THE ROLE OF COMPREHENSION IN HOLISTIC EVALUATION OF SECOND
LANGUAGE WRITING PROFICIENCY AT THE UNIVERSITY LEVEL

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of the Ohio State University

By
Michael Janopulos, B.S., M.A., M.A.

****************
The Ohio State University
1987

Dissertation Committee:

E.D. Allen
E.B. Bernhardt
R. Kantor

Approved by

E.D. Allen
Advisor
College of Education
Copyright by
Michael Janopoulos
1987
THE ROLE OF COMPREHENSION IN HOLISTIC EVALUATION OF SECOND
LANGUAGE WRITING PROFICIENCY AT THE UNIVERSITY LEVEL

By

Michael Janopoulous, Ph.D.

The Ohio State University, 1987

Professor Edward D. Allen, Adviser

Holistic evaluation is generally accepted as a
reliable and valid means of assessing writing proficiency
in both L1 and L2. Yet questions persist concerning how
well holistic raters comprehend what they read, especially
where texts generated by non-native speakers (NNS) are
concerned. The researcher assumes that any measure of
proficiency must provide for the extent to which meaning
can be accessed by the receiver. However, the nature of
texts generated by NNS authors may create task demands for
the holistic reader that complicate the reading task to
the extent that comprehension is significantly hindered.
The purpose of this experiment was to test the construct
validity of holistic scoring of L2 writing proficiency.

In this experiment, twelve experienced ESL composition
instructors holistically evaluated several essays written
by NNS graduate students attending an American university.
After satisfactory interrater reliability was established,
subjects generated recall protocols based on two NNS essays, one better written, the other of lesser quality. Reader orientation (naive or focused) and order of presentation were varied so that six protocols were produced for each of four recall conditions. Protocols were scored using a weighted propositional count and data were analyzed by means of a 2 X 2 ANOVA.

Data analysis revealed that naive holistic readers (e.g., raters reading only for the purpose of evaluation) recalled significantly more of the content of the better written text than of the lower quality text. Results also indicated that focused holistic readers (e.g., raters who were aware that a recall task was to follow) recalled significantly more than naive readers overall, but did not recall more of the better written text.

On the basis of the data, the researcher concluded that holistic scoring provides a valid measure of L2 writing proficiency insofar as a meaning-based construct of writing proficiency is concerned. The failure of focused holistic scorers to recall significantly more of the better written text is hypothesized as being the result of a variety of factors, including a possible ceiling effect and the influence of multiple task demands on depth-of-processing.
To my wife, Debbie, whose encouragement, support and, above all, understanding made this endeavor possible
ACKNOWLEDGEMENTS

Any endeavor of this magnitude is, of necessity, a cooperative one. It is therefore not out of any sense of false modesty that I admit to owing a large number of friends, faculty, and colleagues an immense debt of gratitude. Indeed, I state in all candor that this task could not have been accomplished without them.

To my advisor, Edward D. Allen, I owe much. Without his encouragement and support, I would not have even come to the Ohio State University. Once enrolled, Dr. Allen guided me down a path that led me to my ultimate goal. Through a combination of wisdom borne of experience, gentle wit, and keen insight, he was my advisor, mentor, and, above all, my friend.

To the members of my reading committee I owe equally as much. Elizabeth B. Bernhardt and Robert Kantor, along with Dr. Allen, provided me with a wealth of guidance, information, and constructive criticism. Their forebearance and willingness to give freely of their time and expertise were instrumental in my work.
In particular, I am indebted to Dr. Bernhardt for help in conceptualizing this study and Dr. Kantor for providing me access to the human and material resources of the ESL Composition unit of the Ohio State University.

I wish to thank my colleagues in the ESL Composition unit who cooperated in supplying student compositions, serving as subjects in the study, and assisting in the validation procedures. In a very real sense, these people were more than just "representative samples" of a typical ESL composition program. Rather, their patience, encouragement, and willingness to give of their time and effort made them a special group indeed.

I also wish to thank Majorie Demel for her willingness to share of her time and expertise in using the Meyer text analysis system. Without her assistance, my task would have been immeasurably more difficult.

Finally, I would like to thank my family. We entered into this pursuit together, but with no clear understanding of either the effort it would take or the emotional and physical toll it would exact on us. We come to the end of this task with a sense of closeness and mutual commitment that results from facing a challenge as one. Thank you, Gregory, for giving focus and incentive to my work. All I do, I do for you. Most of all, thank you, Debbie. Your unselfishness, patience, sensitivity, and love made this project not only possible, but worth the effort.
VITA

March 2, 1950............ Born – Chicago, Illinois

1972.................... B.S., Northwestern University, Evanston, Illinois

1972.................... M.A., Northwestern University, Evanston, Illinois


1973-1974................ Reading Resources Specialist, Keokuk Community Schools, Keokuk, Iowa

1976.................... M.A., University of Northern Iowa, Cedar Falls, Iowa

1976-1980................ ESL Instructor, Cairo American College, Ma'adi, Egypt

1980-1983................ Chair, ESL Department, Abqaiq School, ARAMCO School System, Abqaiq, Saudi Arabia

1983-1984................ Senior Instructor in ESL, Industrial Training Center, College Preparatory Program, ARAMCO, Dhahran, Saudi Arabia

1984-1985................ University Fellow, The Ohio State University, Columbus, Ohio

1985-1987................ Graduate Teaching Associate, ESL Composition Unit, The Ohio State University, Columbus, Ohio
PUBLICATIONS


FIELDS OF STUDY

Major Field: Foreign Language Education


# TABLE OF CONTENTS

DEDICATION .............................................................. ii  
ACKNOWLEDGEMENTS ................................................... iii  
VITA ........................................................................... v  
LIST OF TABLES ............................................................ xii  
LIST OF FIGURES ........................................................... xiii  

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. PROBLEM STATEMENT</td>
<td>1</td>
</tr>
<tr>
<td>Introduction ..................</td>
<td>1</td>
</tr>
<tr>
<td>Significance ..................</td>
<td>1</td>
</tr>
<tr>
<td>Purpose ........................</td>
<td>3</td>
</tr>
<tr>
<td>Assumptions ..................</td>
<td>4</td>
</tr>
<tr>
<td>Definition of Terms ..........</td>
<td>5</td>
</tr>
<tr>
<td>Limitations ...................</td>
<td>9</td>
</tr>
<tr>
<td>Research Question ............</td>
<td>10</td>
</tr>
<tr>
<td>II. LITERATURE REVIEW</td>
<td>12</td>
</tr>
<tr>
<td>Overview ........................</td>
<td>12</td>
</tr>
<tr>
<td>Research in Second Language (L2) Writing Proficiency ..................</td>
<td>16</td>
</tr>
<tr>
<td>Generalizations from First Language (L1) Research ....................</td>
<td>16</td>
</tr>
</tbody>
</table>
L1/L2 Research .................................................. 16
Shortcomings of L2 Research ............................... 18
Evaluation in L2 Writing .................................. 19
Holistic Evaluation of L2 Writing
Proficiency: The Untested Premise ...................... 20
Writing Proficiency ......................................... 20
  Elements of Proficiency ................................ 20
  The Writer Variable ......................................... 22
  The Reader Variable ........................................ 22
  Proficiency and Processing .............................. 23
  Definitions of Proficiency ............................... 28
Components of L1 Writing Proficiency ................. 29
Components of Writing Proficiency: An
L2 Perspective ............................................... 34
  A Meaning-based Definition of L2
  Writing Proficiency ....................................... 35
Holistic Scoring .............................................. 36
  Rationale for Use in L2
  at the University Level ................................ 36
  History ......................................................... 36
  Definition ..................................................... 37
  Variations ..................................................... 38
  Reliability .................................................... 39
  Validity ......................................................... 42
    Introduction ............................................... 42
    Dimensions of Validity .................................. 43
    Claims for the Validity of Holistic
    Assessment of Writing Proficiency .................. 44
    Some Unanswered Questions Vis-a-Vis
    the Validity of Holistic Evaluation ................ 46
Holistic Scoring and Comprehension:
  The Case Against ......................................... 47
  Application to Assessment of L2 Texts .............. 49
Recall Protocols ............................................. 49
  Definition, Applications, And Procedures .......... 49
Recall Protocols and Comprehension ..................... 51
  Widespread Use of the Procedure ...................... 51
  Underlying Theory .......................................... 52
  Empirical Support ......................................... 52
Content Structure as Presented and Perceived ...... 53
  Background .................................................. 53
Studies in First Language
Composition and Recall ................................... 54
Studies in Second Language
Composition and Recall ................................... 55
Implications for Holistic Evaluation
of L2 Writing Proficiency ................................ 55

viii
Scoring Procedures........................................ 58
Introduction and Purpose.................................. 58
The Kintsch System......................................... 59
The Frederiksen System.................................... 60
The Meyer System.......................................... 60
Identification of Idea Units............................... 61
Problems in Scoring Recall Protocols.................... 64
The Issue of Focus......................................... 64
Task Awareness and Depth of Processing.................. 64
Shared Goals and Holistic Assessment.................... 71
Research Hypotheses...................................... 72

