significant correlation exists between scoring method A and the sum of the rankings of the other five methods (where A is replaced by each of the methods).

Alternative--A significant positive correlation exists between scoring method A and the sum of the rankings of the other five methods (where A is replaced by each of the methods).

4. Null--No significant inter-method correlation exists among the six methods of essay scoring.

Alternative--A significant overall correlation exists between the six methods of essay scoring.

5. Null--No significant difference exists between ratings by pre-service English teachers and expert readers using the holistic scoring method.

Alternative--A significant difference exists between ratings by pre-service English teachers and expert readers using the holistic scoring method.

6. Null--No significant difference exists between ratings by pre-service English teachers and expert readers using the atomistic scoring method.

Alternative--A significant difference exists between ratings by pre-service English teachers and expert readers using the atomistic scoring method.

Analysis and Presentation of Data

All four groups of raters (i.e., the student and expert groups used for the holistic and atomistic methods) were tested for inter-rater reliability. Ebel (1951) described two formulae for such intra-class correlation. One yields the reliability coefficient for average ratings, while the other produces a reliability coefficient for individual ratings. The choice of formula depends upon the use of the coefficient.

If decisions are based upon average ratings, it of course follows that the reliability with which one should be concerned is the reliability of those averages. However, if the raters ordinarily work individually, and if multiple scores for the same theme or student are only available in experimental situations, then the reliability of individual ratings is the appropriate measure. (Ebel, 1951:408)

The available sources either do not mention the individual reliability formula at all or stress the results obtained from the average reliability formula (Follman and Anderson, 1967; Cooper, 1977).

Because the decisions regarding the correlation of the six scoring methods were based on average ratings for the holistic and atomistic methods, the use of only the average reliability formula can be justified. However, as multiple ratings do not generally occur outside of

an experimental setting, the individual reliability formula can also be justified. Ebel's criteria thus produce ambiguous results, so the reliability for both individual and average scores are listed.

Ebel's (1951) formula for the reliability of average ratings is

$$r = \frac{M_p - M_e}{M_p}$$

and his formula for the reliability of individual ratings is

$$r = \frac{M_p - M_e}{M_p + (k-1)M_e}$$

For both formulae, M_p is the mean square for papers, M_e is the mean square for error, and k is the number of raters. In order to eliminate the adverse effects upon the correlation coefficient arising from a difference in the level of ratings between raters, the "between-raters" variance was removed from the error term as suggested by Ebel (1951).

It should be noted that a Kuder-Richardson (1937) formula is suggested in a later reference of Ebel (1972) as a means of calculating the reliability coefficient. This formula produces results equal to the method of intraclass correlation used to determine the reliability of average ratings found in Ebel's earlier work. McNemar (1969) also gave a helpful discussion of the intraclass correlation method. Both the Ebel and the Kuder-Richardson procedures involve the analysis of variance by which the variation in scores between essays

is compared to the variation within essays. This further supports the usage mentioned above of the average as opposed to the individual reliability coefficient.

Each method of evaluation provided a single score for each paper. For the holistic and atomistic methods (i.e., those methods using raters) the average score of all raters was used, while the other methods produced one score per essay by design. The scores from each method were correlated with the scores from each of the other methods. There are two correlations each for the holistic and atomistic methods--one using the expert readers and one using the pre-service teachers. Thus, an eight-by-eight correlation matrix presented when the results are discussed in Chapter IV, shows all correlations in a manner similar to the five-by-five matrix of Follman and Anderson (1967). The Pearson product-moment correlation coefficient was used to determine these correlations.

The scores obtained from each method (and rater group) were then used to obtain rank orderings of essays for each method (and rater group). These rankings were correlated by pairs using Spearman's coefficient of rank correlation and are also displayed in an eight-by-eight correlation matrix in Chapter IV.

The modified Diederich rating scale used in the study permitted the easy identification of several traits which were determined to be important to writing quality by means of factor analysis (see Chapter

II). These factors were correlated with each method using the Pearson correlation coefficient. The results appear in a seven-by-eight matrix (i.e., factors by methods and rater groups). Another such matrix is used to display the Spearman rank correlations of factors and methods. These matrices appear in Chapter IV.

Correlations were also made between each method and the sums of scores of all other methods. Again, the Pearson and the Spearman rank correlation methods were used.

The rankings were also used to obtain measures of the overall agreement of the different methods. There were two such comparisons, one using the expert rater groups and the other using the pre-service teacher groups. Both groups were compared with all objective methods. Kendall's coefficient of concordance was used for this purpose.

Finally, analyses of variance were performed to test for significant differences between the pre-service and the expert readers for the holistic and atomistic methods.

The t-test used for the student groups was tested for significance using a two-tailed test, while Pearson product-moment correlation coefficients were tested for significance using one-tailed tests (Nie and others, 1975). Spearman coefficients of rank correlation were also tested for significance using one-tailed tests of t as described by Nie and others (1975). Kendall coefficients of concordance were tested for significance by a Chi Square test as suggested

by Ferguson (1976). Analyses of variance were tested for significant F ratios as described in SPSS (Nie and others, 1975).

All of the above correlations and analyses of variance were tested for significance at the .05 level. Tuckman (1972:224) identifies the .05 level as "an arbitrary level that many researchers have chosen as a decision point in accepting a finding as reliable or rejecting it as sufficiently improbable to have confidence in its recurrence." This level permits clear relationships to be recognized. A more stringent requirement in early research might have lessened the chance of identifying the relationships to be explored in further research.

The t-test used to demonstrate student group comparability was tested for significance at the .10 level. The significance level was raised in this case because of the severe consequences of a type II error: the two student groups would have been assumed to have been comparable when they were not, and this invalid assumption would have influenced the conclusions drawn in the rest of the study. A type I error on the other hand, would have forced the selection of other rating groups but would not have influenced the conclusions of the study.

Calculations

All basic calculations were performed by computer. The word lists used for the MSI measures and the calculation tokens were done

with programs written by the researcher (see Appendix A). Analyses of variance were calculated using the SPSS (Nie and others, 1975) subprogram "ANOVA." Pearson product-moment correlation coefficients were calculated using the SPSS subprogram "PEARSON CORR," and the Spearman coefficients of rank were calculated using the SPSS subprogram "NONPAR CORR." Interrater correlations and Kendall coefficients of concordance were calculated using programs written by the researcher. These programs also appear in Appendix A.

Summary

The specific procedures used in the study were discussed in this chapter. Essays were selected from two classes of juniors and seniors at the high school in Belgrade, Montana and were typed to remove the contaminating effects of handwriting. Groups of expert raters and pre-service teacher raters were selected to participate in the holistic and atomistic scoring methods. These two methods together with mature word choice, syntactic complexity, mean T-unit length, and the type/token index formed the categories of the study.

Null and alternative hypotheses were stated which provided the statistical reference points needed to answer the general questions of the study. The grade point averages of the students were analyzed using a t-test to determine the comparability of the two student groups. Holistic and atomistic methods were tested for interrater reliability using Ebel's (1951) formula, and pairs of methods were

correlated both by raw score (Pearson correlation) and by rank (Spearman correlation). The factors of the Diederich scale were correlated with methods using Pearson correlation, and two overall measures of concordance were determined--one for the expert groups, the other for the student groups--using Kendall's coefficient. In addition, the expert raters were compared to the pre-service teachers using analyses of variance.

CHAPTER IV

RESULTS

The results of the study are presented in this chapter. First, the two groups of student raters are shown to be comparable. This is necessary in order that further tests which depend upon this comparability can be interpreted meaningfully. Next, the intraclass correlations for the Holistic and Atomistic methods and all categories of the Atomistic method are presented for both the student and expert groups. Then, the various Pearson correlations of scores and the Spearman correlations of ranks are shown followed by the Pearson and Spearman correlations for each score with the average of all other scores. Next, the Kendall coefficients of concordance are examined for overall correlation between methods. Finally, the analyses of variance between rater groups and methods are discussed. The appropriate statistical hypothesis is addressed in each of these sections.

Comparability of Student Rater Groups

The first task of the study was to show that the two groups of student raters were equivalent. In order to demonstrate this, several facts were considered. First, all of the students were English majors or minors and had similar backgrounds in terms of college level coursework in English. Second, the students were enrolled in a junior level course entitled "Composition and the Teaching of English" and had

satisfied all of the prerequisites for this course. Finally, the grade point averages for the students were obtained and the mean grade point averages of the groups were tested for significance by a t-test. The results are shown in Table 2.

Table 2
Comparison of Grade Point Averages for Student
Groups Using Holistic and Atomistic Scoring

Method of Scoring		
	Holistic	Atomistic
	3.76	3.86
	3.61	3.73
	3.59	3.54
	3.54	3.20
	3.30	3.13
	3.23	2.96
	3.18	2.87
	3.17	2.84
	2.99	2.32
	2.91	1.98
	2.87	
	2.58	
	2.50	
	2.27	
Mean	3.11	3.04
St. Dev.	.45	.59
df = 22	t = .30	Probability = .77

An F test of variances was first performed which yielded a two-tailed probability of .35, indicating that the pooled-variance estimate for the common variance should be used. A two-tailed t-test was then

performed (degrees of freedom = 22) which resulted in a probability of .77, far above the probability of .10 or less which was required to demonstrate significance. Thus, there was no significant difference between the two student groups when considering grade point averages. As a result of the t-test calculation and the other comparisons made, the student groups were found to be equivalent.

Intraclass Reliabilities

The amount of correlation between raters within each rating group is presented in Table 3.

Table 3

Reliability of Average Ratings of Holistic and Atomistic Methods and Each Category of Atomistic Method

	Experts	Students
Methods		
Holistic	.96	.96
Atomistic	.91	.89
Categories of the Atomistic Method		
Ideas	.88	.84
Organization	.84	.79
Wording	.79	.85
Flavor	.73	.54
Usage	.83	.67
Punctuation	.84	.82
Spelling	.95	.94

Although all of the groups were given explicit instructions to use the full range of the appropriate grading scale, within no group was this restriction strictly adhered to. Thus, the reliability figures given are slightly lower than those which would have resulted had the scores been spread across the scales to the full degree. The actual scores assigned to each paper by each rater are given in Appendix B.

Holistic Scoring.--The correlations of .96 for both student and expert groups using the Holistic method are extremely high, indicating very uniform agreement among the raters in these groups.

Atomistic Scoring.--Most categories of the Atomistic method as scored by the student group have reliabilities of .79 or higher. The two categories below this level are "Usage" with a reliability of .67 and "Flavor" with a reliability of .54. These relatively lower reliabilities seem in large part to be results of an imprecision in the definition of these terms by Diederich. Upon returning the scored papers, several students commented on one or both of these definitions as being "vague," "too broad," or even "insulting." Even with improved definitions, however, it would be expected that these categories would have lower correlations than a category such as "Spelling" in which a direct quantitative comparison of errors can be made.

The total score for a rater using the Atomistic method is simply the sum of the scores for all categories of that method. The reliability of these totals for the student group is .89. Such a high level

of correlation indicates that raters within this group tend to give total scores which are essentially the same. Also, this total reliability is higher than all of the categories of the atomistic method with the exception of the "Spelling" category.

Like the student group, the expert group using atomistic scoring achieved slightly lower reliabilities on total score and all categories than its holistic counterpart. All reliability coefficients are .73 or higher, with the "Wording" and "Flavor" categories producing the only scores below .83. The reliability of total scores is .91. Again, the only category with a higher reliability is the "Spelling" category.

Comparison of Students and Experts Using Atomistic Scoring.--A comparison of the reliabilities of the seven categories and total scores of the student group and those of the expert group shows a generally consistent pattern. The expert raters have higher reliability coefficients than the students in all cases except the "Wording" category. Also, the experts had much greater consistency when scoring the "Flavor" and "Usage" categories than did the students.

Correlations between Methods

The average scores for the holistic and atomistic methods are shown in Table 4.

Table 4
Average Scores for Methods Utilizing Raters

Essay	Holistic Groups		Atomistic Groups	
	Students	Experts	Students	Experts
1	1.36	1.30	37.70	35.30
2	4.79	4.80	17.60	15.30
3	2.29	2.30	30.10	30.60
4	3.93	4.30	23.40	23.00
5	3.43	2.80	23.70	35.60
6	2.79	2.40	30.50	29.00
7	4.00	3.40	22.00	23.00
8	2.64	3.10	27.20	24.90
9	3.29	3.00	25.90	22.50
10	1.71	1.80	35.70	35.30
11	2.29	2.10	33.10	29.20
12	1.79	1.40	35.40	35.40
13	2.71	2.20	26.40	30.00
14	3.21	3.20	25.20	23.80
15	3.64	4.00	20.70	16.80
16	3.14	2.90	23.90	27.40
17	3.64	3.40	24.90	23.00
18	3.64	3.10	25.20	26.90

It should be noted that for the atomistic rater groups, a higher score indicates a greater degree of quality as measured by the method. The holistic rater groups, however, utilized a scale in which the smaller number indicates the better score.

The raw scores from the objective methods appear in Table 5.