III. PROCEDURES.............................................. 74

Population.................................................. 74
Sample Selection.......................................... 74
Research Design.......................................... 74
Variables................................................... 76
Independent................................................ 76
Dependent.................................................. 76
Experimental Procedures.................................. 76
Overview of the Plan....................................... 76
Overview of the Program................................... 80
Instrumentation............................................ 81
Measurement of Reading Comprehension................... 81
  Procedures for Collecting Writing Samples............ 81
  Procedures for Developing Protocol
  Scoring Matrices (PSMs).................................. 83
Measurement of Writing Proficiency....................... 85
Procedures for Training Holistic Raters.................. 86
Procedures for Establishing
  Interrater Reliability.................................. 87
Procedures for Eliciting "Naive"
  and "Focused" Recall Protocols......................... 87
Data Analysis.............................................. 90
Summary of Null Hypotheses............................... 93

IV. RESULTS AND DISCUSSION.............................. 95

Introduction.............................................. 95
  Experimental Data on the Recall
  Dependent Variable..................................... 99
  Results of the Null Hypotheses......................... 107
Discussion............................................... 111
V. SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

Summary of the Study..................................................115
Summary of Findings..................................................120
Conclusions...............................................................121
Limitations...............................................................134
Recommendations for Further Research.........................137

LIST OF REFERENCES....................................................139

APPENDICES...............................................................150

A. Diagnostic Essay Questions (Alternate Forms).....................150

B. Protocol Scoring Matrices for Higher (FORM A) and Lower (FORM B) Texts.................151

C. Instructions Concerning Criteria for Collecting Sample Texts.........................158

D. Original Texts (Levels 107 and 108)..............................160

E. Calibration Samples.................................................164

F. Practice Texts (Levels 106 through Q)..............................172

G. Reliability Texts......................................................177

H. Guidelines for Generating Negotiated Texts..........................191

I. Negotiated Texts (Levels 107 and 108)..............................192

J. Criteria for the Identification of Idea Units........................196

K. Valid Propositions and Their Mean Ratings of Importance to the Two Target Texts........197

L. Holistic Grading Criteria............................................199

M. Subject Score Sheet..................................................201
N. Coded Cover Sheet for Recall Protocols....... 202
O. Instructions for Writing Recall Protocols......................... 203
<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Means and Standard Deviations of Recall Scores by Reader Orientation and Proficiency Assessment</td>
<td>100</td>
</tr>
<tr>
<td>2. Analysis of Variance of Reading Comprehension by Reader Orientation and Proficiency Assessment</td>
<td>102</td>
</tr>
<tr>
<td>3. Means and Standard Deviations of Individual Cells</td>
<td>104</td>
</tr>
<tr>
<td>4. Analysis of Variance of Reading Comprehension (Lower/Naive by Higher/Naive)</td>
<td>105</td>
</tr>
<tr>
<td>5. Analysis of Variance of Reading Comprehension (Lower/Naive by Lower/Focused)</td>
<td>106</td>
</tr>
<tr>
<td>6. Analysis of Variance of Reading Comprehension (Focused/Higher by Focused/Lower)</td>
<td>108</td>
</tr>
<tr>
<td>7. Analysis of Variance of Reading Comprehension (Naive/Higher by Focused/Higher)</td>
<td>110</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. 2 X 2 Factorial Design</td>
<td>75</td>
</tr>
<tr>
<td>2. Order of Administration of Naive and Focused Reading Tasks with Respect to Target Texts</td>
<td>89</td>
</tr>
</tbody>
</table>
CHAPTER I

PROBLEM STATEMENT

Introduction

Holistic evaluation of written discourse enjoys a great deal of support among practitioners and researchers alike as a reliable (Cooper, 1977) and valid (Perkins, 1983) tool for assessing both first (L1) and second (L2) language writing proficiency. Yet anecdotal evidence suggests that it is possible for holistic raters to achieve an extremely high level of both interrater and intrarater reliability without actually comprehending the content of what they have read. The question that arises from such an apparent anomaly is one that has an impact upon the construct validity of holistic evaluation; namely, to what extent does a holistic rater's evaluation of written discourse reflect what he/she has comprehended?

Significance

The issue of rater comprehension is of importance in the realm of evaluating ESL compositions for the purpose of placing non-native speakers (NNSs) in remedial writing
programs in American universities. This is so because
the heavy emphasis placed upon being able to communicate
effectively in a written medium within one's area of
academic concentration makes quick and accurate
assessment of overall writing proficiency imperative.

Because it involves a direct means of rank-ordering
students by overall attained writing proficiency,
holistic evaluation of L2 writing proficiency employing
trained and experienced L1 raters is a commonly used
procedure for the purpose of placement of foreign
students in American universities. Moreover, a
consistently high level of both interrater and intrarater
reliability can be achieved and maintained, provided the
guidelines set down by Jacobs, et al. (1981) are
faithfully adhered to. Indeed, interrater reliability
coefficients of as high as .90 have been reported, and
although correlations of such magnitude are admittedly
rare (Braddock, et al., 1963 and Follman and Anderson,
1967), the levels of reliability that can be consistently
achieved by both researchers and practitioners are
considered to be quite adequate for the purpose of
initial placement into an ESL composition course sequence.

But while it can be argued convincingly that
holistic scoring - if properly done - possesses not only
reliability (Perkins, 1983) but also concurrent validity
(Kaczmarek, 1980), the case for construct validity rests on somewhat more shaky ground, assuming that the construct of writing proficiency includes a significant component of comprehension. Notwithstanding assertions like the one by Perkins (1983) that "holistic scoring has the highest construct validity when overall attained writing proficiency is the construct to be assessed" (p.652), it remains to be demonstrated that raters spending on average no more than two minutes per paper (Cooper, 1977) comprehend what they have read.

**Purpose**

While it cannot be denied that holistic evaluation of written discourse can provide a highly reliable and valid measure of something within the writing process, whether that "something" is indeed writing proficiency has yet to be established. The research study that follows sought to provide a framework within which this issue could be systematically addressed.

This experiment focused upon experienced and trained native speaking raters of ESL compositions at the university level as they engaged in a holistic evaluation task. The objective of this inquiry was to determine the extent to which holistic raters engaged in evaluating L2 compositions for the purpose of ascertaining the level of
writing proficiency demonstrated by authors of those compositions comprehended what they had read.

The theoretical basis for investigating the degree of covariance shared by indicators of the constructs of writing proficiency and comprehension is drawn from Cronbach's (1971) notion of convergence of indicators. Proceeding from the assumption that "persons who score high on the test ought to score high on other indicators of the same construct" (p. 466), this study tested the contention that measures of comprehension should evidence considerable covariance with measures of writing proficiency, if holistic scoring does, indeed, measure what it purports to measure. Specifically, if comprehension is an indicator of writing proficiency, then holistic evaluators of L2 texts should recall significantly more of the content of texts composed by more proficient authors than of texts composed by less proficient authors.

Assumptions

For the purposes of this study, the following assumptions were made:

1. subjects would provide an accurate measure of comprehension in their recall protocols;

2. adequate interrater reliability (e.g., a
minimum of .90) would be established during the calibration session;

3. the scoring system applied to recall protocols would yield valid and reliable data pertaining to rater comprehension;

4. sample compositions selected from the original corpus would accurately reflect a range of proficiency that could be reliably assessed on a four-point continuum from low to high;

5. writing samples assessed as being lower or higher in quality relative to one another would reflect the level of assessed proficiency of their authors.

**Definition of Terms**

For the purposes of this study, the following operational definitions were applied:

1. **Comprehension** - the degree to which a correspondence exists between the object to be understood and its internal representation; the extent to which a coherent, connected representation is achieved between the message as sent and the message as received (Greeno, 1977); measured on the basis of the amount of information correctly recalled in a written recall protocol of a text generated by a non-native speaker of English;
2. **Written Recall Protocol** - procedure wherein a subject reads a piece of prose, or written discourse (Voss, et al., 1982), and then recounts in as much detail as he/she can the contents of what was read;

3. **Writing Proficiency** - the amount of control of a writing system demonstrated by a non-native speaker responding to an essay topic; evaluated holistically on a continuum from level 106 (roughly Intermediate-Low on the ACTFL Proficiency Guidelines) to level Q (Superior on the ACTFL Proficiency Guidelines); reflected in the assessed quality of a writing sample, so that a text graded as high in relation to other texts is considered to be the product of a highly proficient NNS author;

4. **Holistic Scoring** - method of direct assessment of writing proficiency wherein the rater reads an ESL composition quickly and impressionistically, so as to form a judgment of the overall quality of the composition. In this study, scoring was done on a four-point scale, from 106 (reflecting the performance of a low proficiency NNS author) to Q (reflecting the performance of a high proficiency NNS author). These correspond to the three levels of ESL Composition instruction at the Ohio State University (106, 107, 108), plus Q (Qualified), which is the score assigned to compositions written by NNS students judged as proficient in academically-oriented
writing of the sort generally required at the graduate school level in American universities;

5. **Trained Evaluator** - ESL instructor at the Ohio State University with either current or recent experience teaching and evaluating ESL composition who has demonstrated satisfactory performance in a calibration session designed to establish interrater reliability;