Table 5
Raw Scores for Methods Not Utilizing Raters

Essay ,	Mature Word Index	Type/Token Index	Mean T-unit Length	Syntactic Complexity
1	.113	6.83	21.1	12.3
2	.015	5.63	17.1	8.0
3	.045	5.38	16.0	8.4
4	.047	5.17	18.5	8.4
5	.000	4.56	10.7	5.4
6	.092	6.06	15.3	7.7
7	.045	4.46	14.9	7.3
8	.034	4.81	17.9	11.6
9	.018	4.72	11.0	6.3
10	.097	6.72	15.5	10.1
11	.049	5.35	13.9	6.2
12	.081	6.54	14.5	8.0
13	.036	6.06	19.1	12.2
14	.065	5.14	13.8	6.3
15	.033	4.63	17.3	11.0
16	.009	4.15	18.0	10.0
17	.058	4.95	20.6	15.7
18	.066	4.74	11.8	13.3

More detailed information concerning how words were categorized for the Mature Word Index may be found in Appendix C. To avoid the negative correlations which would result if the objective methods were directly correlated with the holistic rater groups, the scores of the holistic rater groups were subtracted from 6 before the rankings and correlations were determined. In effect, this translated the Holistic scores into a scale where the higher score indicated higher quality.

Table 6 is a listing of the rank ordering of the essays as obtained from each method.

Table 6
Rank Ordering of Methods and Rater Groups

Essay	Ex Hol	Ex Atm	St Hol	St Atm	MWI	TTI	T-U	Syn Com
1	1	2.5	1	1	1	1	1	3
2	18	18	18	18	16	6	8	11.5
3	6	4	4.5	6	10	7	9	9
4	17	14	16	15	9	9	4	10
5	8	10	12	14	18	16	18	18
6	7	7	8	5	3	5	11	13
7	14.5	14	17	17	11	17	12	14
8	11.5	11	6	7	13	12	6	5
9	10	16	11	9	15	14	17	15
10	3	2.5	2	2	2	2	10	7
11	4	6	4.5	4	8	8	14	17
12	2	1	3	3	4	3	13	11.5
13	5	5	7	8	12	4	3	4
14	13	12	10	10.5	6	10	15	16
15	16	17	14	16	14	15	7	6
16	9	8	9	13	17	18	5	8
17	14.5	14	14	12	7	11	2	1
18	11.5	9	14	10.5	5	13	16	2

The correlations between methods are presented in the following discussion and the accompanying tables. There are 16 degrees of freedom for all Pearson and Spearman correlations in the study, and the corresponding critical value for both Pearson and Spearman coefficients is .40.

Table 7 presents a matrix which shows the Pearson correlations of all methods and rater groups.

Table 7

Pearson Correlation Matrix of Methods and Rater Groups

	Ex Hol	Ex Atm	St Hol	St Atm	MWI	TTI	T-U	Syn Com
Ex Hol	---							
Ex Atm	.94*	---						
St Hol	.94*	.90*	---					
St Atm	.92*	.92*	.95*	---				
MWI	.57*	.64*	.62*	.75*	---			
TTI	.61*	.66*	.66*	.74*	.75*	---		
T-U	-.04	.04	.12	.06	.23	.28	---	
Syn Com	.03	.09	.08	.05	.27	.09	.64*	---

df = 16 Critical Value of $r = .40$

* significant at .05 level of confidence

All correlations followed by an asterisk are significant at the .05 level. In fact, all of these correlations are significant at levels of .006 or less, indicating that those methods and groups which correlate within the tested level have extremely high correlation coefficients. Those correlations which are not significant have probability values greater than .13. Thus, the correlations fall dramatically into two groups--very high correlations and very low correlations--with no borderline values.

In addition, the correlations between all rater groups are significant far beyond even the .0001 level. The relationships between rater groups are discussed in greater detail in a later section. Finally, Mean T-Unit Length and Syntactic Complexity scores correlate significantly only with each other.

The matrix of Spearman rank order correlations which appears in Table 8 presents similar results.

Table 8

Spearman Rank Order Correlation Matrix of Methods and Rater Groups

	Ex Hol	Ex Atm	St Hol	St Atm	MWI	TTI	T-U	Syn Com
Ex Hol	---							
Ex Atm	.93*	---						
St Hol	.91*	.88*	---					
St Atm	.89*	.85*	.93*	---				
MWI	.46*	.57*	.49*	.68*	---			
TTI	.58*	.61*	.60*	.69*	.67*	---		
T-U	-.07	.05	.09	.00	.08	.29	---	
Syn Com	.00	.14	.10	.12	.26	.19	.70*	---
df = 16	Critical Value of $\rho = .40$							
* significant at .05 level of confidence								

The correlations tend to be slightly less than the corresponding values for the Pearson correlations, but the same strong patterns hold.

These correlations by raw scores and by rank orderings indicate: (1) the average scores of all rater groups are highly correlated with each other; (2) the average scores of all rater groups are less highly

but still significantly correlated with the Mature Word Index and the Type/Token Index; and (3) the Mean T-Unit Length and Syntactic Complexity scores are significant only with respect to each other and do not even approach significance with any other method or rater group.

These results allowed the following decisions to be made regarding the acceptance or rejection of each aspect of hypothesis 1.

Null--No significant correlation exists between the following pairs of scoring methods:

- Holistic (Experts) - Mean T-unit Length
- Holistic (Experts) - Syntactic Complexity
- Holistic (Students) - Mean T-unit Length
- Holistic (Students) - Syntactic Complexity
- Atomistic (Experts) - Mean T-unit Length
- Atomistic (Experts) - Syntactic Complexity
- Atomistic (Students) - Mean T-unit Length
- Atomistic (Students) - Syntactic Complexity
- Mature Word Index - Mean T-unit Length
- Mature Word Index - Syntactic Complexity
- Type/Token Index - Mean T-unit Length
- Type/Token Index - Syntactic Complexity

Alternative--A significant positive correlation exists between the following pairs of scoring methods:

- Holistic (Experts) - Holistic (Students)
- Holistic (Experts) - Atomistic (Experts)
- Holistic (Experts) - Atomistic (Students)
- Holistic (Experts) - Mature Word Index
- Holistic (Experts) - Type/Token Index
- Holistic (Students) - Atomistic (Experts)
- Holistic (Students) - Atomistic (Students)
- Holistic (Students) - Mature Word Index
- Holistic (Students) - Type/Token Index
- Atomistic (Experts) - Atomistic (Students)
- Atomistic (Experts) - Mature Word Index
- Atomistic (Experts) - Type/Token Index
- Atomistic (Students) - Mature Word Index

Atomistic (Students) - Type/Token Index
 Mature Word Index - Type/Token Index
 Mean T-unit Length - Syntactic Complexity

Correlations of Atomistic Categories with Methods

The scores from each of the categories of the atomistic scoring procedure were correlated with the scores of each method. The Pearson and Spearman Correlation Coefficients using the student group are shown in Tables 9 and 10, respectively.

Table 9

Pearson Correlations between Methods and Categories of
 Atomistic Scoring from Students

	Ex Hol	Ex Atm	St Hol	St Atm	MWI	TTI	T-U	Syn Com
Ideas	.85	.86	.89	.93	.73	.86	.22 ⁺	.17 ⁺
Organization	.84	.84	.86	.90	.57	.70	.05 ⁺	-.05 ⁺
Wording	.81	.82	.87	.93	.80	.77	.12 ⁺	.12 ⁺
Flavor	.78	.79	.81	.90	.77	.79	.04 ⁺	.08 ⁺
Usage	.69	.71	.76	.80	.75	.66	-.03 ⁺	-.07 ⁺
Punctuation	.56	.56	.62	.65	.69	.34 ⁺	.02 ⁺	.00 ⁺
Spelling	.74	.72	.74	.75	.33 ⁺	.19 ⁺	-.17 ⁺	-.02 ⁺
df = 16	Critical Value of r = .40							

⁺ not significant at .05 level of confidence

Table 10

Spearman Rank Order Correlations between Methods and
Categories of Atomistic Scoring from Students

	Ex Hol	Ex Atm	St Hol	St Atm	MWI	TTI	T-U	Syn Com
Ideas	.81	.82	.81	.89	.69	.74	.19 ⁺	.31 ⁺
Organization	.90	.84	.88	.88	.44	.61	-.05 ⁺	-.06 ⁺
Wording	.72	.72	.81	.91	.77	.69	.02 ⁺	.10 ⁺
Flavor	.75	.76	.75	.90	.78	.75	.01 ⁺	.14 ⁺
Usage	.67	.67	.72	.80	.76	.70	-.10 ⁺	-.06 ⁺
Punctuation	.41	.44	.51	.56	.72	.38 ⁺	-.05 ⁺	-.12 ⁺
Spelling	.72	.66	.73	.73	.36 ⁺	.23 ⁺	-.17 ⁺	-.08 ⁺
df = 16		Critical Value of $\rho = .40$						

⁺ not significant at .05 level of confidence

Both types of correlation indicate the same patterns. Every category correlates significantly with all rater groups; the "Ideas," "Organization," "Wording," and "Flavor" categories show very high correlations with rater groups, while the other categories show slightly lower correlations.

None of the categories are significantly correlated with either the Mean T-Unit Length or the Syntactic Complexity, and the "Spelling" category is not significantly correlated with any objective method.

The Pearson and Spearman Correlation Coefficients using the expert group are shown in Tables 11 and 12, respectively.

Table 11
Pearson Correlations between Methods and Categories of
Atomistic Scoring from Experts

	Ex Hol	Ex Atm	St Hol	St Atm	MWI	TTI	T-U	Syn Com
Ideas	.91	.95	.85	.86	.63	.76	.04 ⁺	.09 ⁺
Organization	.83	.89	.78	.79	.53	.73	.07 ⁺	-.01 ⁺
Wording	.85	.90	.87	.88	.74	.77	.15 ⁺	.15 ⁺
Flavor	.90	.94	.85	.85	.64	.61	.13 ⁺	.20 ⁺
Usage	.81	.87	.82	.83	.60	.39 ⁺	-.01 ⁺	.13 ⁺
Punctuation	.47	.54	.49	.51	.46	.25 ⁺	.06 ⁺	.05 ⁺
Spelling	.69	.69	.65	.67	.30 ⁺	.14 ⁺	-.12 ⁺	.03 ⁺
df = 16		Critical Value of r = .40						

⁺ not significant at .05 level of confidence

Table 12

Spearman Rank Order Correlations between Methods and
Categories of Atomistic Scoring from Experts

	Ex Hol	Ex Atm	St Hol	St Atm	MWI	TTI	T-U	Syn Com
Ideas	.92	.97	.83	.83	.56	.63	-.03 ⁺	.11 ⁺
Organization	.89	.93	.83	.78	.47	.69	.06 ⁺	.06 ⁺
Wording	.84	.94	.81	.79	.67	.65	.07 ⁺	.21 ⁺
Flavor	.88	.95	.81	.83	.63	.65	.11 ⁺	.22 ⁺
Usage	.83	.89	.87	.91	.65	.57	.07 ⁺	.29 ⁺
Punctuation	.31 ⁺	.41	.36 ⁺	.42	.50	.37 ⁺	-.02 ⁺	-.12 ⁺
Spelling	.66	.63	.62	.62	.29 ⁺	.17 ⁺	-.08 ⁺	.00 ⁺
df = 16		Critical Value of p = .40						

⁺ not significant at .05 level of confidence

These figures are very similar to those of the student group. All categories except "Punctuation" and "Spelling" show extremely high correlations with rater groups but with no objective method, while the "Punctuation" category is marginally significant in some cases and not significant in others. Again, in no case is either Mean T-Unit Length or Syntactic Complexity significantly correlated with any category.

There are two very consistent patterns in all four of these tables. First, the atomistic categories--whether by students or experts--produce scores which correlate to a greater degree with the

rater group scores than with the objective methods. Of course, each category does contribute partially to the overall score for one rater group (in fact, in all cases this group has the highest correlation with each category), but the scores of the other three rater groups are totally independent. Second, Mean T-Unit Length and Syntactic Complexity scores have very insignificant--at times even slightly negative--correlation coefficients with all categories.

These results allowed the following decisions to be made regarding the acceptance or rejection of each aspect of Hypothesis 2.

Null--No significant correlation exists between the following pairs of atomistic categories and scoring methods:

<u>Category</u>	<u>Method</u>
Ideas - Mean T-unit Length	
Ideas - Syntactic Complexity	
Organization - Mean T-unit Length	
Organization - Syntactic Complexity	
Wording - Mean T-unit Length	
Wording - Syntactic Complexity	
Flavor - Mean T-unit Length	
Flavor - Syntactic Complexity	
Usage - Type/Token Index ¹	
usage - Mean T-unit Length	
Usage - Syntactic Complexity	
Punctuation - Holistic (Experts) ²	
Punctuation - Holistic (Students) ²	
Punctuation - Type/Token Index	
Punctuation - Mean T-unit Length	
Punctuation - Syntactic Complexity	
Spelling - Mature Word Index	
Spelling - Type/Token Index	
Spelling - Mean T-unit Length	
Spelling - Syntactic Complexity	

¹Only for Pearson correlations for experts.

²Only for Spearman correlations for experts.

Alternative--A significant positive correlation exists between the following pairs of atomistic categories and scoring methods:

<u>Category</u>	<u>Method</u>
Ideas - Holistic (Experts)	
Ideas - Holistic (Students)	
Ideas - Atomistic (Experts)	
Ideas - Atomistic (Students)	
Ideas - Mature Word Index	
Ideas - Type/Token Index	
Organization - Holistic (Experts)	
Organization - Holistic (Students)	
Organization - Atomistic (Experts)	
Organization - Atomistic (Students)	
Organization - Mature Word Index	
Organization - Type/Token Index	
Wording - Holistic (Experts)	
Wording - Holistic (Students)	
Wording - Atomistic (Experts)	
Wording - Atomistic (Students)	
Wording - Mature Word Index	
Wording - Type/Token Index	
Flavor - Holistic (Experts)	
Flavor - Holistic (Students)	
Flavor - Atomistic (Experts)	
Flavor - Atomistic (Students)	
Flavor - Mature Word Index	
Flavor - Type/Token Index	
Usage - Holistic (Experts)	
Usage - Holistic (Students)	
Usage - Atomistic (Experts)	
Usage - Atomistic (Students)	
Usage - Mature Word Index	
Usage - Type/Token Index ³	
Punctuation - Holistic (Experts) ⁴	
Punctuation - Holistic (Students) ⁴	
Punctuation - Atomistic (Experts)	
Punctuation - Atomistic (Students)	
Punctuation - Mature Word Index	
Spelling - Holistic (Experts)	

³Except for Pearson correlations for experts.

⁴Except for Spearman correlations for experts.

Spelling - Holistic (Students)
 Spelling - Atomistic (Experts)
 Spelling - Atomistic (Students)

Correlations Between Categories of the Atomistic Method

Although the correlations between the categories of the Atomistic method were not originally to be included in the study, these correlations do provide some interesting data. Table 13 shows the Pearson correlations for these categories as scored by the group of experts.

Table 13

Pearson Correlations between Categories of Atomistic
 Scoring for Experts

	Ideas	Organ	Wording	Flavor	Usage	Punct	Spelling
Ideas	---						
Organization	.95	---					
Wording	.94	.85	---				
Flavor	.89	.82	.85	---			
Usage	.71 ⁺	.60 ⁺	.73 ⁺	.84	---		
Punctuation	.33 ⁺	.21 ⁺	.38 ⁺	.58	.75	---	
Spelling	.51	.43	.46	.56	.76	.46	---
df = 16		Critical Value of r = .40					

⁺ not significant at .05 level of confidence

The scores for the category of "Punctuation" do not correlate significantly with three other categories. These are the only correlations which are not significant. In general, the four categories of "Ideas," "Organization," "Wording," and "Flavor" form a group within which the correlations are quite high.

The Pearson correlations for Atomistic categories as scored by the group of students are shown in Table 14.

Table 14
Pearson Correlations between Categories of Atomistic
Scoring for Students

	Ideas	Organ	Wording	Flavor	Usage	Punct	Spelling
Ideas	---						
Organization	.88	---					
Wording	.85	.80	---				
Flavor	.92	.82	.86	---			
Usage	.67	.59	.83	.68	---		
Punctuation	.46	.32 ⁺	.61	.48	.77	---	
Spelling	.51	.63	.62	.52	.48	.58	---
df = 16 Critical Value of r = .40							

⁺ not significant at .05 level of confidence

The same trends appear as in the expert group's scores, although there is only one correlation that is not significant ("Punctuation" with "Organization").

Pearson correlations between the Atomistic categories as scored by experts and the categories as scored by students are shown in Table 15. Very high correlations appear along the diagonal matching the same category for each group. For neither group do the scores for the "Punctuation" category correlate significantly with the scores for either the "Ideas" or the "Organization" categories. Again, the first four categories have generally higher correlations within their group than any other categories.

Table 15

Pearson Correlations between Categories of Atomistic
Scoring for Experts and Those for Students

Students	Experts						
	Ideas	Organ	Wording	Flavor	Usage	Punct	Spelling
Ideas	.90	.82	.88	.84	.67	.38 ⁺	.45
Organization	.85	.87	.78	.77	.64	.19 ⁺	.55
Wording	.79	.72	.83	.76	.75	.46	.52
Flavor	.80	.72	.80	.77	.68	.42	.44 ⁺
Usage	.65 ⁺	.56 ⁺	.79	.67	.74	.59	.37 ⁺
Punctuation	.38 ⁺	.25 ⁺	.53	.51	.74	.80	.53
Spelling	.54	.46	.52	.58	.80	.47	.96
df = 16		Critical Value of r = .40					

⁺ not significant at .05 level of confidence

Correlations of Methods with Sum of Rankings of all Other Methods

In order to investigate the relationship of each method to all other methods, the scores of each method were correlated with the sum of ranks of the other five methods. Again, there are 16 degrees of freedom and the critical values of the Pearson and Spearman coefficients are .40. The results of these correlations appear in Table 16.

Table 16

Correlations between Each Method and the Sum of Rankings of All Other Methods

Method	Pearson Correlation	Spearman Correlation
Using Expert Rater Groups		
Holistic	.57	.58
Atomistic	.69	.71
Mature Word Index	.65	.57
Type/Token Index	.72 ⁺	.70 ⁺
Mean T-unit Length	.31 ⁺	.31 ⁺
Syntactic Complexity	.29 ⁺	.39 ⁺
Using Student Rater Groups		
Holistic	.67	.63
Atomistic	.71	.70
Mature Word Index	.67	.59
Type/Token Index	.71 ⁺	.69 ⁺
Mean T-unit Length	.34 ⁺	.25 ⁺
Syntactic Complexity	.31 ⁺	.35 ⁺

df = 16

Critical Value of $r = .40$ Critical Value of $\rho = .40$

⁺ not significant at .05 level of confidence

Two sets of correlations appear in the table: in the first, the scores for the holistic and atomistic methods were taken from the expert groups; in the second, these scores were taken from the student groups. In both cases the Holistic, Atomistic, Mature Word Index, and Type/Token Index scores correlate significantly with their respective sums of ranks of all other methods. In fact, these correlations are all significant beyond the .01 level. On the other hand, Mean T-unit length and Syntactic Complexity scores do not show significant correlations.

These results allowed the following decisions to be made regarding the acceptance or rejection of each aspect of Hypothesis 3.

Null--No significant correlation exists between the following scoring methods and the sum of the rankings of the other five methods (regardless whether expert or student groups were used):

Mean T-unit Length
Syntactic Complexity

Alternative--A significant positive correlation exists between the following scoring methods and the sum of the rankings of the other five methods (regardless whether expert or student groups were used):

Holistic
Atomistic
Mature Word Index
Type/Token Index

Overall Correlations

The most general question investigated was whether or not a significant overall correlation exists between the six methods. Two correlations were calculated in order to answer this question. One used the expert groups as indicative of the Holistic and Atomistic methods, while the other used the student groups. The results are presented in Table 17.

Table 17

Kendall Coefficients of Concordance for All Methods

Using Expert Groups	Using Student Groups
Kendall W = .47	Kendall W = .50
df = 17	df = 17
$\chi^2 = 47.80$	$\chi^2 = 50.52$
p < .001	p < .001

There is a highly significant degree of agreement between the six methods. This is true no matter which raters (experts or students) are selected for the comparison.

This result allowed rejection of Hypothesis 4. That is, this study showed that a significant overall correlation exists between the six methods of essay scoring.

Because of the insignificant relationships found between the Mean T-Unit Length and the Syntactic Complexity scores with each of the other methods, the concordance levels were recomputed without using the scores from these two methods. The results are shown in Table 18. As was expected, these correlations are higher than those using all six methods. These differences between the correlations with all methods and those without the two methods mentioned, however, are slight.

Table 18

Kendall Coefficients of Concordance for Holistic, Atomistic,
Mature Word Index, and Type/Token Index Methods

Using Expert Groups	Using Student Groups
Kendall W = .72	Kendall W = .76
df = 17	df = 17
$\chi^2 = 49.21$	$\chi^2 = 51.53$
p < .001	p < .001

Analysis of Variance between Expert and Student Raters

A basic question to be answered was whether a difference exists in the ways experts and students score written composition. To provide support for answering this question, two analyses of variance were conducted: one for the Holistic method and one for the Atomistic method. The results of the analysis using the scores of the holistic rating groups are summarized in Table 19. The interaction effect is not significant, indicating that the differences among essays were not changing with the rater groups. The essay effect is highly significant. This was as expected: there is a significant difference in the rated quality of the papers.

Table 19

Analysis of Variance for Holistic Rating Groups by Essays

Source of Variation	SS	DF	MS	F	Significance
Rater groups	2.52	1	2.52	4.27	.039
Essays	333.02	17	19.59	33.24	.000
Interaction	11.09	17	.65	1.11	.345
Error	233.36	396	.59		

Grand Mean = 2.95

Expert Group Mean = 2.86

Student Group Mean = 3.02

The rater group effect is significant at the .05 level. This suggests that a difference exists between the experts and the students in the mean of scores assigned to the papers by each group. The student mean is higher than that of the experts, indicating that--for the Holistic method--students graded the papers more harshly than did the experts. The reader will recall that the scores of these two groups are highly correlated ($r = .94$). Thus, a given paper would be expected to receive more favorable scores from expert raters using Holistic scoring than from student raters using Holistic scoring. Differences in training sessions and perceptions between these two groups which may account for this variation are discussed in Chapter V.

This result required the rejection of Hypothesis 5. The study showed that a significant difference exists between ratings by pre-service English teachers and expert readers using the Holistic scoring method.

The results of the analysis of variance conducted using the scores of the atomistic rating groups are summarized in Table 20. Again, the interaction effect is not significant and the essay effect is highly significant. However, these groups did not demonstrate the difference in means found in the groups using Holistic scoring. Together with the high correlation between these groups discussed above ($r = .92$), this indicates that both experts and students assigned essentially the same scores to the same paper when using Atomistic scoring.

Table 20

Analysis of Variance for Atomistic Rating Groups by Essays

Source of Variation	SS	DF	MS	F	Significance
Rater groups	23.51	1	23.51	.64	.424
Essays	10456.63	17	615.10	16.78	.000
Interaction	576.69	17	33.92	.93	.544
Error	11873.80	324	36.65		
Grand Mean = 26.82					
Expert Group Mean = 26.49					
Student Group Mean = 27.14					

This result allowed Hypothesis 6 to be accepted. The study showed that no significant difference exists between ratings by pre-service English teachers and expert readers using the Atomistic scoring method.

Summary

The results of the statistical analyses required for the study were presented in this chapter. The principle findings were:

- (1) Based on previous training, course of study, membership in a specified English course, and college grade-point average, the two student groups were found to be comparable in make-up. Grade-point averages of the groups were analyzed with a t-test, and the results showed no significant difference in group means.
- (2) Intraclass correlations were computed for each rater group, and for each category of the Atomistic method. The resulting reliability figures are very high for the methods (.89 or higher). The figures for the categories show considerably more variability (.54 to .95) with "Spelling" having the highest reliability for both groups, and "Flavor" having the lowest for both groups.
- (3) Correlations between pairs of methods show significant relationships between all rater groups (i.e., the Holistic and Atomistic methods). The Mature Word Index and Type/Token Index are significantly correlated with all rater groups, while the Mean T-Unit Length and Syntactic Complexity scores are correlated only with each other.

- (4) The scores of each of the rater groups correlate very significantly with the combined scores of all other methods and groups.
- (5) The categories of the Atomistic method generally correlate significantly with all rater groups. No category for either students or experts correlates significantly with the Mean T-Unit Length or Syntactic Complexity.
- (6) There is a very significant overall correlation between the six methods when either the student groups or the expert groups are used to provide the scores for the Holistic and Atomistic methods.
- (7) Analyses of variance for Holistic groups by essays and for Atomistic groups by essays showed the interaction between essays and rating groups to be not significant. The essay effect in both cases was highly significant. For the Holistic rating groups, group membership did have a significant ($p = .039$) bearing on score, while for the Atomistic rating groups, group membership was not significant.

The statistical hypotheses presented in Chapter III are accepted or rejected based on the results.

CHAPTER V

DISCUSSION

In this final chapter, the problems of the study are re-examined in light of the knowledge obtained from the investigations reported in previous chapters. The chapter begins with a brief summary of the study in order to reacquaint the reader with the principle problems and procedures contained herein. Then, the conclusions of the study are stated, followed by a general interpretive discussion of all major findings of the study. These conclusions were made in light of the results of the various statistical analyses as they applied to the hypotheses stated in Chapter III. Next, recommendations for applications of the findings are made. Finally, a section of suggestions for further research concludes the chapter.

Summary of the Study

This study was undertaken to compare various methods of evaluating student writing as well as to determine if experienced teachers and students differed in their judgments of writing quality. The methods studied were: holistic scoring, atomistic scoring, and measures of mature word choice, syntactic complexity, mean T-unit length, and vocabulary diversity.

The Holistic scoring procedure required raters to score each essay using a five-point scale after a single, rapid reading. In

contrast, the Atomistic procedure required raters to read each essay more carefully in order to judge the quality of the paper based on seven distinct categories. Each category was scored on a five-point scale and the scores summed to give a total score for each paper. The Mature Word Index is a measure of the frequency with which mature words appear in an essay. For the Syntactic Complexity score, various syntactic structures were assigned values from 0 to 3; the values of all such structures appearing in an essay were added together and then divided by the number of sentences in the essay. The result is a measure of the complexity of syntactic structure employed by the writer. Mean T-unit Length is a measure of the length of a particular syntactic structure closely related to an independent clause, and the Type/Token Index is a measure of vocabulary diversity.