6. **Calibration Session** - practice evaluation session in which subjects familiarize themselves with model compositions that reflect the range of proficiency with which they will be dealing. Results of this session are also used to compute interrater reliability;

7. **ESL Compositions** - writing samples provided by foreign graduate students during the first week of classes of the Autumn, 1986 term at the Ohio State University. The writing sample was a timed exercise (48 minutes) in which students were asked to respond to a predetermined question (see Appendix A). Also known as a Diagnostic, this writing task is administered to students placed into the ESL Composition course sequence for the purpose of checking the accuracy of the original placement assessment;

8. **Protocol Scoring Matrix (PSM)** - an adaptation of the Meyer (1975) text analysis instrument used to measure the extent to which the content included in a written recall protocol matches the content present in the
writing sample from which the protocol was derived. The PSM measures idea units on a weighted scale of one to seven according to their relative importance to the meaning of the text. Scores are reported on a percentage basis (see Appendix B);

9. Idea Unit - the most important idea in each sentence, plus other ideas that describe or give information about the main idea; also termed a "proposition"; determined subjectively, on the basis of whether the chunk of discourse under consideration is "felt to be a single, meaningful piece of information conveyed by the passage, whether it consist(s) of a word, a definition, or a phrase in the passage" (Meyer & McConkie, 1972, p.13);

10. Naïve Readers - subjects asked to read ESL compositions for the purpose of holistically evaluating them, but not told beforehand that they would be asked to provide recall protocols of texts they had just read;

11. Focused Readers - subjects who had just completed recall protocols as Naïve readers, and were told that the next compositions they would read and holistically rate would also serve as the basis for a second recall protocol task;

12. Lower Quality Composition - ESL writing sample chosen from a corpus of Level 107 Diagnostics to serve as one of the two target texts for which subjects would
write recall protocols. Lower quality composition is assumed to reflect the output of a less proficient NNS author;

13. Higher Quality Composition - ESL writing sample chosen from a corpus of Level 108 Diagnostics to serve as the target text assumed to have been produced by the more proficient NNS author.

Limitations

Limitations of the study included the following:

1. subjects were asked to rate compositions drawn from Levels 107 and 108 only as representative of NNS authors of relatively Lower or Higher proficiency. While it is quite possible that raters would utilize different reading strategies to rate texts drawn from more homogeneous or more heterogeneous samples, the researcher chose to limit the scope of the investigation to samples drawn from the mid range of ESL writing proficiency;

2. proficiency was measured on a four-point scale. Although a more detailed scale could have been used, it was decided that subjects could be conditioned to a more realistic rating environment if a four-point scale was employed;

3. use of a predetermined topic restricted ESL students' freedom of response in their compositions;
4. the Diagnostic writing exercise was produced under time constraints;

5. twelve holistic raters took part in the study; hence, a relatively limited number of written recall protocols were generated;

6. length of sample compositions used to calibrate raters and establish interrater reliability was noted but not controlled for;

7. the PSM measured comprehension in terms of "idea units correctly recalled" only. No adjustments were made in scoring totals for incorrectly identified or omitted idea units in rater-generated protocols (Murphy & Puff, 1982). In addition, whether or not the order in which they were presented in the sample compositions was not a consideration in the scoring of the protocols.

**Research Question**

This research study sought to answer the following question: What is the role of the reader in the holistic assessment of second language writing proficiency? From this general question, two specific questions were derived:

1. Will holistic raters of ESL writing proficiency recall compositions generated by more proficient non-native speaking (NNS) authors better than they recall compositions written by less proficient NNS authors?
2. Will reader orientation (task awareness) have an impact upon the amount of content recalled by holistic raters of ESL compositions?
CHAPTER 2
LITERATURE REVIEW

Overview

Four premises are central to the execution of this study. First, holistic evaluation by L1 raters is used to provide a measure of L2 writing proficiency. That is to say, a holistic rater's assessment of the quality of a writing sample serves as an assessment of the demonstrated writing proficiency of that sample's author. Second, writing proficiency, whether in L1 or L2, must include a comprehension component. Put another way, any measurement of a writer's proficiency must include an indication of how well his/her message is understood by the reader. Third, a recall protocol can provide a quantifiable measurement of how well an L1 holistic rater comprehends a text generated by an L2 author. Fourth, prior knowledge of such a recall protocol task may influence how a holistic rater approaches a holistic evaluation task and, by extension, play a significant role in determining the amount of content recalled in a written protocol.
The assumption that holistic scoring is a reliable and valid means of assessing L2 writing proficiency has not been directly tested in terms of comprehension. Accordingly, the following literature review will focus upon five areas of research that have a direct impact on holistic assessment of L2 writing proficiency:

1. Writing Research
2. Writing Proficiency
3. Holistic Scoring
4. Recall Protocols
5. Task Awareness

The first part of the review provides an overview of research in second language writing proficiency. In this section, it is demonstrated that a large part of the methodology, assumptions, and findings incorporated into the second language research base originate in first language research. Next, the emphasis in current L2 writing research upon process models of composing is detailed. This section concludes with a discussion of the shortcomings of L2 writing research in general, and L2 writing assessment research in particular.

In the second part of the review, the construct of writing proficiency is explored, with the purpose of developing an operational definition that includes the element of reader comprehension. Moving from a discussion of the multifaceted nature of the construct, this section
surveys the various components cited by researchers in both L1 and L2 as being elements of writing proficiency and discusses the shortcomings of those elements in providing a comprehensive definition of writing proficiency. This section concludes with an argument for the adoption of a definition of writing proficiency that acknowledges the salient importance of meaning in the assessment of writing proficiency.

Once a definition of writing proficiency is operationalized and the underlying assumptions of holistic scoring are presented, it becomes apparent that the most effective means by which to measure the degree to which a reader understands what he/she had read while engaged in a holistic evaluation task is to use a written recall protocol. Accordingly, the third section defines and discusses what recall protocols are, how they work, what they measure, and how they are scored. Particular attention is paid to research findings concerning the use of recall protocols to measure organization of content in memory. Such findings are especially relevant in a study whose major premise is that the better an L2 text is written, the better it will be comprehended. The final portion of this section discusses various procedures used in scoring recall protocols. In it, a rationale for adoption of the Meyer (1975) system for use in this study is provided.
The next section of the literature review offers a discussion of holistic scoring and its uses in evaluating writing proficiency. In this section, the researcher reviews the theoretical foundations upon which holistic scoring is based, focusing on empirical findings pertaining to the reliability and validity of holistic scoring as a means of assessing writing proficiency. This section concludes with a review of research findings that challenge the usefulness of holistic evaluation as a procedure to assess a meaning-based construct of writing proficiency. In particular, the influence of time constraints on reading comprehension is highlighted, thereby calling into question the ability of holistic raters to attend to meaning.

In the final section of the literature review, the influence of task awareness on depth of processing of a written text is explored. Despite recent evidence to the contrary (e.g., Lee, 1986a), it is generally assumed that "focused" (i.e., forewarned) subjects will attend more to meaning than their naive counterparts when advised that they will be tested on the contents of the text they are about to read. Consequently, any study seeking to investigate what is comprehended by the reader while engaged in a holistic evaluation task must account for potential differences between holistic reading in the naive mode and holistic reading in the focused mode.
Research in Second Language (L2) Writing Proficiency

Generalizations from First Language (L1) Research

With few exceptions, the research base concerning the evaluation of L2 writing proficiency is nonexistent. Instead, decisions concerning writing instruction in general, and evaluation in particular, seem to be based primarily "on rumor, fashion, or tradition" (Krashen, 1984, p.1). Moreover, much of what is generally accepted in L2 writing research is an extension of research in L1, as can be seen in the title of one of the seminal works in L2 writing research, Zamel's (1976) Teaching composition in the ESL classroom: What we can learn from research in the teaching of English. Indeed, a review of the literature reveals that virtually all research into both the production and evaluation of L2 writing has its origins in research based on native language writers.

L1/L2 Research

Krashen (1984) has made one of the few comprehensive attempts to enumerate and compare L1 and L2 writing research in a systematic way. From Krashen, we learn of the similarities between the composing processes of L1 and L2 writers.

Studies contrasting the composing processes of skilled and unskilled high school and college students make up the bulk of the L1 research of Emig (1971),
Sommers (1980), Flower and Hayes (1981), and Perl (1979, 1980). In L2, a great deal of research involving skilled and unskilled ESL college students has been done by Zamel (1982, 1983, 1985) and Raimes (1985a, 1985b).

It is true that some differences between L1 and L2 have been noted. Raimes (1985a, 1985b, 1986), for example, points out that skilled L2 writers frequently utilize so-called "unskilled" composing strategies. Scarcella (1984) demonstrates the relative lack of skill shown by L2 writers in using effective means by which to orient the reader when composing expository prose. Despite the existence of a few studies like those just cited, Krashen (1984) concludes that L2 research results "have been remarkably similar to those found in first language writing"(p.40).

Another area in which L2 results have been in general agreement with those of L1 is the usefulness of feedback during the writing process. Both L1 researchers, such as Beach (1979), and L2 researchers, including McKay (1983), Zamel (1985, 1986), and Raimes (1983, 1986), concur in the position that teacher feedback between drafts has generally positive effects on writing quality.