In order to compare the methods, a set of papers was obtained from high school juniors and seniors and graded using each of the six methods. The Holistic and Atomistic methods required subjective evaluations and produced different scores for different raters, while the other four methods are objective in nature and required only a careful adherence to a set procedure. Furthermore, since one of the purposes of the study was to compare the ratings of experienced teachers with those of pre-service teachers, groups of experts and groups of college English majors and minors were recruited to score the

papers either holistically or atomistically. The scores for these methods were the average scores assigned by each group of raters.

Each of the methods was correlated with every other method and each category of the atomistic method was correlated with every method. Both of these correlations were carried out with raw scores as well as with the rank orders of the essays as determined from the various methods. Also, each method was correlated with the sum of the rankings of all other methods. A correlation of all methods was then performed to determine the degree of overall agreement among the methods. Finally, expert raters using the Holistic method were compared with student raters using that method, and a similar comparison was made for the Atomistic method.

Conclusions

The following conclusions were drawn from the results of the statistical tests performed in the study.

(1) The Atomistic scoring method is more time-consuming and no more reliable or informative than Holistic scoring.

(2) Many of the factors generated by Diederich to score writing do not provide reliable results between different raters.

(3) The Mature Word Index is an appropriate measure of writing quality.

(4) The Type/Token Index is an appropriate measure of writing quality.

(5) The Mean T-unit Length is not an appropriate measure of writing quality.

(6) The Syntactic Complexity Index is not an appropriate measure of writing quality.

(7) Writers do not misuse or misplace mature words as they often do syntactic structures.

(8) Student raters judge writing as a whole in essentially the same manner as do expert raters.

(9) Student raters are slightly less able to distinguish the various factors of quality writing than are experts.

The following sections of this chapter explain and expand upon these conclusions.

Holistic Versus Atomistic Scoring

One of the more interesting results of the study was the level of agreement among the various rater groups--very high reliability scores were found for both the Holistic and Atomistic methods and for both expert and student groups. The groups using Holistic scoring, however, had somewhat higher reliability scores than those using Atomistic scoring. This result would seem to make the selection of the rating method for use by teachers an easy task: the Holistic method is both faster and more reliable and hence ought to be chosen. It is difficult to argue with these facts. Still, it must be remembered that the Holistic method provides little more than a score on each

paper. When the chief goal is to grade large quantities of papers accurately and rapidly, this is of little consequence. But when students are to receive feedback on their writing, the method's weakness becomes apparent--no specific information is available to the student about his writing.

The Atomistic method, on the other hand, did provide a degree of more specific information through the scores on the various categories. However, the reliability scores of some of these categories--especially among the student group--were considerably lower than the reliability of the total Atomistic scores. Thus, even if writers in general were to read the relatively lengthy definitions of the various categories, they would have a high probability of using a different definition of a category such as "Flavor" than did the grader of the writer's paper. Even for a relatively mechanical category such as "Punctuation," the writer would only have a score, with no hint as to what punctuation might be improper and what alternatives might exist. Thus, the supposed advantage of the Atomistic scale disappears; it really provides little, if any, additional corrective information to the writer.

The one remaining possible defense of the Atomistic scale is that it may be more valid than the Holistic measure. The high correlations between the two methods, however, tend to belie this claim, demonstrating that the methods produce essentially the same scores. It

would appear, then, that in situations where themes are used as aids to placement into the appropriate level of multi-level English curriculum or for other gross measurement purposes, the Holistic method is to be preferred. In classroom situations where the improvement of writing skills is the goal, neither the Holistic nor the Atomistic method is appropriate.

The Diederich (1966) scale and similar Atomistic scales do not appear to be directly useful to the classroom teacher. There is certainly much merit in Diederich's approach to separating distinct writing qualities. However, much work remains to be done before his scale will provide results which are meaningful to students.

Within the Atomistic method, there was a substantial variation across categories in the degree to which raters agreed in their ratings. As would be expected, the "Spelling" category recorded highly reliable results. This is a category which allows an almost mechanical comparison of the number and types of spelling errors. Most of the variation which did occur could probably have been eliminated if the raters would have been required to keep tallies of incorrect spellings on each paper.

In contrast, the category of "Flavor" generated much less reliable scores. As was mentioned earlier, there was a general uneasiness among raters concerning this category. In the work explaining the scale which was adapted for use in this study, Diederich (1966)

defined the high, middle, and low points of each of the categories. These definitions--particularly in the case of "Flavor"--are less than adequate. While Diederich was correct to identify major factors of writing, the extension of these factors to a rating scale does not appear to have provided the solution Diederich hoped for.

Correlations Between Methods

The matrices showing correlations between pairs of methods (Tables 7 and 8, Chapter IV) are very revealing. Mean T-unit Length and Syntactic Complexity correlate significantly only with each other. This was very surprising since both of these scores would seem, on an intuitive level, to provide a much more precise measure of writing quality than either the Mature Word Index or the Type/Token Index. Furthermore, since its development by Hunt in 1965, the T-unit has been a standard measure employed almost unquestioningly by researchers. It now appears that the average length of the T-units in a writing sample has no bearing on the perceived quality of that sample. Similarly, the Syntactic Complexity Score obtained from Botel and Granowsky's (1972) formula is irrelevant to the perceived quality of writing.

Explanations of these facts are difficult in light of the past use of T-unit and syntactic complexity measures (see Cooper, 1975), but it may be that sheer weight of syntactic structure is not of any

importance; for, after all, many complex structures are simply incorrect or even incomprehensible. If this is so, Christensen's (1968) criticisms were accurate--we ought to be focusing more on the positioning and appropriateness of the features of the syntactic landscape rather than merely looking to see what is there.

It should be noted that Mean T-unit Length and Syntactic Complexity scores correlate to a highly significant degree. Both methods are apparently measuring the same thing: complexity of syntactic structure. Nonetheless, this is not--according to all rater groups--an important factor to be measuring.

The scores from the Mature Word Index are highly significantly correlated with the scores from each of the rater groups. Clearly, the use of mature words is an important factor in the judgment of writing quality. It is natural at this point to question why the Mature Word Index does not suffer from the same weaknesses as the T-unit and syntactic complexity measures; that is, do not writers often load a paper with mature words which are misplaced or misused? The answer seems to be "no." The writers used in this study rarely misused words in the same way that they misused syntactic structures. Furthermore, this is probably not an isolated occurrence. Writers are generally conscious of the meaning of a word they wish to employ; if they are not reasonably sure of its semantic value, they will choose another word. The analogous process apparently does not occur

when syntax is involved. That is, writers will frequently use improper syntactic structures without knowing they are doing so. So, a quantitative measure of mature words is adequate, while for syntactic structures a qualitative measure is required.

The scores from the final objective measure, the Type/Token Index also correlated to a highly significant degree with all rater groups. The level of vocabulary diversity is important in the perception of writing quality. Much of the above discussion concerning mature words is relevant here, as well. That is, vocabulary diversity generally results from competence--if a writer commands a substantial vocabulary, he will tend to use a wider range of words in his writing than another writer who possesses a smaller store of words. Also, it is very difficult (if not impossible) to use words one does not know. The use of a relatively large set of words was clearly valued by all of the rater groups.

Even higher correlations were obtained between the scores of the various rater groups, indicating substantial agreement among the groups as to what constituted good writing. Somewhat surprising was the degree to which student raters agreed with their expert counterparts. Coupled with the high reliability of student scores, this strongly suggests that student raters--at least by the time they have reached the junior year as English majors or minors--possess essentially the same skills in judging writing as do those with considerable

teaching experience. Perhaps, in the area of evaluation at least, teaching experience is not of the importance we have been led to believe.

The most likely explanation of these high student-expert correlations is that students are generally quite literate and have substantial backgrounds in reading and writing about quality literature. Thus, they have standards of excellence, so to speak, which they have come to recognize. While they have not read as many student papers as the experts, they nevertheless do have substantial criteria against which to measure writing quality. The results of the study show that expert raters have gained little if any additional competence in judging writing quality since they were upperclassmen in college.

Correlations Between Atomistic Categories and Methods⁵

In the last section, the lack of correlation between the scores from the Mean T-unit Length and Syntactic Complexity methods with all individual rater groups was examined. When the scores from these methods and scores from each of the categories of the Atomistic method are compared, an even more striking lack of correlation is evident:

⁵Pearson and Spearman correlations using the categories of the Atomistic method provided very nearly the same results. Discrepancies are minor and do not influence the major findings. Thus the following discussion assumes the Pearson correlations. Also, in most cases the expert and student groups closely agreed concerning correlation values. Those instances which are exceptions are noted in the discussion.

neither of these two methods even approaches significance with any of the categories of the Atomistic scale. This is true even for the "Usage" category which is largely concerned with the use of proper syntactic structures. Thus, neither Mean T-unit Length nor Syntactic Complexity measured the same quality as any of the factors (disregarding handwriting) which Diederich (1966) identified as important aspects of writing quality. This further supports the idea that neither method effectively measures any significant portion of writing quality; since Diederich claimed to have found the most important factors of writing maturity and none of them correlated with the methods at hand, those methods do not measure any major aspects of writing.

Several other patterns emerge when considering the scores in the Atomistic categories. As would be expected, the Mature Word Index showed its highest correlations with the "Wording" category, indicating that this category does allow raters to discriminate writing which contains uncommon words. The Mature Word Index showed an insignificant correlation only with the "Spelling" category. Thus, raters did not find spelling to be related to mature word usage. With one exception (discussed below), this method provided consistently high significant correlations with all other categories.

The Type/Token Index was not significantly correlated with either "Punctuation" or "Spelling" scores. Clearly, vocabulary diversity has little or no relation to these very mechanical aspects of writing and

the scores of both groups of raters reflected this fact. On the other hand, this method correlated to a highly significant degree with the first four categories. Interestingly, however, for the "Usage" factor there was considerable disagreement between the expert raters and the student raters. Apparently, the students found vocabulary diversity to be an important aspect of the "Usage" factor, while there was no significant correlation for the experts. This is one case where the experts seem to have performed better than the students; for in Diederich's definition of the "Usage" category, vocabulary diversity is not mentioned. The students were probably somewhat uncomfortable with the term "Usage" and ascribed qualities to it that were not intended by Diederich. The other explanation--that students with good control over usage also have a more diverse vocabulary--seems less plausible. Both the Mature Word Index and the Type/Token Index as predictors of perceived writing quality are discussed in the section "Recommendations."

When examining Tables 9 and 11 from Chapter IV, it will be noticed that the range of correlations from the expert group is broader than that from the students (discounting Mature Word Index and Syntactic Complexity). In particular, the "Punctuation" and "Spelling" correlations for the experts are generally lower than for the students. This indicates that the experts were better able to eliminate the more mechanical aspects of a piece of writing from the total impression of

the writing. While spelling and punctuation did contribute to the overall score of a writing sample when judged by the experts, these factors were more strongly related to total scores when students scored the writing.

There are several possible reasons for this situation. Perhaps the student raters were less confident about their scoring practices and, having scored the first five factors, hesitated to diverge from these scores for the final two factors. That is, having committed themselves by marking a paper in a general area (high, middle, low), they were influenced by these marks and tended to conform to the same area for the remaining factors. It is also possible that the lines between all of the categories were slightly blurred for the students. Certainly the "Usage"-Type/Token Index correlation discussed above lends support to this interpretation. Finally, the students may believe that mechanics are, after all, important enough to contribute more strongly to the final score. As in all such situations, the true reason is probably a combination of all of the above explanations.

Further light is shed on this topic by an examination of the correlations between categories. For the experts, the first five factors correlated to a higher degree among themselves than they did for the students. The reverse is true for the "Punctuation" and "Spelling" factors. All of this supports the idea that the students view all of the factors as slightly more homogeneous than do the

experts, the students being less able to differentiate various factors which influence their perception of the quality of the writing. Still, as it has been shown, the correlations of total scores for students with those for experts were exceptionally high.

The salient point is this: students are equally as capable as experts of providing a single indicator of quality for each piece of writing in a set; they are somewhat less capable of determining why that piece of writing is given a particular score. Experts were better able to deal with the various categories of writing (as evidenced by their higher reliability scores on all factors but one), and to differentiate between the mechanical and the more creative aspects of writing.

Correlations of Methods with Sum of Rankings of All Other Methods

More evidence for the unsuitability of the Mean T-unit Length and Syntactic Complexity methods as measures of writing quality was obtained from the correlations of each of these methods with the sum of the rankings of all other methods. These correlations were not significant, indicating that neither method is capable of providing to a significant degree the same scores as the combination of several varied methods. On the other hand, each of the other four methods did show highly significant correlations with all other methods, even including Mean T-unit Length and Syntactic Complexity. Presumably,

these correlations would have been higher still had the Mean T-unit Length and Syntactic Complexity scores been removed from the total.

Overall Correlations

The Kendall Coefficients of Concordance showed highly significant correlations between all methods no matter if the expert groups or the student groups were used to represent the Holistic and Atomistic methods. Since these correlations include all six methods, even the unpromising Mean T-unit Length and Syntactic Complexity measures, the strength of the relationship between the other four methods is obvious. Thus, the Holistic, Atomistic, Mature Word Index and Type/Token Index are all highly inter-related methods of rating student writing.

Comparison of Expert and Student Raters

One of the more important results of the study concerns the similarities and differences between the expert and student rater groups. The results presented thus far for the groups using the Holistic method indicated that both groups maintained the same relative ordering and intervals among the essays. That is, if a paper was the fourth best as scored by the experts, it would tend to be in the same position, at the same relative distance from the best paper when scored by the students. What has not yet been discussed is whether, on the absolute scale of 1 to 5, both groups tended to assign the same

score to the paper. Results of the analysis of variance suggest that they did not. The student group was less forgiving in its grading than the experts. This result is somewhat surprising as it was anticipated that if the students differed at all from the experts, they would have been more lenient. Some reasons for this fact may be suggested. One possible reason may be that the experts were used to dealing with less than perfect papers and were thus somewhat less severe in grading, while students, not used to reading high school writing were less willing to appreciate its relative merits. Another possible explanation concerns the training sessions which were conducted for each group. Each group was encouraged to develop a group standard of quality in disregard of any other group or personal norm. The discussions which ensued after each training paper was graded were designed to gradually lead to acceptance of a standard of grading determined by and unique to each particular group of raters--to alter the makeup of the group would be to alter the concomitant grading standard. Thus, it is easy to see how this difference could result. (In fact, it is surprising that the correlation between these groups was so high given this procedure.)

Probably some combination of these two factors was at work in this case. It should also be noted that the significance of the rater group's F ratio was .039, only slightly below the chosen confidence level of .05. Thus, the difference here does not seem to be severe

enough to suggest any major effort to train students into a more lenient mode of grading.

When the analysis of variance was applied to the total scores obtained from the Atomistic method, the means of the expert and student groups were found not to be significantly different. Because the total scores of these groups correlated so highly and the analysis of variance showed no significant difference between scores of the groups, the Atomistic method produced scores which tended to be the same for each group on an absolute scale.

Finally, considering all results which bear on the comparability of scores from the expert groups with those from the student groups, it is clear that a remarkable degree of accord exists.

Recommendations

The major results of the study have several implications for research and teacher training.

(1) Because Atomistic scoring is more time-consuming and no more reliable or informative than Holistic scoring, Holistic scoring should be the method of choice for research, placement, and other evaluative tasks not requiring feedback to students.

(2) Any use of Mean T-unit Length in research should be suspect until thoroughly justified. Because it is not an accurate measure of any major factor of perceived writing quality, it ought not to be taken as an independent indicator of writing quality. Studies which

show an increase in mean T-unit length with chronological age should be re-examined to determine more qualitative bases which may underlie the increase in T-unit length.

(3) A similar skepticism should surround the use of measures of syntactic complexity. Again, studies using such measures should be re-examined to find recognizable qualitative syntactic differences between different levels of writing.

(4) College methods courses should assist prospective teachers in defining the various factors which are important parts of the writing task. If these future teachers are to be able to help their students, they must learn to identify those areas within a piece of writing which need improvement. Thus, English methods courses should include activities to train prospective teachers to recognize specific factors and how those factors contribute to the whole impression generated by a piece of writing.

(5) Students in the method courses used in the study were surprisingly lacking in confidence concerning their abilities to properly evaluate written work. This study should provide a great deal of reassurance and stimulate their confidence because of the high correlation between scores by students and scores by experts. These results should be discussed with the students and the implications brought firmly to their attention. A great deal of anxiety may thereby

be relieved and leave such these students to concentrate on more difficult aspects of English education.

(6) Professors of English education methods courses should not be overly concerned with developing consistent grading patterns among their students. These patterns are already well in place.

Suggestions for Future Research

This study was an initial step in the area of comparative evaluation of written composition. As such it has exposed many more problems than it has solved. Some of the more interesting of these problems are presented in this section as possible topics of future research.

(1) There is a great need to redefine the factors discovered by Diederich. A replication of his study reported in 1966 with a special commitment to providing lucid definitions for the factors found to be important is in order.

(2) A reliable qualitative measure of syntactic complexity is desperately needed. It is not enough to catalogue syntactic structures--a method needs to be found which will consider effectiveness of the structures as primary.

(3) This study used relatively sophisticated people as raters. It would be of interest to extend the sample of raters to include groups of high school students, teachers in other disciplines, non-teacher adults, professional writers, etc. In this way a more

complete picture of how writing is perceived by different groups could be obtained.

(4) The study could also be extended to include writing samples from different groups throughout the range of beginning writers to professionals.

(5) This study showed that in most important respects, teaching experience was not a factor in how raters scored writing samples. It would be of great benefit to discover what other aspects of a teacher's job similarly are not enhanced by experience. This information could greatly assist college professors of English methods courses in the planning of their instruction. Much of this type of information could probably be transferred to other content fields within education, as well.

(6) Eldridge (1981) found that college instructors of English composition tended to stress mechanics and organization to a greater extent during the seventies than in the sixties. The present study provided a base of data which may make changes such as these more readily apparent and more easily quantified if the study were replicated at intervals.

(7) This study eliminated the contaminating effects of handwriting. It would be informative to include this factor in a replication of the study. It may be that the effect of handwriting is so powerful that other factors lose their importance.

REFERENCES CITED

- Applebee, Arthur N. "Looking at Writing." Educational Leadership, 38 (1980-81), 458-462.
- Belanger, J. F. "Calculating the Syntactic Density Score: A Mathematical Problem." Research in the Teaching of English, 12 (1978), 149-153.
- Bishop, Arthur, ed. Focus 5: The Concern for Writing. Princeton, N.J.: Educational Testing Services, 1978.
- Botel, Morton and Alvin Granowsky. "A Formula for Measuring Syntactic Complexity: A Directional Effort." Elementary English, 49 (1972), 513-516.
- Carlson, R. K. Sparkling Words: Two Hundred Practical and Creative Writing Ideas, rev. ed. Berkeley, California: Wagner Printing Company, 1973.
- Carroll, John B. Language and Thought. Englewood Cliffs, New Jersey: Prentice-Hall, 1964.
- Carroll, John M., Peter Davies, and Barry Richman. Word Frequency Book. Boston: Houghton Mifflin, 1971.
- Chaucer's Poetry: An Anthology For The Modern Reader, ed. E. T. Donaldson, 2d ed. New York: John Wiley & Sons, 1975.
- Christensen, Francis. "The Problem of Defining a Mature Style." English Journal, 57 (1968), 572-579.
- Coffman, William E. "On the Reliability of Ratings of Essay Examinations in English." Research in the Teaching of English, 5 (1971), 24-36.
- Cooper, Charles R. "Measuring Growth in Writing." English Journal, 64 (1975), 111-120.
- _____. "Holistic Evaluation of Writing." Evaluating Writing, eds. Charles R. Cooper and Lee Odell. Urbana, Ill.: National Council for Teachers of English, 1977.
- Cooper, Charles R. and Lee Odell. "Introduction." Evaluating Writing, eds. Charles R. Cooper and Lee Odell. Urbana, Ill.: National Council of Teachers of English, 1977.

- Diederich, Paul B. "How to Measure Growth in Writing Ability." English Journal, 55 (1966), 435-449.
- _____. Essentials of Educational Measurement, 2d ed. Englewood Cliffs, New Jersey: Prentice-Hall, 1972.
- _____. Measuring Growth in English, Urbana, Ill.: National Council of Teachers of English, 1974.
- Dixon, Edward A. "Syntactic Indexes and Student Writing Performance: A Paper Presented at NCTE-Las Vegas, 1971." Elementary English, 49 (1972), 714-716.
- Ebel, Robert L. "Estimation of the Reliability of Ratings." Psychometrika, 16 (1951), 407-424.
- _____. Essentials of Educational Measurement, 2d ed. Englewood Cliffs, N.J.: Prentice-Hall, 1972.
- _____. "Measurement and the Teacher." Educational and Psychological Measurement, eds. David A. Payne and Robert F. McMorris, 2d ed. Morristown, N.J.: General Learning Press, 1975.
- Eldridge, Richard. "Grading in the 70s: How We Changed." College English, 43 (1981), 64-68.
- Endicott, Anthony L. "A Proposed Scale for Syntactic Complexity." Research in the Teaching of English, 7 (1973), 5-12.
- Ferguson, George A. Statistical Analysis in Psychology & Education, 4th ed. New York: McGraw Hill Book Company, 1976.
- Finn, Patrick J. "Computer-Aided Description of Mature Word Choices in Writing." Evaluating Writing, eds. Charles R. Cooper and Lee Odell. Urbana, Ill.: National Council of Teachers of English, 1977.
- Follman, John C. and James A. Anderson. "An Investigation of the Reliability of Five Procedures for Grading English Themes." Research in the Teaching of English, 1 (1967), 190-200.
- Fowles, Mary E. Basic Skills Assessment: Manual for Scoring the Writing Sample. Princeton, N.J.: Educational Testing Services, 1978.

- Fox, Sharon E. "Syntactic Maturity and Vocabulary Diversity in the Oral Language of Kindergarten and Primary School Children." Elementary English, 49 (1972), 489-496.
- Gebhard, Ann O. "Writing Quality and Syntax: A Transformational Analysis of Three Prose Samples." Research in the Teaching of English, 12 (1978).
- Godshalk, Swineford, and Coffman. The Measurement of Writing Ability. Princeton, N.J.: College Entrance Examination Board, 1966.
- Golub, Lester S. Syntactic Density Score (SDS) with Some Aids for Tabulating. ERIC Document ED 091 741, 1973.
- Golub, Lester S. and Carole Kidder. "Syntactic Density and the Computer." Elementary English, 51 (1974), 1128-1131.
- Green, John A. Teacher-Made Tests. New York: Harper & Row, 1963.
- Grose, Lois M., Dorothy Miller, and Erwin R. Steinberg, eds. Suggestions for Evaluating Junior High School Writing. Urbana, Ill.: National Council of Teachers of English, 1963.
- Herdan, G. Quantitative Linguistics. Washington, D.C.: Butterworth Inc., 1964.
- Hillard, Helen, ed. chairman. Suggestions for Evaluating Senior High School Writing. Urbana, Ill.: National Council of Teachers of English, 1963.
- House, Ernest R., Wendell Rivers, and Daniel L. Stufflebeam. "An Assessment of the Michigan Accountability System." Phi Delta Kappan, 55 (1973-74), 663-669.
- Hunt, Kellogg W. Grammatical Structures Written at Three Grade Levels. NCTE Research Report, no. 3. Urbana, Ill.: National Council of Teachers of English, ERIC Document ED 113 735, 1965.
- _____. "Early Blooming and Late Blooming Syntactic Structures." Evaluating Writing, eds. Charles R. Cooper and Lee Odell. Urbana, Ill.: National Council of Teachers of English, 1977.
- Hunting, Robert and others. Standards for Written English in Grade 12. Crawfordsville, Indiana: Indiana Printing Company, 1960.

- Judine, Sister M., ed. A Guide for Evaluating Student Composition. Urbana, Ill.: National Council of Teachers of English, 1965.
- Kuder, G. F. and M. W. Richardson. "The Theory of the Estimation of Test Reliability." Psychometrika, 2 (1937), 151-160.
- Lindquist, E. F. Design and Analysis of Experiments in Psychology and Education. Boston: Houghton Mifflin Company, 1953.
- Lloyd-Jones, Richard. "Primary Trait Scoring." Evaluating Writing, eds. Charles R. Cooper and Lee Odell. Urbana, Ill.: National Council of Teachers of English, 1977.
- Lorge, Irving. "Word Lists as Background for Communication." Teachers College Record, 45 (1944), 543-552.
- Lundsteen, Sara W., ed. Help for the Teacher of Written Composition (K-9). Urbana, Ill.: ERIC Clearinghouse on Reading and Communication Skills, 1976.
- Maybury, B. Creative Writing for Juniors. London: B. T. Batsford, 1967.
- McNemar, Quinn. Psychological Statistics, 4th ed. New York: John Wiley and Sons, 1969.
- Moffett, James and Wagner, Betty Jane. Student-Centered Language Arts and Reading, K-12: A Handbook for Teachers, 2d ed. Boston: Houghton Mifflin Company, 1976.
- Moslemi, Marlene H. "The Grading of Creative Writing Essays." Research in the Teaching of English, 9 (1975), 154-161.
- Morris, William, ed. The American Heritage Dictionary of the English Language. Boston: Houghton Mifflin Company, 1978.
- Nail, Pat and others. A Scale for Evaluation of High School Student Essays. Urbana, Ill.: National Council of Teachers of English, 1960.
- Nemanich, Donald. "Passive Verbs in Children's Writing." Elementary English, 49 (1972), 1064-1066.
- Nie, Norman H. and others. Statistical Package for the Social Sciences, 2d ed. New York: McGraw-Hill Book Company, 1975.