Krashen (1984) also notes the link in L1 between amount of pleasure reading and writing quality that has been established by the research of Applebee (1978), Ryan (1977), McNeil (1976) and Thorndike (1973). Although Krashen confesses to being unaware of comparable studies
in L2, findings reported by Elley and Mangubhai (1983) and Janopoulou (1986) have given strong indications that a similar relationship holds for pleasure reading and writing in L2.

Shortcomings of L2 Research

While much attention has been paid recently to the L2 writing process, the focus of such inquiries as those conducted by Zamel (1976, 1982, 1983, 1984, 1985, 1986) and Raimes (1979, 1983, 1985a, 1985b, 1986) has been upon how the L2 writer generates text rather than upon how his/her audience processes the text. Thus, despite the growing consensus that reader response and reader/writer negotiation of meaning are crucial to the development of writing proficiency in a second language, L2 writing research is essentially writer-based. Such an emphasis, while reflecting a growing awareness of the salient role of the writing process, has been at the expense of evaluating the writing product, especially insofar as that product is used to measure a second language writer's proficiency.

There are some signs, however, that this trend is changing, as can be seen in the direction of studies such as the one done by Scarcella (1984), in which she sought to compare how L1 and L2 writers orient their audience in an expository essay. Indeed, such a staunch advocate of
the process model of ESL composition as Zamel (1986) has recently acknowledged the need for product assessment in addition to process development in the realm of second language writing.

Evaluation in L2 Writing

In keeping with the emphasis given to investigating the writing process in a second language, research into the evaluation of ESL compositions has, for the most part, concentrated on evaluation as a tool of instruction rather than assessment. Accordingly, discussion of evaluation procedures is primarily concerned with the issues of error analysis (e.g., Kroll & Schafer, 1978), error correction (e.g., Hendrickson, 1980), and responding to student writing (e.g., Zamel, 1985; Raimes, 1983, 1986). Such empirical research as exists is even more limited, focusing as it does on a particular method (i.e., Think-aloud Protocols) to explore the nuances of the composing/responding process (e.g., Zamel, 1982, 1983, 1985; Raimes, 1985a, 1985b). Thus, a situation exists wherein, as Krashen (1984) points out, "While reasonable suggestions have been made on the basis of first language research, empirical investigation is lacking"(p.38).
Holistic Evaluation of L2 Writing Proficiency: The Untested Premise

Nowhere is the lack of empirical investigation more apparent than in the realm of holistic evaluation of L2 writing proficiency by L1 raters. Holistic evaluation of writing is a common tool by which proficiency is measured. However, while the procedure enjoys widespread acceptance as a reliable means by which to measure L2 writing ability (Perkins, 1983; Cooper, 1977), it is not at all clear that what it measures is what it purports to measure; namely, writing proficiency.

It is the contention of the researcher that the issue of comprehension must be addressed in order to test the assumption that holistic evaluation offers a valid measurement of L2 writing proficiency. This contention derives from the assumption that any such measurement of a writer's proficiency must be based, at least in part, on a reader's ability to comprehend what has been written. The remainder of this literature review will provide a framework within which such a premise can be tested.

Writing Proficiency

Elements of Proficiency

Identifying elements of writing proficiency is a task that must be approached with caution. What constitutes "good writing" is an issue complicated on the one hand by
the interactive and subjective nature of the reader-writer relationship (de Beaugrande, 1978; Carrell, 1984a; Pearson & Tierney, 1984), and on the other by the fact that we seek to measure writing quality across a wide spectrum of options, from skills such as persuasiveness (Stiggins, 1981), to rhetorical relationships like causation or description (Meyer, 1982), to amount of cohesive density (Witte & Faigley, 1981). What, then, is writing proficiency, and how can it be measured on a consistently accurate basis?

That writing proficiency is a multifaceted construct has been acknowledged repeatedly. Odell (1977a), for example, notes the variety of dimensions within the rubric of "Writing Proficiency." Spandel and Stiggins (1980) refer to the ephemeral nature of the construct, pointing out that owing to "the broad range of potentially relevant writing competencies and the difficulties in setting standards of acceptable performance." the characteristics of "good writing" remain elusive (pp.5 & 6). Cronbach (1973) reminds us that, as with any construct, writing proficiency cannot be defined in terms of what he describes as "specific responses-to-situations." Rather, the definition must be open-ended and susceptible to "surplus meaning."
The Writer Variable

Research into how raters respond to prose reveals yet another dimension of the assessment of writing proficiency. Braddock et al. (1963) conducted a metaanalysis of 485 studies dealing with research in written composition. The results of their analysis pointed out the need for English teachers to make a distinction between a student's writing ability (or competence) and his/her actual performance on the writing task. That is, in any attempt to evaluate writing proficiency, the teacher must realize that "If a student's writing is consistently low, one may say that he has demonstrated poor ability, but often one cannot say positively that he has poor ability" (p.6). Moreover, Braddock et al. cited Kinkaid's (1953) study of college students and Anderson's (1960) examination of eighth grade students in support of the assertion that writing performance can fluctuate widely from day to day and topic to topic. From this, Braddock et al. concluded that such a "writer variable" must be taken into account in any assessment of writing skill, noting that "Although it is obvious the writer variable cannot be controlled, certainly allowances should be made for it" (1963, p.6).

The Reader Variable

Another variable in rating compositions discussed by Braddock et al. (1963) was the reader variable, which they
characterized as "the tendency of a rater to vary his own standards of evaluation" (p. 10). One of the specific manifestations of this variable, which Braddock et al. termed "personal feelings" (p. 10), is the halo effect. They caution against raters falling victim to a mindset that "suffuses their general ratings" (p. 14), citing as evidence references to a "grammatical halo effect" in the research of Starring (1952), which employed pre-trained raters using an analytical scoring method, and Diederich et al. (1961), which used untrained holistic raters. Braddock et al. (1963) conclude by stating that the effectiveness of assessment procedures is dependent upon "the degree of commitment which each rater feels toward the criteria being employed" (p. 15).

Proficiency and Processing

More recently researchers have demonstrated that another dimension of this "reader variable" resides in how the reader approaches his task. Research by both Just and Carpenter (1980) and Kintsch and van Dijk (1978), for example, has emphasized the importance of the reader's goals in determining how visual input is processed.

In a study involving 14 university undergraduates reading 15 scientific texts, Just and Carpenter (1980) sought to validate their model of reading comprehension. The subjects were instructed to recall each passage
immediately after reading it. They were further instructed to read each text as they would normally read, without trying to memorize the contents. Data concerning allocation of eye fixations and gaze durations were also collected, by means of eye-tracking equipment. These data were then interpreted in light of the model posited by Just and Carpenter in order to determine if the model provided a satisfactory fit with the data.

In their description of the features of their model of reading comprehension, Just and Carpenter listed Get Next Input, Interclause Integration, and Sentence Wrap-up as stages of the reading process that are directly affected by reader goals. In the Get Next Input stage, information is located, encoded, and processed. The point in the cycle at which the eye is directed to a new place in the text is determined by a list of conditions, including some that can reflect specific reading goals. According to Just and Carpenter (1980), "By setting the conditions properly, the reader can adjust his processes to the situation at hand" (pp. 336-337), whether it be skimming or reading to memorize.

In the Interclause Integration cycle, the reader relates clauses and sentences to each other in order to discover coherence in a text. In this phase, the importance of the reader's goals in processing the text are apparent because, as Just and Carpenter note, "Text
roles that are usually more important to the text and to the reader's goals... are integrated differently than less important units," while, "By contrast, details are often less important to the reader's goals and to the text"(p.344) and, thus, are assigned a lower position in the organizational hierarchy of the text structure.

During the Sentence Wrap-up phase of reading comprehension, the reader searches for unassigned referents, constructs interclause relationships with the aid of inferences, and attempts to resolve any inconsistencies that remain at the sentence, clause, or paragraph level. According to Just and Carpenter:

The decision of when and if to do a wrap-up may be controlled by the depth of processing. For example, skimming may require wrap-up only at paragraph terminations, whereas understanding a legal contract may require wrap-up at clause boundaries(p.346).

Thus, reader goals are posited to play a significant role at this stage of the reading comprehension process as well.

Data was collected from both eye-tracking and scoring of recall protocols. Analysis of both sources of data revealed that the Just and Carpenter model did, indeed, fit the data, especially with respect to the above-mentioned three features of the processing model.

Employing a procedure wherein 63 university undergraduates were asked to produce both immediate and
delayed (one-month and three-month interval) written recall protocols and summaries of a one paragraph text entitled "Bumperstickers and the Cops," Kintsch and van Dijk (1978) found empirical evidence in support of a model of text comprehension in which the use of macroprocesses are made explicit by appropriate control schemata.

In their model of text comprehension, Kintsch and van Dijk (1978) link the instantiation of appropriate schemata with the reader's goals. According to Kintsch and van Dijk, each schemata, which is a formal representation of the macro-operators (mechanisms used to delete or generalize all redundant and/or irrelevant propositions), "determines which micropropositions or generalizations are relevant and, thus, which parts of the text will form the gist" (p. 373). They concluded that in the absence of such a controlling schema, whether due to vague reader goals or unconventional text structure, the outcome of the comprehension process becomes haphazard and indeterminate. On the other hand, when the reading task is approached by subjects "who read a text with a specific problem-solving set" (p. 373), the special purpose "overrides whatever text structure there is" (p. 373), even if the text is not conventionally structured.