- O'Donnell, Roy C. "A Critique of Some Indices of Syntactic Maturity." Research in the Teaching of English, 10 (1976), 31-38.
- Page, Ellis B. "The Imminence of Grading Essays by Computer." Phi Delta Kappan, 47 (1966), 238-243.
- Slotnick, Henry B. and Knapp, John V. "Essay Grading by Computer: A Laboratory Phenomenon?" English Journal, 60 (1971), 75-87.
- Thorndike, Edward L., and Lorge, Irving. The Teacher's Word Book of 30,000 Words. New York: Bureau of Publications, Teachers College, Columbia University, 1944.
- Thorndike, Robert L. "Reliability." Perspectives in Education and Psychological Measurement, eds. Glenn H. Bracht, Kenneth O. Hopkins, and Julian C. Stanley. Englewood Cliffs, New Jersey: Prentice-Hall, 1972.
- Tuckman, Bruce W. Conducting Educational Research. New York: Harcourt Brace Jovanovick, 1972.
- Veal, L. Romon, and Murray Tillman. "Mode of Discourse Variation in the Evaluation of Children's Writing." Research in the Teaching of English, 5 (1971), 37-45.

APPENDIXES

APPENDIX A
COMPUTER PROGRAMS

***** SPITSOL WORD COUNT PROGRAM - A *****

*
*
*
*
*
*

BARRY DONAHUE
MONTANA STATE UNIVERSITY
JUNE 15, 1981

*** THIS PROGRAM OUTPUTS AN ALPHABETIZED LIST OF WORDS AND
* AND THE NUMBER OF TIMES EACH APPEARS, AS WELL AS
* COUNT OF TYPES AND TOKENS
*

INPUT(.INPUT,105)
OUTPUT(.OUTPUT,108)
&ANCHOR = 1
SPACER = ' _-/()&"',;:?!'
HYPHEN = '-'
BLANKS = SPAN(SPACER)
PAT = BREAK(SPACER) . WORD SPAN(SPACER) . CH
NUMBER = ANY('123456789')

*

* Function which extracts words from a file and counts
* the total number of words and the number of distinct
* words.

```

      DEFINE('READ()')      : (PR)
READ      TOKEN = 0
          NUMWORD = TABLE(100,10)
NEXTL     TEXT = INPUT ' ' : F(BACK)
          TEXT NUMBER : S(NEXTL)
          TEXT BLANKS =
GOOD      TEXT PAT = : F(NEXTL)
HY        IDENT(CH,HYPHEN) : F(KEEP)
          SAVE = WORD
          TEXT PAT =
          WORD = SAVE '- ' WORD : (HY)
KEEP      NUMWORD<WORD> = NUMWORD<WORD> + 1
          TOKEN = TOKEN + 1 : (GOOD)
BACK      READ = NUMWORD : (RETURN)

```

*

* Function which outputs list and word counts

```

PR      DEFINE('PRINT(OUT) I') : (GO)
PRINT   OUTPUT = 'WORD          OCCURRENCES'
          OUTPUT = '-----'
          OUTPUT =
NO       I = I + 1
          OUTPUT = OUT<I,1> DUPL(' ',18 - SIZE(OUT<I,1>))
          -SIZE(OUT<I,2>)) OUT<I,2> : S(NO)

```

121

```
OUTPUT = '-----'  
OUTPUT =  
OUTPUT = 'NUMBER OF TYPES = ' I - 1  
OUTPUT = 'NUMBER OF TOKENS = ' TOKEN      : (RETURN)
```

```
*-----  
***** Main body of program  
*
```

```
GO      WDLIST = READ()  
        ALPHA = SORT(WDLIST)  
        PRINT(ALPHA)  
END
```

***** SPITBOL WORD COUNT PROGRAM - B *****

*

*

*

*

*

*** THIS PROGRAM OUTPUTS AN ALPHABETIZED LIST OF WORDS
* AND THE NUMBER OF TIMES EACH APPEARS, AS WELL AS
* COUNT OF TYPES AND TOKENS FOR EACH ESSAY IN A SET
* OF ESSAYS.
*

INPUT(.INPUT,105)

OUTPUT(.OUTPUT,108)

&ANCHOR = 1

SPACER = ' .-/()&" , ; : ? ! ' .

HYPHEN = '-'

BLANKS = SPAN(SPACER)

PAT = BREAK(SPACER) . WORD SPAN(SPACER) . CH

NUMBER = ANY('123456789')

*

* Function which extracts words from a file and counts
* the total number of words and the number of distinct
* words.

```

      DEFINE('READ()')      : (PR)
READ  TOKEN = 0
      NUMWORD = TABLE(100,10)
NEXTL TEXT = INPUT ' '      : F(BACK)
      TEXT NUMBER      : S(BACK)
      TEXT BLANKS =
GOOD  TEXT PAT =           : F(NEXTL)
HY    IDENT(CH,HYPHEN)     : F(KEEP)
      SAVE = WORD
      TEXT PAT =
      WORD = SAVE '- ' WORD : (HY)
KEEP  NUMWORD<WORD> = NUMWORD<WORD> + 1
      TOKEN = TOKEN + 1     : (GOOD)
BACK  READ = NUMWORD       : (RETURN)

```

* Function which outputs list and word counts

```

PR    DEFINE('PRINT(OUT) I') : (GO)
PRINT OUTPUT = 'ESSAY # ' ESSAY
      OUTPUT = '-----'
      OUTPUT =
      OUTPUT =
      OUTPUT = 'WORD          OCCURRENCES'

```

123

```

        OUTPUT = '-----'
        OUTPUT =
NO      I = I + 1
        OUTPUT = OUT<I,1> DUPL(' ',18 - SIZE(OUT<I,1>))
        - SIZE(OUT<I,2>)) OUT<I,2>      :S(NO)
        OUTPUT = '-----'
        OUTPUT =
        OUTPUT = 'NUMBER OF TYPES = ' I - 1
        OUTPUT = 'NUMBER OF TOKENS = ' TOKEN      : (RETURN)
*-----*
***** Main body of program
*
GO      DUMMY = INPUT
NEXTE   ESSAY = LT(ESSAY,18) ESSAY + 1      :F(END)
        WDLIST = READ()
        ALPHA = SORT(WDLIST)
        PRINT(ALPHA)      : (NEXTE)
END
```

```

C***** COEFFICIENT OF CONCORDANCE (KENDALL)
C
C      BARRY DONAHUE
C      MONTANA STATE UNIVERSITY
C      APRIL 1, 1982
C
C      SSD = SUM OF SQUARES OF DEVIATIONS ABOUT RANK MEAN
C      ERS = EXPECTED RANK SUM
C      TS = TOTAL SUM OF RANKS
C      X2 = CHI SQUARED
C
  DIMENSION SUM(20)
  DIMENSION K(20,10)
  REAL K
  DO 5 I=1,20
5    SUM(I) = 0
  SSD = 0
  TS = 0
  M = 4
  N = 18
  DO 10 I=1,N
  DO J=1,M
  READ(105,50) K(I,J)
  SUM(I) = SUM(I) + K(I,J)
10  CONTINUE
  DO 15 I=1,N
  TS = TS + SUM(I)
15  CONTINUE
  ERS = TS/N
  DO 30 I=1,N
  SSD = SSD + (SUM(I) - ERS)**2
30  CONTINUE
  W = (12*SSD)/((M**2)*(N**3-N))
  X2 = M*(N-1)*W
  WRITE(108,75)
  OUTPUT ' METHODS'
  DO 40 I=1,N
  WRITE(108,60) I,(K(I,J),J=1,M)
40  CONTINUE
  OUTPUT '-----'
  WRITE(108,70) W,N-1,X2
75  FORMAT(2/)
60  FORMAT(/,' ESSAY ',I2,10X,10(F4.1,5X))
70  FORMAT(/,' KENDALL W = ',F7.6,/, ' DEGREES OF FREEDOM
  * = ',I2,/, ' CHI SQUARED = ',F5.2)
50  FORMAT(F4.1)
  END

```

```

C***** RELIABILITY COEFFICIENT FOR RATERS OF ESSAYS
C
C      BARRY DONAHUE
C      MONTANA STATE UNIVERSITY
C      JUNE 15, 1981
C
C      SSR=SUM OF SQUARED RATINGS
C      SR=SUM OF A RATER'S SCORES
C      SE=SUM OF AN ESSAY'S SCORES
C      S2R=SUM SQUARED OF A RATER'S SCORES
C      S2E=SUM SQUARED OF AN ESSAY'S SCORES
C      TS2R=TOTAL OF SUMS SQUARED OF A RATER'S SCORES
C      TS2E=TOTAL OF SUMS SQUARED OF AN ESSAY'S SCORES
C      SOSR=SUM OF SQUARES FOR RATERS
C      SOSES=SUM OF SQUARES FOR ESSAYS
C      SOST=SUM OF SQUARES FOR TOTAL
C      SOSER=SUM OF SQUARES FOR ERROR
C      MSES=MEAN SQUARE FOR ESSAYS
C      MSER=MEAN SQUARE FOR ERROR
C      RELI=RELIABILITY FOR INDIVIDUAL RATINGS
C      RELA=RELIABILITY FOR AVERAGE RATINGS
C      HEADER=HEADING OF THE FILE
C      K=NUMBER OF RATERS
C      N=NUMBER OF ESSAYS
C      First record of input file must contain the heading
C      of the file; the second record must contain the
C      number of raters; the third record must contain the
C      number of essays.
C
C      DIMENSION S2E(20),S2R(20),L(20,20),SR(20),SE(20)
C      DIMENSION AVE(20)
C      INTEGER  HEADER(5)
C      REAL     MSER,MSES
C
C      TS2R=0
C      TS2E=0
C      SSR=0
C      TOTALSUM=0
C
C      initialize all arrays to 0
C      DO 200 I=1,20
C          S2E(I)=0
C          S2R(I)=0
C          SR(I)=0
C          SE(I)=0
200  CONTINUE

```

C input header, number of raters, and number of essays

100 FORMAT(12)

110 FORMAT(20I2)

133 FORMAT(5A4)

READ(105,133) (HEADER(I), I=1,5)

READ(105,100) K

READ(105,100) N

C

C input a rater's scores and process

DO 10 I=1,K

READ(105,110) (L(J,I),J=1,N)

C

DO 15 J=1,N

SR(I)=SR(I)+L(J,I)

SE(J)=SE(J)+L(J,I)

TOTALSUM=TOTALSUM+L(J,I)

SSR=SSR+L(J,I)**2

15 CONTINUE

S2R(I)=SR(I)**2

TS2R=TS2R+S2R(I)

10 CONTINUE

C

C calculate TS2E and essay averages

DO 30 J=1,N

S2E(J)=SE(J)**2

TS2E=TS2E+S2E(J)

AVE(J)=SE(J)/K

30 CONTINUE

C

C perform necessary calculations

Z=TOTALSUM**2/(N*K)

SOSR=TS2R/N-Z

S0SES=TS2E/K-Z

SOST=SSR-Z

SOSER=SOST-S0SES-SOSR

MSES=S0SES/(N-1)

MSFR=SOSER/((N-1)*(K-1))

RELI=(MSES-MSER)/(MSES+(K-1)*(MSER))

RELA=(MSES-MSER)/MSES

C

C output

WRITE(108,75) (HEADER(I), I=1,5)

WRITE(108,65) K, (I, I=1,K)

WRITE(108,82) K

DO 50 J=1,N

WRITE(108,60) J,K,(L(J,I), I=1,K),AVE(J)

```
50 CONTINUE
   WRITE(108,82) K
82  FORMAT(' ',22,'-',N('-----'))
   WRITE(108,70) RELI,RCLA
65  FORMAT(16X,N(I2,3X),'AVERAGE')
75  FORMAT(2//,29X,'RELIABILITY COEFFICIENT',//X,5A4,5//
*38X,'RATERS')
60  FORMAT(' ESSAY #',I2,6X,N(I2,3X),F7.4)
70  FORMAT(//,' RELIABILITY OF INDIVIDUAL RATINGS = ',F7.6,
*//,' RELIABILITY OF AVERAGE RATINGS = ',F7.6)
   END
```

APPENDIX B
RAW SCORES OF RATERS

Table 21

Reliability of Ratings by the Student Group Using Holistic Scoring

Essay	Raters														Average
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
1	1	2	1	1	1	1	1	2	1	1	1	2	2	2	1.36
2	5	5	5	4	5	4	4	5	5	5	5	5	5	5	4.79
3	2	2	3	2	2	2	2	1	2	3	2	4	2	3	2.29
4	5	4	4	3	4	3	4	4	5	3	4	5	4	3	3.93
5	3	4	2	3	3	4	4	2	4	3	3	5	4	4	3.43
6	3	3	2	3	3	3	3	2	3	2	3	4	3	2	2.79
7	4	4	4	4	4	5	3	5	5	4	3	5	4	2	4.00
8	3	3	3	3	3	1	3	3	2	3	2	4	2	2	2.64
9	2	3	3	3	3	5	3	3	3	3	3	5	4	3	3.29
10	2	1	2	1	3	2	1	1	1	2	1	3	1	3	1.71
11	2	3	1	3	4	3	2	3	1	3	2	3	1	1	2.29
12	1	1	2	1	1	4	1	2	3	3	1	2	2	1	1.79
13	2	4	3	1	3	3	2	3	3	3	2	4	3	2	2.71
14	3	4	3	3	3	3	2	3	4	3	2	5	4	3	3.21
15	4	3	3	4	3	4	3	4	5	3	3	5	3	4	3.64
16	2	3	3	2	3	5	3	3	4	3	2	5	3	3	3.14
17	4	4	3	2	2	5	4	4	4	4	3	5	3	4	3.64
18	3	4	3	5	4	4	3	3	4	3	2	5	5	3	3.64