The conclusions contained in both the Kintsch and van Dijk (1978) and Carpenter and Just (1980) reports underscore the element of subjectivity in evaluation of
prose material. In particular, both studies provide empirical support for an assumption that is critical to this dissertation; namely, that the reader's goals are a crucial determinant in the reading process.

Another aspect of the interactive relationship between reader and text has been explored by Baker and Anderson (1982). In their study, Baker and Anderson investigated the effects of inconsistent information on text processing. In particular, Baker and Anderson pursued the issue of whether or not readers engaged in a focused reading task monitor their own comprehension in order to determine if ideas expressed in a text are consistent. They concluded that while subjects did, indeed, engage in such self-monitoring when reading logically inconsistent texts, the subjects could not distinguish well between main ideas and details. The results obtained by Baker and Anderson also indicated that specific instructions to read passages carefully in preparation for subsequent questions did not seem to influence how the subjects approached the reading task.

Unfortunately, the fact that Baker and Anderson used a multiple choice testing instrument to measure reading comprehension makes drawing firm conclusions difficult, owing to the drawbacks inherent in the multiple choice format noted by Oller (1979). Nevertheless, the results of this study add more weight to the argument that the
organization of a text affects both how quickly one can read it and to what extent one can comprehend it.

Definitions of Proficiency

To date, a universally accepted definition of the elements contained in "good writing" has eluded scholars, teachers, and researchers, alike. Indeed, attempts to isolate specific components of writing proficiency have ranged from definitions that trace their origins from Classical rhetoric to the ACTFL Proficiency Guidelines.

Some recent researchers have chosen to define writing proficiency in such general terms as one global phrase, "Writing Effectiveness" (Odell, 1977), or two criteria, "Does it make sense and is it interesting?" (Cooper & Purvis, 1973). Others view the writing process as a pragmatic task which should be judged on the basis of communicative effect (Oller, 1979). Such pronouncements, however, are for the most part lacking in either empirical justification or specific means by which to operationalize the construct of writing proficiency.

The view of writing as the production of a text that serves the function of communicative interaction is also popular (de Beaugrande, 1978; Moffett, 1983; Carrell, 1984a). Pearson and Tierney (1984) provide a detailed discussion of how writing functions as communicative interaction. They perceive the writing process as one
"of negotiating between a reader and an author through the medium of a text" (p.145). They point out that because "texts are written by authors with the intention that readers will create meaning" (p.145), "... every reader must, at every instant during reading, satisfy herself about what the current text she has composed (during the process of reading) means" (p.150). In his definition of writing as a "purposeful activity involving language use in special modes" (p.10), deBeaugrande (1977), too, stresses the importance of interaction between reader, writer, and real world, in both the production and comprehension of writing.

In line with this viewpoint is the position taken by Spandel and Stiggins (1980) that proficiency is best measured on a continuum whereupon the narrower the gap between message as sent and message as received, the more proficient the writer. It is also reasonable to assume that a proficiently written text allows for a maximum of decoding of the message with a minimum of effort at overcoming interference in its transmission. Such an assumption is crucial to the position that meaning is an integral component of writing proficiency.

Components of L1 Writing Proficiency

In the past few years, several L1 researchers, including Coffman, Diederich et al., and Witte and Faigley,
have offered detailed descriptions of the components of writing proficiency. In his study, Coffman (1968), for example, listed as aspects of writing proficiency ideas of content, organization, style, mechanics, and creativity. An earlier study by Diederich et al. (1961) investigated factors involved in judgments of writing proficiency by means of a factor analysis of comments made by expert raters engaged in direct assessment of writing samples submitted by 300 liberal arts college freshmen. On the basis of this analysis, they concluded that expert raters measure writing ability primarily in terms of ideas (e.g., relevance, clarity, development, and persuasiveness), form (e.g., organization and analysis), and flavor (e.g., style, interest, and sincerity). Other categories of rater response mentioned by subjects included skill in effecting smooth transitions, ability to paraphrase, and variety of sentence structure.

Both the Coffman (1968) and Diederich et al. (1961) descriptions of the components of writing proficiency are extremely global in nature, to the extent that the factors they enumerate are, in and of themselves, general enough to cover everything but too general to explain anything. What is needed to establish measurable criteria is something more specific.

Witte and Faigley (1981) used a much more specific and explicit system to investigate the nature of writing
quality. In their study, analysis of holistically-rated compositions written by 90 American college freshmen was performed by means of a general four-point index of cohesive density based on the work of Halliday and Hasan (1976). Results indicated that writers of high-rated essays were more proficient than writers of low-rated essays at expansion and connection of ideas.

However, while offering a much more explicit treatment of the components of writing proficiency, the system upon which Witte and Faigley (1981) base their analysis is not without its critics. Specifically, their concentration upon cohesive density seems to deny the interactive and communicative nature of writing by employing a system that is based on the assumption that meaning resides at the text level.

Spiro (1980), for example, criticizes the acceptance of the Chomskyan notion that meaning is text-based. He points out that a "bottom-up" view of processing that focuses on the individual sentence cannot provide a complete semantic description of what is contained in the sentence. Moreover, he holds that concepts of cohesion, while possessed of properties that may be amenable to decomposition and analytical examination, are nevertheless concepts that are basically experiential and gestalt-like. In other words, to Spiro (1980), "Meaning does not reside in words, sentences, or even entire passages considered in
isolation" (p. 245). Instead, Spiro views text as "a skeleton, a blueprint for the creation of meaning" that cannot be analyzed at any discrete level without resulting in "an incomplete understanding of that level's meaning in use" (p. 245).

Carrell (1982) offers perhaps the most detailed critique of the sort of system used by Witte and Faigley (1981) in a discussion of the inadequacies inherent in such text-analytic procedures as they are applied to second language research. Using as the basis of her argument empirical evidence supporting a schema-based model of text processing that sees reading as an interactive process between the reader and the text, she demonstrates through a survey of pertinent studies the fallacy of using cohesion as a measure of the coherence of a text. In reference to researchers who seek to use cohesion to measure coherence in a text, Carrell concludes:

... these authors appear to be making the assumptions Halliday and Hasan made. They seem to be assuming that the features of a text which, like cohesion, may be subjected to textual analysis techniques, determine a text's coherence. They seem to believe that coherence is located in the text and can be defined as the result of specific textual features... Cohesion is not the cause of coherence; if anything, it's the effect of coherence (p. 286).

In addition to the shortcomings already detailed, what all of the above-mentioned definitions of proficiency lack is explicit acknowledgement of the role of meaning in the
reader-writer contract. If the author of a piece of written discourse is to be judged as proficient, there should be a reasonably close correspondence between meaning as intended by the writer and meaning as understood by the reader. Yet the researchers just cited focus on the parts at the expense of the whole. Arguing that writing proficiency involves ideas, form, and flavor, or content, organization, style, mechanics, and creativity, or cohesive density without mentioning meaning confuses the end with the means.

Just how important meaning is to judgments of writing proficiency can be seen in the results reported by Johns (1985) of a study designed to identify the types of writing tasks assigned and methods of evaluation employed by instructors in two-year and four-year educational institutions. In her study Johns polled faculty at both San Diego State University and a nearby junior college to determine what teachers thought constituted "good writing" in their respective disciplines. Specific responses were solicited concerning writing genres required, cognitive demands imposed, and factors considered important in evaluating student writing. Results indicated that a vast majority of respondents (86% of the two-year college faculty and 92% of university faculty) listed the ability to convey meaning (ideas) as the single most important feature they were looking for in student writing.
Components of Writing Proficiency: An L2 Perspective

Although most traditional discussions of the assessment of L2 writing proficiency omit explicit reference to meaning, some researchers have pointed out that attention to ideas is an integral part of effective writing. Taylor (1986), for example, lists among the characteristics of a process-oriented ESL writing class the central role of content and ideas. Raimes (1983) is another who defines the proficiency of an ESL writer in terms of his/her ability to communicate, noting that "Composing means expressing ideas, conveying ideas" (p. 83).

Despite the attempts of the above-cited researchers to develop a meaning-based definition of writing proficiency - at least with respect to instruction - most of what has been written about the assessment of L2 writing proficiency is based upon the assumption that the L1 model can be directly applied to an L2 context. Indeed, the only systematic attempt to identify the components of writing proficiency in a second language has been the ACTFL Proficiency Guidelines, which were first published in their provisional form in 1982. Without commenting on the degree to which L1 research can be applied to an L2 environment, it is nevertheless interesting to note that while the ACTFL Proficiency Guidelines for writing share many of the same features as those enumerated by L1 researchers, these Guidelines differ in that they contain explicit references
to such terms as meaning, understanding, and comprehension.

The ACTFL Proficiency Guidelines, however, are sufficiently vague that an instrument for the measurement of writing proficiency cannot be easily derived from them. Therefore, we must return to the realm of first language research to find a definition that is both quantifiable and applicable to L2. The basis of such a definition can be found in Greeno's (1977) criteria of "good comprehension."

A Meaning-based Definition of L2 Writing Proficiency

Greeno's (1977) criteria of "good comprehension" allow us to derive a definition of writing proficiency that does not ignore the salient importance of meaning. These three criteria include (a) achievement of a coherent, connected representation; (b) correspondence between the object to be understood and its internal representation and; (c) connectedness between one's general knowledge and the concepts contained in the message. These criteria are consistent with what is generally accepted as being in the mainstream of opinion among researchers in both L1 and L2. Moreover, according to Voss et al. (1982), Greeno's criteria have two added advantages, in that (1) comprehension is defined as readability, or the ease with which prose material may be
read and understood, and (2) it is possible to measure readability "by assessing how well subjects perform tests based upon the contents of a passage"(p.363). Such a definition of writing proficiency gives us a powerful tool for measuring writing as communication.

Holistic Scoring

Rationale for Use in L2 at the University Level

Once the construct of writing proficiency is defined and its components identified and operationalized, a composition may be assessed by means of one of three scoring techniques: holistic, analytical, or primary trait (Perkins, 1983). Holistic scoring, by virtue of the fact that it offers a means of direct assessment of actual composing skill, is the method of choice among practitioners if the goal is to measure "less tangible writing skills... for which the concept of right and wrong is less relevant"(Stiggins, P.3). Consequently, holistic scoring is in widespread use, not only for placement and grade-level evaluation purposes in university ESL composition programs, but also as a scoring procedure for the newly-developed TOEFL Writing Test (1986).

History

Contrary to popular belief, holistic scoring of writing proficiency is not a new arrival on the assessment
scene. Indeed, as early as 1912, Ayres developed a handwriting scale that was a precursor of modern-day holistic scoring methods that employ calibration samples (Cronbach, 1971). That holistic scoring offers a viable alternative to analytic scoring methods was noted by Wiseman in 1949. In one of the first of many anecdotal observations by practitioners in writing evaluation, Wiseman reported the experience of using an analytical scoring method and consistently having what was clearly the "best" paper turn up some place other than at the top of the pile. He conjectured that this was so because analytical scoring methods are not sufficiently sensitive to the nuances and subtle interactions that allow a piece of written discourse to be acknowledged as "good" (Cronbach, 1971).

Definition

Holistic scoring can be defined in generic terms as assessment of the overall quality of a composition (Davis, Scriven, & Thomas, 1981). This assessment is accomplished by means of a rank ordering of subjects on a continuum of writing skills rather than according to a predetermined set of criteria for, as Spandel and Stiggins (1980) note, "... it is important to realize that in holistic scoring, there is no preconceived notion of the 'ideal' paper" (p.20).
Specific recommendations for the holistic scoring procedure that are widely accepted include those of Godschalk et al. (1966), Cooper (1977), Braddock et al. (1963), and Jacobs et al. (1981). Both Godschalk et al. (1966) and Cooper (1977) recommend that raters read quickly to form general impressions. Braddock et al. (1963) suggest that during the evaluation process, "papers should be graded by at least two raters using a mutually-understood and accepted rating system with which they have previously practiced" (p. 45). Jacobs et al. (1981) agree, adding that raters should be selected from the same background, that they be thoroughly trained until they can achieve close agreement in their assessments on a consistent basis, and that they should be monitored periodically during evaluation sessions to assure that they are applying the criteria and standards of evaluation on a consistent basis.

Variations

While most varieties of holistic scoring share several common features, holistic scoring should be viewed as a family of techniques that are similar in general outline rather than as a method, per se. Spandel and Stiggins (1980) underscore this point when they explain that the purpose of direct assessment is "to gather samples of writing and to evaluate them according to prespecified
criteria" (p.7). They also point out that while holistic scoring is commonly used for placement, "because there are different purposes, the assessment strategies used to achieve them will differ" (p.12).

Perkins (1983) describes several variations in holistic evaluation procedure. For example, the rating scale may vary from as few as two points (e.g., better/worse or high/low) to ten or more. According to Spandel and Stiggins (1980), most systems employ an even-numbered scale of either four or eight levels, so as to prevent raters from opting for a mid-range score on essays for which assigning a score is difficult. Furthermore, the scale may incorporate widely differing criteria upon which holistic scores are based. Once the rater has read the composition to form a general impression, he/she assigns a mark to that composition based on a standard established by means of model compositions, the rater's own experience, a list of holistic criteria, or any combination of the above. Each composition is usually assessed by two independent raters, but the number can vary.

Reliability

It is a maxim of psychometrics that any instrument of assessment must be shown to possess reliability, or internal consistency, as well as validity if it is to be used with confidence (Oller, 1979). Results of research
into the reliability of holistic evaluation have ranged from positive (Follman & Anderson, 1967; Cooper, 1977) through inclusive (Braddock et al., 1963; Mullin, 1980) to negative (Diederich et al., 1961). With regard to the issue of interrater reliability, researchers such as Braddock et al. (1963) and Follman and Anderson (1967) have reported coefficients as high as .90. Mullin (1980), on the other hand, reported inconclusive results in a study which sought to investigate interrater reliability of evaluators of writing proficiency in ESL by using an analytical scoring procedure on a five-point scale.

On the other end of the spectrum, the classic Diederich et al. (1961) study demonstrated that raters coming from a variety of professional backgrounds, including English teachers, exhibited a marked lack of agreement in holistic evaluations of papers written by college freshmen. The Diederich et al. report, which used 60 raters and compositions written by 300 college students, noted that "out of the 300 essays graded, 101 received every grade from one to nine, 94% received either seven, eight, or nine different grades; and no essay received less than five different grades from 53 readers" (p. 5).

Even advocates of holistic scoring recognize some potential weaknesses in the approach. Braddock et al. (1963), in a generally positive treatment of holistic evaluation of writing proficiency, acknowledge the danger
of the halo effect as a threat to reliability in holistic scoring procedures. They also concede that the effectiveness of assessment procedures is dependent upon "the degree of commitment which each rater feels toward the criteria being employed" (p.15). Perkins (1983), while also generally in favor of holistic scoring, notes that such evaluation can be extremely subjective, and hence unreliable, owing to such factors as "bias, fatigue, internal lack of consistency, previous knowledge of the student, and/or shifting standards from one paper to the next" (p.653).

Both Cronbach (1971) and Coffman (1968, 1971a, 1971b) have pointed out the difficulty in achieving consistency in grading essay responses. Cronbach (1971) cites Coffman's (1968) objections to essay exams, all of which center on the issues of reliability and subjectivity, as telling drawbacks of holistic evaluation. According to Cronbach, advocates of holistic scoring maintain that the primary advantage of essay examinations is that they provide a direct measure of educational achievement. Coffman (1971a), on the other hand, disagrees. He states in response to such claims that "There is extensive evidence that it is difficult to achieve consistency in grading essay responses" (p.277).

Granted, different raters may tend to assign different grades to the same papers (Wiseman, 1949);
Coffman, 1968; Diederich et al., 1961; Perkins, 1983). And differences do show a tendency to increase as the essay topics become more open-ended (Coffman, 1968, 1971a, 1971b; Cronbach, 1971). But while these drawbacks pose a potential threat to reliability and consistency of measurement, it should be stressed that the problem is far from impossible to solve. As is pointed out by Perkins (1983), careful application of the basic rules of preparation for holistic evaluation can keep potential reliability problems to a minimum. In sum, if safeguards such as those advocated by Henning (1984) and Jacobs et al. (1981) are taken throughout the evaluation process, holistic scoring can provide an extremely reliable means of direct assessment of attained writing proficiency.

Validity

Introduction

While most research has centered on questions of reliability, there are some indications that lack of consistency is not the primary issue. Rather, an argument can be made for the position that the problem lies in determining what it is we are measuring on a consistent basis, even when consistency among raters is achieved. But while the validity of holistic evaluation of writing
proficiency may be difficult to verify, such verification should by no means be thought of as impossible.

Dimensions of Validity

Researchers have spoken of dimensions of validity that can be applied to the direct assessment of writing proficiency, noting that many of the problems involved in establishing the validity of direct assessment methods arise as a result of the confusion between the major uses of such tests (Cronbach, 1971; Spandel & Stiggins, 1980). Those uses may include decision-making in such areas as placement, special instruction, grading, or graduation (Cooper, 1977).

Direct methods used in placement are what Cronbach (1971) calls decision-making tests, and need to be based on criteria of performance, while descriptive instruments rely more on the assumption of content or construct validity for their use. Confusion between purposes, according to Cronbach, results in the blurring of the distinction between the two.

Cronbach further asserts that a construct cannot be defined in terms of specific "responses-to-situations;" rather, it should be measured with an instrument that is open-ended and susceptible to what he terms "surplus meaning." Hence, to measure the construct validity of
holistic scoring of writing proficiency, the instrument should be descriptive in nature, or at least possessed of a demonstrated concurrent validity with a descriptive instrument. Holistic schemes of the sort typically used by ESL composition programs in American universities for placement of students according to proficiency levels, while based on criteria of performance, are sufficiently open-ended to accommodate surplus meaning. This is especially true if the criteria focus more on features which emphasize the communicative over the prescriptive (Perkins, 1983).