129

Reliability of Individual Ratings = .61

Reliability of Average Ratings = .96

Table 22

Reliability of Ratings by the Expert Group Using Holistic Scoring

Essay	Raters										Average
	1	2	3	4	5	6	7	8	9	10	
1	1	1	1	1	2	1	1	1	2	2	1.30
2	5	5	4	5	5	5	5	5	4	5	4.80
3	2	3	3	2	3	2	2	2	1	3	2.30
4	4	5	3	5	4	5	3	4	5	5	4.30
5	2	3	3	3	3	3	3	2	2	4	2.80
6	3	4	3	2	2	2	3	2	2	1	2.40
7	4	4	3	3	4	4	4	2	3	3	3.40
8	4	3	2	3	4	3	3	3	3	3	3.10
9	3	4	3	3	4	4	3	2	2	2	3.00
10	1	2	2	1	2	3	2	2	2	1	1.80
11	2	2	3	2	3	3	1	1	2	2	2.10
12	1	2	2	1	2	1	2	1	1	1	1.40
13	3	4	1	2	2	1	2	2	2	3	2.20
14	4	5	3	2	4	3	3	2	3	3	3.20
15	5	5	3	4	4	4	4	3	4	4	4.00
16	2	4	2	3	3	3	3	3	3	3	2.90
17	4	4	3	3	3	3	4	3	3	4	3.40
18	3	5	2	4	3	2	4	2	3	3	3.10

Reliability of Individual Ratings = .68

Reliability of Average Raings = .96

Table 23

Reliability of Ratings for the Category "Ideas" by the
Student Group Using Atomistic Scoring

Essay	Raters										Average
	1	2	3	4	5	6	7	8	9	10	
1	10	8	8	8	10	10	10	10	10	10	9.40
2	2	10	2	6	4	2	6	2	4	6	4.40
3	8	6	8	6	8	2	6	6	6	6	6.20
4	8	6	2	6	2	2	8	5	2	10	5.00
5	2	6	6	4	6	4	4	6	4	6	4.80
6	4	10	10	6	6	8	6	6	8	4	6.80
7	4	6	6	6	4	2	4	2	6	6	4.60
8	2	8	10	6	4	6	8	2	6	8	6.00
9	2	6	10	4	4	4	4	6	6	2	4.80
10	10	10	8	6	10	8	6	8	8	10	8.40
11	10	8	9	8	8	2	10	8	8	6	7.70
12	8	10	8	8	8	10	10	4	8	10	8.40
13	8	6	6	10	10	3	8	8	6	8	7.30
14	6	6	6	2	6	2	4	4	6	4	4.60
15	6	4	6	6	6	2	2	4	2	10	4.80
16	6	4	6	2	4	4	2	6	6	6	4.60
17	4	6	6	4	6	2	4	4	4	10	5.00
18	4	6	8	2	6	2	10	8	2	8	5.60

Reliability of Individual Ratings = .35

Reliability of Average Ratings = .84

Table 24

Reliability of Ratings for the Category "Organization" by
the Student Group Using Atomistic Scoring

Essay	Raters										Average
	1	2	3	4	5	6	7	8	9	10	
1	10	8	10	8	8	10	10	10	8	8	9.00
2	2	10	10	2	4	4	6	2	4	2	4.60
3	6	8	10	4	8	2	8	4	8	2	6.00
4	2	6	4	4	6	2	10	4	4	8	5.00
5	6	8	6	4	6	6	6	4	6	8	6.00
6	6	8	8	8	6	10	4	6	8	4	6.80
7	4	6	4	2	4	2	8	4	4	4	4.20
8	4	8	8	10	4	6	6	4	8	6	6.50
9	4	8	10	2	4	4	6	4	8	4	5.40
10	8	10	8	4	8	4	6	6	8	8	7.00
11	8	8	6	6	10	2	10	10	10	4	7.40
12	6	10	8	6	6	10	10	6	8	6	7.60
13	6	6	4	6	10	4	2	4	8	6	5.60
14	6	6	6	2	4	4	6	2	6	4	4.60
15	6	6	4	2	2	2	4	2	2	6	3.60
16	4	6	8	2	4	2	6	6	6	10	5.40
17	6	6	2	6	4	2	4	2	4	8	4.40
18	6	6	2	4	6	2	10	4	6	6	5.20

Reliability of Individual Ratings = .28

Reliability of Average Ratings = .79

Table 25

Reliability of Ratings for the Category "Wording" by the
Student Group Using Atomistic Scoring

Essay	Raters										Average
	1	2	3	4	5	6	7	8	9	10	
1	5	3	4	4	5	4	5	5	5	5	4.50
2	1	2	3	2	2	2	2	2	2	2	2.00
3	2	4	4	3	4	3	4	4	4	3	3.50
4	3	2	1	4	2	1	5	1	1	4	2.40
5	3	2	2	1	3	3	3	1	2	2	2.20
6	3	3	5	1	3	5	2	2	3	3	3.00
7	2	2	2	2	3	2	3	2	4	2	2.40
8	1	3	4	4	3	3	2	1	4	5	3.00
9	1	3	4	4	4	3	2	2	3	3	2.90
10	5	5	2	3	5	5	2	3	5	5	4.00
11	3	3	3	3	4	2	4	4	4	2	3.20
12	4	4	5	5	5	5	4	4	4	2	4.20
13	3	3	1	3	3	3	3	2	3	3	2.70
14	2	3	3	3	3	3	3	2	4	4	3.00
15	2	2	1	2	2	1	1	2	3	3	1.90
16	2	2	2	2	3	3	2	1	2	1	2.00
17	3	2	3	2	3	3	4	2	3	5	3.00
18	2	2	3	2	3	2	5	2	3	3	2.70

Reliability of Individual Ratings = .37

Reliability of Average Ratings = .75

Table 26

Reliability of Ratings for the Category "Flavor" by the
Student Group Using Atomistic Scoring

Essay	Raters										Average
	1	2	3	4	5	6	7	8	9	10	
1	5	3	3	2	4	3	4	5	4	5	3.80
2	1	3	4	4	3	2	3	2	2	1	2.50
3	2	4	4	4	4	2	5	3	3	2	3.30
4	4	2	2	4	3	1	5	2	1	5	2.90
5	2	2	3	3	2	2	3	2	3	3	2.50
6	2	3	4	5	4	4	2	4	4	3	3.50
7	1	3	3	3	3	2	2	4	3	3	2.70
8	2	3	5	4	3	3	3	1	3	4	3.10
9	1	3	5	4	3	2	1	3	4	3	2.90
10	4	5	3	1	4	4	1	3	5	5	3.50
11	4	3	4	5	5	1	4	4	4	2	3.60
12	4	5	4	3	4	4	5	3	4	2	3.80
13	3	3	3	4	5	3	2	2	3	4	3.20
14	2	3	2	5	3	2	2	2	3	3	2.70
15	2	3	2	4	2	1	3	2	3	4	2.60
16	3	3	2	3	2	1	2	2	3	2	2.30
17	2	2	4	4	3	2	3	1	2	4	2.70
18	2	2	4	4	3	2	5	3	3	4	3.20

Reliability of Individual Ratings = .10

Reliability of Average Ratings = .54

Table 27

Reliability of Ratings for the Category "Usage" by the
Student Group Using Atomistic Scoring

Essay	Raters										Average
	1	2	3	4	5	6	7	8	9	10	
1	5	3	2	2	3	4	5	5	4	3	3.60
2	1	3	3	3	1	3	2	1	1	1	1.90
3	2	4	5	2	3	3	4	3	4	3	3.30
4	3	2	3	3	2	2	4	2	3	3	2.70
5	3	2	4	1	2	3	2	2	2	4	2.50
6	2	3	5	1	2	4	3	3	4	2	2.90
7	2	3	1	1	2	3	3	3	3	2	2.30
8	2	3	3	2	2	1	2	1	4	4	2.40
9	2	3	3	2	2	3	4	2	3	3	2.70
10	5	4	2	5	4	5	1	2	5	5	3.80
11	4	3	3	2	3	2	3	5	4	1	3.00
12	3	5	2	2	2	3	5	4	3	2	3.10
13	3	3	1	3	4	2	3	2	3	3	2.70
14	3	3	4	3	4	5	4	3	3	4	3.60
15	1	3	3	2	1	2	2	3	3	3	2.30
16	1	3	3	1	2	3	2	1	3	2	2.10
17	3	3	4	2	2	1	3	2	2	4	2.60
18	2	3	2	2	2	3	5	2	2	3	2.60

Reliability of Individual Ratings = .17

Reliability of Average Ratings = .67

Table 28

Reliability of Ratings for the Category "Punctuation" by the
Student Group Using Atomistic Scoring

Essay	Raters										Average
	1	2	3	4	5	6	7	8	9	10	
1	5	2	2	1	2	3	5	5	4	4	3.30
2	1	1	2	1	1	3	1	1	1	1	1.30
3	3	2	3	2	4	3	5	2	4	4	3.20
4	4	2	2	2	3	3	4	3	2	3	2.80
5	1	2	3	2	1	3	4	2	2	2	2.20
6	3	5	5	3	2	3	3	3	4	4	3.50
7	2	4	2	3	2	3	4	3	4	4	3.10
8	3	2	2	2	3	2	3	1	4	4	2.60
9	1	2	3	2	2	3	4	2	4	4	2.70
10	5	5	3	3	4	5	3	5	5	4	4.20
11	4	3	4	2	4	3	3	5	4	4	3.60
12	3	5	4	2	3	4	4	3	3	4	3.50
13	2	3	2	2	4	2	2	2	3	3	2.50
14	3	4	4	5	5	5	4	3	3	4	4.00
15	2	5	4	2	3	2	3	3	3	3	3.00
16	2	3	3	2	4	3	3	2	4	3	2.90
17	3	4	5	4	2	2	4	2	3	5	3.40
18	2	2	5	2	1	2	5	2	2	3	2.60

Reliability of Individual Ratings = .31

Reliability of Average Ratings = .82

Table 29

Reliability of Ratings for the Category "Spelling" by the
Student Group Using Atomistic Scoring

Essay	Raters										AVERAGE
	1	2	3	4	5	6	7	8	9	10	
1	4	3	5	4	2	5	4	4	5	5	4.10
2	1	1	2	1	1	1	1	1	1	1	1.10
3	4	5	5	5	4	5	5	4	4	5	4.60
4	2	3	2	2	3	2	5	2	3	2	2.60
5	3	4	3	3	3	3	4	5	3	4	3.50
6	1	4	5	5	4	5	5	3	4	4	4.00
7	1	3	3	2	3	3	3	2	3	4	2.70
8	4	3	5	2	5	5	4	2	4	4	3.80
9	3	5	5	5	5	5	4	4	4	5	4.50
10	5	5	5	5	5	5	3	5	5	5	4.80
11	5	5	5	5	5	5	3	4	4	5	4.60
12	5	5	5	5	5	5	3	4	4	5	4.60
13	1	2	3	1	2	3	4	3	2	3	2.40
14	1	2	3	1	3	4	4	2	3	4	2.70
15	2	4	4	1	2	3	2	3	2	2	2.50
16	5	5	5	5	5	5	4	3	4	5	4.60
17	3	5	5	5	3	3	4	2	3	5	3.80
18	3	2	5	2	4	4	5	2	3	4	3.40

Reliability of Individual Ratings = .61

Reliability of Average Ratings = .94

Table 30

Reliability of the Total of All Categories by the Student
Groups Using Atomistic Scoring

Essay	Raters										Average
	1	2	3	4	5	6	7	8	9	10	
1	44	30	34	29	34	39	43	44	40	40	37.70
2	9	30	26	17	16	17	21	11	15	14	17.60
3	27	33	39	26	35	20	37	26	33	25	30.10
4	26	23	16	25	21	13	41	18	16	35	23.40
5	20	26	27	18	23	24	26	22	22	29	23.70
6	21	36	42	29	27	39	25	27	35	24	30.50
7	16	27	21	19	21	17	27	20	27	25	22.00
8	18	30	37	29	24	26	28	12	33	35	27.20
9	14	30	40	23	24	24	25	23	32	24	25.90
10	42	44	31	27	40	36	22	32	41	42	35.70
11	38	33	34	31	39	17	37	40	38	24	33.10
12	33	44	36	31	33	41	43	29	33	31	35.40
13	26	26	20	29	38	20	24	23	28	30	26.40
14	23	27	28	21	28	25	27	18	28	27	25.20
15	21	27	24	19	18	13	17	19	18	31	20.70
16	23	26	29	17	24	21	21	21	28	29	23.90
17	24	28	29	27	23	15	26	14	21	41	24.90
18	21	23	28	18	25	17	45	23	21	31	25.20

Reliability of Individual Ratings = .45

Reliability of Average Ratings = .89

Table 31

Reliability of Ratings for the Category "Ideas" by the
Expert Group Using Atomistic Scoring