Claims for the Validity of Holistic Assessment of Writing Proficiency

As was noted above, claims for validity have proved to be even more difficult to verify than claims for reliability, owing in part to the ephemeral nature of the construct of proficiency (see Literature Review, Section III above for details). Despite the elusiveness of the construct, however, claims for the validity of holistic writing proficiency have been advanced by Kaczmarek (1980) and Flahive and Snow (1980), whose claims for concurrent validity are supported by their research, as well as by Cooper (1977) and Perkins (1983), whose claims for construct validity are less well supported by research.
Kaczmarek (1980) examined both teacher ratings of proficiency and trained rater evaluations of comprehension in a study designed to investigate the concurrent validity of subjective and objective measures of writing skill. She concluded teacher and rater measures of evaluating essays correlated with each other at the .67 level. Moreover, these two subjective evaluation methods were both strongly correlated with other measures of ESL writing proficiency having independent claims to validity.

Flahive and Snow (1980) also found evidence of concurrent validity in a study involving analysis of writing samples supplied by ESL students at the university level. In their study, they reported significant correlations between objective measures of syntactic complexity and holistic evaluation of ESL compositions.

Cooper (1977), in offering a concise summary of holistic evaluation, asserts that "Where there is commitment and time to do the work required to achieve reliability of judgment, holistic evaluation remains the most valid and direct means of rank-ordering students by writing ability"(p.3). Finally, Perkins (1983), in a review of composition scoring techniques in ESL, concludes that holistic scoring possesses a higher degree of construct validity than any other composition evaluation scheme in use today "when overall attained writing proficiency is the construct to be assessed"(p.652).
Some Unanswered Questions vis-a-vis the Validity of Holistic Evaluation

In a general discussion of the characteristics of language tests, Oller (1979) points out that while any instrument of educational evaluation must be reliable before it can be valid, reliability alone is no guarantee of validity. In other words, a test instrument may be reliable in the sense that it is accurately and consistently measuring something, but that "something" being measured is not necessarily the construct it is purported to measure.

Granted holistic evaluation of writing proficiency can be made to yield quite reliable results; nevertheless, whether the procedure yields a valid assessment of writing proficiency is open to question. Assertions by Cooper (1977) and Perkins (1983) notwithstanding, one must still ask if a procedure that emphasizes a speedy and generalized perusal of a composition written by an unfamiliar non-native speaking author on a possibly unfamiliar topic can in fact render an accurate assessment of writing proficiency based on the criteria set down by Greenco (1977). This question is especially pertinent in light of a growing conviction on the part of researchers and practitioners, alike, that "If readers are to develop control over the models of meaning they build, they must
approach a text with the same deliberateness, time, and reflection that a good author employs as he revises his text" (Pearson & Tierney, 1984, p.151).

Holistic Scoring and Comprehension: The Case Against

If comprehension is a prerequisite to the assessment of writing proficiency, then the reader must attend to meaning. Yet holistic evaluation of the writing proficiency of non-native authors is enormously complicated by the fact that most texts produced by L2 writers contain errors of such type and magnitude that meaning is, at least to some extent, obscured. While there are no firm answers to the question posed above, research in the related fields of cognitive psychology and information processing offer revealing insights into the issue. For example, questions involving the impact of time constraints upon reading comprehension have been addressed by L1 researchers seeking to test assumptions made by hypotheses concerning text processing and the representation of meaning in memory (e.g., Meyer, 1982; Kintsch et al., 1977; Baker & Anderson, 1982).

In an L1 study seeking to find evidence of the use of one of five basic writing plans by unskilled writers, Meyer (1982) required subjects to read essays written by poor writers and then generate immediate recall protocols.
She found that if readers were able to identify the author's plan in a target composition, they could generate reasonably complete recall protocols. When unable to identify the author's plan, however, readers produced recalls that were inferior in both quality and quantity. From her findings, Meyer concluded that while skilled readers make heavy use of top-level content in processing a text, "unskilled writers may depart from this normal order and thereby obscure their content hierarchies" (p.44).

In a similar study by Kintsch et al. (1977), readers' ability to comprehend paragraphs arranged in normal sequences was compared to their ability to comprehend scrambled paragraphs. The researchers reported that while summaries of both versions of the target paragraphs were much the same if subjects were allowed unlimited time to read and decode them, under severe time constraints readers no longer had adequate time to unscramble the altered passages. Without sufficient time to unscramble the content of these altered passages, subjects consequently produced inferior summaries.

A related study by Baker and Anderson (1982) sought to investigate the effects of inconsistent information on text processing at the paragraph level. The researchers found that subjects spent a longer amount of time processing sentences containing inconsistent information. In
addition, the researchers concluded that most of their subjects had engaged in comprehension monitoring while reading, although most could not distinguish well between main ideas and details in a follow-up multiple choice test.

Application to Assessment of L2 Texts

While these studies involved texts generated and processed by native speakers, the conclusions drawn within them are nevertheless germane to any study of how a reader processes a text written in his/her native language, regardless of whether the author of that text is writing in his/her L1 or L2. This is so for two reasons. First, these studies point out the importance of allowing sufficient time for the reader to process unfamiliar or ambiguous text. Second, they also strongly imply that the more difficult the text is to process, or, to borrow from Meyer (1982), the more obscured the content hierarchy, the more time-consuming the process of comprehension becomes.

Recall Protocols

Definition, Applications, and Procedures

Recall protocols are a powerful tool in the study of prose comprehension. The range of applications of recall
protocols includes analysis of text structure, organization of information in memory, and study of how text is processed from one task situation to another (Voss et al., 1982). Of particular relevance to the present study is the frequent and effective use of recall protocols to measure comprehension in both L1 and L2 (see below for further discussion).

Most recall studies have employed the subtype of recall protocol generally termed unit recall, in which "The subject... is asked to remember the contents of a prose passage without benefit of any external aids" (Voss et al., 1982, p. 374). In this procedure, recall is measured in terms of the number of "idea units" contained in the subject's unaided retrieval that correspond to units of meaning contained in the original text. The purpose of such a protocol is to determine whether or not units of information judged by the researcher to be in the target text are present in the protocol. Modality of presentation of the text can vary according to the purpose of the research from oral to written, as can modality of recall, itself, although a written recall protocol has the advantage of allowing the subject to examine what he/she has written and make changes in his/her output (Voss et al., 1982).

The researcher may insist upon verbatim recall, although the preferred procedure is to allow paraphrasing,
so long as the original intent of the passage is preserved (Frederiksen, 1977; Mandler, 1978). Instructions given prior to a recall task generally direct that subjects attempt to use the ordering and, where possible, the exact wording contained in the target text (Johnson, 1970).

Widespread Use of the Procedure

Underlying Theory

The theoretical underpinnings of the use of recall protocols to measure comprehension can be found in Ausubel's assimilation theory, which links a concept's meaningfulness to the ease with which it is learned. Ausubel (1963,1978) holds that learning is potentially meaningful to the degree that concepts are relatable to the learner's cognitive structure on a nonarbitrary and substantive basis. Hence, it follows that meaningful material presented meaningfully (e.g., in a way in which it is relatable to prior learning) to a subject possessing a meaningful learning set (e.g., the optimum combination of willingness and readiness to learn) should be better comprehended than material of a more arbitrarily-organized nature. In terms of reading, this suggests that a properly motivated and experienced reader should have little trouble comprehending a text, providing it is sufficiently well written that meaning is not appreciably obscured.

Empirical Support

In a study designed to investigate the relationship between meaningfulness and the recall of textual material, Johnson (1973) derived a measurement of meaningfulness
based on Ausubel's (1963) assimilation theory. Utilizing both an immediate and a delayed recall instrument to measure comprehension, Johnson found a strong relationship between meaningfulness and recall of expository prose. Use of a stepwise multiple regression analysis confirmed the positive relationship between meaningfulness as it was operationally defined and recall of linguistic subunits in textual prose, thereby providing strong support to the contention that comprehension can be measured by means of recall protocols.

Content Structure as Presented and Perceived

Background

To date, research studies employing recall protocols have used texts generated by expert native writers (Connor, 1986). Because of the relatively transparent nature of those texts, questions of author intent have not been an issue. Instead, research has focused exclusively on the reader and how he/she processes the content structure of a coherent text generated by a proficient native speaker.

In holistic evaluation of L2 writing proficiency, however, the ability of the reader to access the author's meaning is by no means a foregone conclusion. Indeed, one needs only to read a few compositions written by non-native
speakers to appreciate the amount of effort involved in understanding what the author is trying to convey. Consequently, an understanding of the connection between comprehension and assessment of proficiency is crucial to any discussion of holistic evaluation of L2 writing proficiency.