Essay	Raters										Average
	1	2	3	4	5	6	7	8	9	10	
1	6	10	10	10	8	8	10	10	8	6	8.60
2	4	2	4	2	2	2	6	4	4	6	3.60
3	4	6	4	8	6	6	10	8	6	6	6.40
4	4	2	2	2	2	2	2	6	6	10	3.80
5	4	4	8	6	4	4	8	4	6	6	5.40
6	4	4	4	2	6	10	6	6	8	10	6.00
7	2	6	2	4	4	6	6	4	6	2	4.20
8	4	4	2	6	6	6	4	4	6	4	4.60
9	4	2	4	4	4	2	4	6	6	2	3.80
10	10	10	6	4	10	10	10	8	6	6	8.00
11	6	6	4	4	10	8	2	10	8	2	6.00
12	8	8	8	8	8	10	10	6	10	6	8.20
13	8	6	8	6	6	4	6	6	10	10	7.00
14	4	4	4	4	6	8	4	2	4	4	4.40
15	2	4	2	4	2	4	2	4	2	2	2.80
16	4	4	4	8	4	4	4	8	8	4	5.20
17	4	4	2	2	4	6	4	2	6	2	3.60
18	2	4	4	4	6	8	8	8	8	8	6.00

Reliability of Individual Ratings = .43

Reliability of Average Ratings = .88

Table 32

Reliability of Ratings for the Category "Organization" by the
Expert Group Using Atomistic Scoring

Essay	Raters										Average
	1	2	3	4	5	6	7	8	9	10	
1	6	10	10	8	8	8	8	10	6	6	8.00
2	4	2	4	4	2	4	8	4	4	6	4.20
3	6	4	6	8	6	6	10	6	2	6	6.00
4	4	2	4	4	4	2	2	6	4	8	4.00
5	6	4	8	4	6	4	8	6	8	4	5.80
6	4	4	6	6	8	10	4	2	6	10	6.00
7	2	2	2	4	4	6	6	8	2	4	4.00
8	4	2	6	6	6	6	8	6	4	4	5.20
9	4	4	2	4	2	2	4	4	4	2	3.20
10	8	8	6	2	10	10	10	6	4	4	6.80
11	6	6	4	2	8	8	2	6	10	2	5.40
12	8	8	6	10	8	10	8	4	10	4	7.60
13	10	8	4	4	6	4	6	6	8	8	6.40
14	4	2	4	2	6	6	4	6	8	2	4.40
15	2	2	2	2	2	2	2	2	2	2	2.00
16	4	4	4	6	6	6	4	8	6	4	5.20
17	2	4	2	4	4	6	2	2	4	2	3.20
18	2	4	4	2	6	10	6	2	6	8	5.00

Reliability of Individual Ratings = .34

Reliability of Average Ratings = .84

Table 33

Reliability of Ratings for the Category "Wording" by the
Expert Group Using Atomistic Scoring

Essay	Raters										Average
	1	2	3	4	5	6	7	8	9	10	
1	3	5	5	5	3	4	5	5	2	5	4.20
2	1	1	2	2	3	3	3	2	3	2	2.20
3	3	3	3	5	3	4	4	3	3	4	3.50
4	2	1	2	3	3	3	3	2	3	4	2.60
5	3	3	3	1	3	3	3	2	2	3	2.60
6	3	2	3	3	3	3	3	3	3	4	3.00
7	2	1	1	2	3	3	3	3	3	3	2.40
8	3	3	2	3	3	1	3	1	3	3	2.50
9	3	1	3	3	3	2	3	2	2	1	2.30
10	2	5	4	3	5	5	5	5	1	3	3.80
11	3	3	3	3	4	4	3	2	4	2	3.10
12	4	5	3	4	3	3	3	4	3	3	3.70
13	3	3	3	4	3	3	3	3	5	2	3.20
14	2	3	2	3	3	4	3	3	4	2	2.90
15	3	3	3	3	2	3	2	2	3	1	2.50
16	2	2	3	4	3	2	3	3	3	3	2.80
17	2	1	3	3	2	3	3	1	3	3	2.40
18	2	3	3	3	4	4	3	2	3	4	3.10

Reliability of Individual Ratings = .27

Reliability of Average Ratings = .79

Table 34

Reliability of Ratings for the Category "Flavor" by the
Expert Group Using Atomistic Scoring

Essay	Raters										Average
	1	2	3	4	5	6	7	8	9	10	
1	4	4	4	5	3	5	5	5	2	3	4.00
2	1	1	2	1	2	2	3	2	1	1	1.60
3	3	2	2	4	2	3	5	3	3	3	3.00
4	2	2	1	3	5	3	1	2	4	5	2.80
5	2	2	3	3	3	3	4	2	3	3	2.80
6	3	2	2	3	3	4	3	2	4	5	3.10
7	2	3	1	3	3	3	3	2	3	3	2.60
8	4	3	3	3	2	3	2	2	3	3	2.80
9	3	2	4	3	3	3	3	2	2	1	2.60
10	4	4	3	3	4	4	4	4	1	3	3.40
11	3	3	3	3	4	5	1	3	5	2	3.20
12	4	4	3	4	3	3	4	4	5	3	3.70
13	3	3	5	3	4	3	3	3	5	5	3.70
14	2	2	3	3	4	4	3	2	3	2	2.80
15	2	3	3	3	2	1	1	2	3	1	2.10
16	1	2	3	4	2	3	3	3	5	3	2.90
17	2	2	3	2	3	4	3	1	4	3	2.70
18	2	3	2	3	5	3	3	1	4	5	3.10

Reliability of Individual Ratings = .21

Reliability of Average Ratings = .73

Table 35

Reliability of Ratings for the Category "Usage" by the
Expert Group Using Atomistic Scoring

Essay	Raters										Average
	1	2	3	4	5	6	7	8	9	10	
1	3	4	5	4	3	4	4	4	3	4	3.80
2	1	1	1	1	3	1	1	1	1	1	1.20
3	3	4	4	5	4	4	4	2	3	4	3.70
4	3	2	2	3	3	5	3	2	3	4	3.00
5	2	2	3	2	4	4	2	2	3	3	2.70
6	3	3	2	3	3	5	3	3	2	5	3.20
7	3	4	2	4	3	4	2	2	4	2	3.00
8	2	3	3	4	3	4	3	4	3	3	3.20
9	3	3	4	3	4	4	3	3	3	1	3.10
10	3	5	3	4	5	5	5	5	2	5	4.20
11	3	4	4	4	5	5	3	4	2	2	3.60
12	3	4	2	4	5	5	3	3	3	3	3.50
13	3	4	2	4	3	3	3	3	4	3	3.20
14	2	3	4	4	3	4	4	4	2	1	3.10
15	3	1	3	4	3	2	1	2	1	1	2.10
16	2	4	3	5	5	3	3	2	4	1	3.20
17	2	3	3	5	4	5	3	2	3	1	3.10
18	2	3	2	5	3	5	4	4	2	3	3.30

Reliability of Individual Ratings = .33

Reliability of Average Ratings = .83

Table 36

Reliability of Ratings for the Category "Punctuation" by the
Expert Group Using Atomistic Scoring

Essay	Raters										Average
	1	2	3	4	5	6	7	8	9	10	
1	2	4	4	3	2	2	3	3	3	3	2.90
2	1	4	1	1	2	1	1	2	1	1	1.50
3	3	3	4	3	5	5	2	4	2	3	3.40
4	3	4	4	4	4	4	3	3	5	4	3.80
5	2	3	3	2	3	2	2	1	2	3	2.30
6	2	3	3	4	3	5	3	2	5	4	3.40
7	3	4	4	3	4	4	4	4	4	1	3.50
8	2	4	4	3	3	4	4	1	4	2	3.10
9	3	4	3	3	5	4	2	3	5	1	3.30
10	3	4	3	5	5	5	5	4	4	4	4.20
11	4	4	4	4	4	5	4	4	4	2	3.90
12	4	5	4	3	4	3	4	4	5	3	3.90
13	3	5	4	3	5	4	4	4	5	2	3.90
14	2	3	5	5	3	5	5	5	4	3	4.00
15	2	4	4	3	3	3	3	3	1	1	2.70
16	2	4	4	3	5	5	3	4	2	1	3.30
17	3	4	4	4	5	3	3	5	4	2	3.70
18	2	4	2	4	3	5	4	2	2	2	3.00

Reliability of Individual Ratings = .34

Reliability of Average Ratings = .84

Table 37

Reliability of Ratings for the Category "Spelling" by the
Expert Group Using Atomistic Scoring

Essay	Raters										Average
	1	2	3	4	5	6	7	8	9	10	
1	4	4	4	5	4	5	3	4	4	3	4.00
2	1	1	1	1	1	1	1	1	1	1	1.00
3	5	5	5	5	5	5	5	3	5	3	4.60
4	3	2	1	4	2	4	3	2	5	4	3.00
5	4	4	4	5	3	5	3	3	5	4	4.00
6	4	5	3	5	5	5	5	2	5	4	4.30
7	1	4	4	5	3	4	3	2	5	2	3.30
8	2	4	4	5	4	5	2	2	4	3	3.50
9	3	4	4	5	4	5	4	4	4	5	4.20
10	5	5	4	5	5	5	5	5	5	5	4.90
11	5	4	4	5	5	5	4	4	5	3	4.40
12	5	5	4	5	5	5	5	5	5	4	4.80
13	2	3	2	2	3	3	2	2	5	2	2.60
14	1	2	1	2	3	3	2	1	4	2	2.10
15	2	3	4	3	3	3	1	2	4	1	2.60
16	3	5	5	5	5	5	5	5	5	5	4.80
17	3	5	4	5	3	5	5	4	5	4	4.30
18	2	3	2	5	3	5	4	4	3	3	3.40

Reliability of Individual Ratings = .67

Reliability of Average Ratings = .95

Table 38

Reliability of the Total of All Categories by the Expert
Group Using Atomistic Scoring

Essay	Raters										Average
	1	2	3	4	5	6	7	8	9	10	
1	26	41	42	40	31	36	38	41	28	30	35.30
2	14	12	15	12	15	14	23	16	15	18	15.30
3	27	27	28	38	31	33	40	29	24	29	30.60
4	21	15	16	23	23	23	17	23	30	39	23.00
5	23	22	32	23	26	25	30	20	29	26	25.60
6	23	23	23	26	31	42	27	20	33	42	29.00
7	15	24	16	25	24	30	27	25	27	17	23.00
8	21	23	24	30	27	29	26	20	27	22	24.90
9	23	20	24	25	25	22	23	24	26	13	22.50
10	35	41	29	26	44	44	44	37	23	30	35.30
11	30	30	26	25	40	40	19	29	38	15	29.30
12	36	39	30	38	36	39	37	30	43	26	35.40
13	32	32	28	26	30	24	27	27	42	32	30.00
14	17	19	23	23	28	34	25	23	29	16	23.70
15	16	20	21	22	17	18	12	17	16	9	16.80
16	18	25	26	35	30	28	25	33	33	21	27.40
17	18	23	21	25	25	32	23	17	29	17	23.00
18	14	24	19	26	30	40	32	23	28	33	26.90

Reliability of Individual Ratings = .51

Reliability of Average Ratings = .91

APPENDIX C

INTERMEDIATE RESULTS FROM CALCULATION
OF MATURE WORD INDEX

Table 39

Mature Words Used in the Essays

affecting	disadvantages	keg	resource
affects	disregard	litter	ruin
alter	distillery	long-range	ruined
alternative	disturbed	manual	security
alternatives	drain	microorganisms	seeps
argue	dump	minimal	shutting
awhile	dumped	nature's	sicken
beer	dumping	nonpollution	sickness
benefits	ecology	operating	someone's
businesses	editorial	outlets	stability
cheaper	elimination	overlooked	starving
cleaner	employs	permanently	summation
community's	encounter	personally	surpass
company's	environmental	personnel	symptoms
compromise	everyone's	petitions	temporarily
consequences	expendable	poisoned	temporary
constructively	extensive	potential	tenous
consumes	facets	preservation	threat
contaminating	faulty	priority	thrive
contamination	feeds	profit	totally
continual	filters	prominent	townspeople's
contractors	fined	punished	toxic
controlling	grocery	purified	tumble
convince	handicaps	purify	ultimately
convincing	harming	pyramid	unemployment
cozy	hatchery	qualified	unfeasible
damaged	hell	reap	unpure
deaths	hire	rebuilding	vicinity
decline	immediate	recourse	wanton
definitely	indefinite	reduction	wastes
destroying	infraction	remodeling	wells
destruction	innumerable	reopens	wholly
deteriorating	insist	repercussions	workless
detriment	installing	representative	year's
devastating	involve	residents	
disadvantage	irrigated	resort	

Table 40

Contractions, Proper Nouns, and Slang Used in the Essays

Bozeman	New York
grandkids	stink
it'll	uptight
junk	

Table 41

Topic Imposed Words Used in the Essays

pollute
polluted
polluting
pollution

Table 42

Number of Types and Tokens Used in the Essays

Essay	Number of Types	Number of Tokens
1	160	274
2	114	205
3	101	176
4	89	148
5	73	129
6	106	153
7	73	134
8	91	179
9	70	110
10	140	217
11	123	264
12	145	246
13	135	248
14	81	124
15	72	121
16	61	108
17	71	103
18	69	106
TOTAL	714	3044

MONTANA STATE UNIVERSITY LIBRARIES
stks D378.D714@Theses RL
Indicators of quality in natural languag



3 1762 00177749 7

D378
D714
Cap 2