Studies in First Language Composition and Recall

During the past twenty years, researchers in psychology, psycholinguistics, text linguistics, reading, and second language acquisition have attempted to generate formal models of text that describe in an explicit manner discourse-level prose structure and organization. In L1, such models include Rumelhart's (1975) generative story grammar, Mandler and Johnson's (1977) story grammar, and van Dijk and Kintsch's (1976) theory of macrostructures and microstructures. Discourse analysts with an interest in expository prose have, for the most part, focused on the analysis of the propositional content of sentences, with the most extensive attempts at validation concentrated in the areas of first language composition and recall (Kintsch, 1974; Frederiksen, 1975; Meyer, 1975; Mandler & Johnson, 1977).
Studies in Second Language Composition and Recall

Research into second language content organization and recall has borrowed heavily from first language research in terms of both methodology and application of models. While by no means as extensive as work done in L1, considerable research into how readers of a second language process text has been executed. Examples of such L2 research can be found in the studies of Carrell (1983, 1984a, 1984b, 1986), Connor (1984), Bernhardt (1983b, 1985, 1986a), and Lee (1986a, 1986b). This research, however, has concentrated on how a non-native reader organizes content within an "accomplished" text generated by a native speaker (Connor, 1986).

Implications for Holistic Evaluation of L2 Writing Proficiency

As was pointed out above, there have been no investigations into how an expert native speaking reader processes a text generated by a non-native speaking writer. The data that do exist, however, can be applied to such a context inasmuch as the hierarchical organization of a text affects both how and how much the reader comprehends. The work of Meyer, in particular, has offered important insights into both reader-based and text-
based variables that affect comprehension, especially in the realm of textual content structure.

Meyer (1980) characterizes content structure as "the text's overall organization and the interrelationships between its ideas and their relative importance" (p. 6). Kintsch (1974) also concerns himself with the content structure of a text, describing it as hierarchical in nature. The higher level organizational patterns posited by Meyer and Kintsch in their respective conceptualizations of content hierarchy are, in turn, analogous to the constructs theorized by still other researchers in the field. For example, the notion of schemata hypothesized by Anderson (1977) and Rumelhart and Ortony (1977), as well as the construct of macrostructures advanced by van Dijk and Kintsch (1978), equates to what Meyer et al. (1978) term top-level patterns of text organization.

In their 1978 study, Meyer et al. found that "high comprehenders" used authors' schemata in recall tasks, but "low comprehenders" did not. In a later paper, Meyer (1981) concluded that the comprehension process of a skilled reader with a desire to understand the author's logic will entail his forming a memorial representation of the text that parallels that of the text's content structure, and that recall will reflect such a parallel. On the basis of such findings, it is therefore reasonable to expect that if a judging proficiency involves
comprehension, it should follow that trained raters who are attending to meaning will recall more of the content of a text than trained raters who do not.

In another study concerning the identification of three variables in prose processing, Meyer (1973a) reported that "the organization in prose as represented by the content structure influences a reader's organization in memory of the passage's information, and thus, his subsequent recall" (p. 13). This finding lends support to the premise that better written papers are recalled better, and hence, provides a link between assessment of proficiency and comprehension.

Subsequent research into the "structure strategy" used by readers has tended to confirm the influence of content structure on comprehension (Mandler & Johnson, 1977; Meyer, 1974, 1980a; Thorndyke, 1977; Meyer & McConkie, 1973, 1974). In addition, the aforementioned studies by Kintsch and van Dijk (1978) and Just and Carpenter (1980) also tend to reinforce the notion that the better written a text, the easier it should be to comprehend, regardless of the level of processing at which it is read.

What these findings mean in terms of this study is that holistic raters of ESL placement compositions should recall the better organized essay better than the essay that is relatively lacking in organizational cohesion.
That is, assuming that holistic judgments of writing proficiency involve comprehension, the better rated composition should be recalled better, provided a true qualitative difference exists between the texts in question. Whether such a generalization does in fact hold true in practice is a central issue of this study.

**Scoring Procedures**

**Introduction and Purpose**

Once a recall protocol is generated, the question arises concerning how that protocol is to be scored. Several schemes have been devised, and while they share many common features, it must be understood that each was developed to probe a unique aspect of a reader's mental organization of the logical structure of a text.

In the following section, the feasibility of scoring recall protocols based on written texts generated by non-native speakers will be discussed. First, the range of scoring procedures that are available to the researcher employing recall protocols will be reviewed. Then, the advantages of using a system that employs a weighted propositional count to measure comprehension during a holistic scoring task will be explained.
Various procedures exist for analyzing the logical structure of a text. Among the better known are those of Kintsch (1974), Frederiksen (1975, 1977), and Meyer (1975). While similar in that analysis in each relies on the notion of propositions, each is distinctly different as well.

The Kintsch System

Kintsch's (1974) system seeks to describe a representation of text structure that is a hierarchy built on repetition of concepts. The basic assumption of the Kintsch model is that each text has a base which is different from its surface structure. In order to actualize meaning at the surface level, an ordered set of propositions must be devised so as to provide textual coherence. This coherence, according to Kintsch, is provided by argument repetition.

To use this system requires less training than for the other two systems. Scoring, however, is said to be more objective and thus limits the researcher's ability to describe relationships among propositions in a text (Meyer & Rice, 1984). According to Meyer and Rice (1984), the major problem with Kintsch's system is that emphasis upon superordinate ideas in the content structure reflects a concern more for explicit than inferred semantic relationships in text.
The Frederiksen System

Frederiksen's (1975, 1977) system was designed to study inferences. The model it was developed to test includes both a semantic structure and a logical structure. The former consists of sentence elements classified according to linguistic rules, while the latter is a non-hierarchical structure based on the logical relationships that exist between the concepts contained within the text. This system employs structure graphs, called networks, rather than hierarchical structures. It attempts to represent the semantic content of the text by means of set relations, which serve to connect two logically and semantically linked concepts into what Frederiksen calls a proposition.

The Meyer System

Meyer's (1975) system, like Frederiksen's, seeks to investigate the semantic relationships between concepts in a text. Meyer's (1875) system, however, goes beyond Frederiksen's in that it acknowledges the role of logical relationships as well. Semantic and logical considerations are combined into what Meyer calls rhetorical relationships. According to Connor (1984), "Meyer's analysis views the text as a set of interrelated simple
propositions that together form a complex proposition...
A proposition is a meaning unit which always consists of
a predicate (relation) and one or more arguments which
belong to the predicate (that is, concepts connected to
each other by the relation)"(pp.240-241).

The Meyer approach, while generally preferred among
researchers of prose structure and its effects on memory,
does require raters to receive considerable training in
identifying rhetorical relationships in text. In its
current form, the Meyer system yields a wealth of
information concerning text structure, and although it is
highly detailed, it is extremely complicated. On the
other hand, it has been demonstrated to provide a
superior measurement of recall when compared to the
systems of Kintsch and Frederiksen (Meyer & Rice, 1984).

Identification of Idea Units

Central to any system that seeks to analyze the
propositional content of prose is a means by which to
divide a text into its component idea units. Once the
text is so dissected, these units of meaning can be
weighted in terms of their relative importance to the
passage and scored against a recall protocol. Various
methods for scoring idea units in a text have been

In the system developed by Johnson (1970), scoring was done on the basis of one point for each idea unit or relation from the original content structure included in the recall protocol. In his study, Johnson used "pause acceptable" units to objectively divide a text into linguistic subunits. Johnson defined "pause acceptable" units as locations in the text where pausing might serve "to catch a breath, to give emphasis to a story, or to enhance meaning" (p. 13).

The system of Meyer and McConkie (1974), though differing from that of Johnson (1970) in that it identified linguistic subunits in terms of "idea units" consisting of "a single, meaningful piece of information conveyed by the passage, whether it consist of a word, a definition, or a phrase" (Meyer & McConkie, 1973, p. 110), was still closely related to Johnson's. Both systems provided a means by which insights could be gained into the effects of position of information on how a passage was recalled. Neither system, however, made allowances for the relative importance of idea units in deriving a comprehension score.

This drawback was remedied by a modified version of the Meyer and McConkie (1974) system produced by Meyer (1980). In this version, scoring of protocols involved
awarding points for each idea unit correctly recalled, with higher level idea units being worth proportionately more than lower level idea units. The resulting index, called the Message Score, showed the degree to which subjects recalled the most important ideas of the passage.

The relative importance of idea units was determined by averaging the scores of three independent judges. In this procedure, Meyer (1973b) had raters identify idea units in a text and then assign a rating to each proposition on a continuum of importance to the meaning of the text. Meyer justified the use of such a procedure on the basis of results reported in a study in which raters achieved a 91.5% average level of agreement in the identification of salient elements of the content structure of target passages (1971). More recently, Pederson (1985) used the same method in her study to obtain a weighted propositional count, with a reported interrater reliability of .91. Thus, this procedure has been proven to be highly effective in obtaining a reliable objective measurement of the amount of content correctly recalled in a recall protocol.
Problems in Scoring Recall Protocols

Despite the existence of a workable procedure by which idea units can be identified and assigned weighted values, problems in scoring recall protocols still exist. Potential problems include the resolution of questions concerning how to handle intrusions (e.g., information in the protocol that is not contained in the target passage), how to score generalized statements, achieving and maintaining scoring consistency (e.g., intrarater and interrater reliability), how strict the scoring criteria should be, and the choice of modality of recall (Voss et al., 1982).

While such problems as have been mentioned above do exist, however, they are by no means insoluble. Voss et al. (1982) suggest that these problems can be overcome by such strategies as using two or more scorers per protocol, using dual scoring criteria (e.g., one strict and one more lenient), and adhering to conventions accepted in the scoring of cloze-acceptable passages.

The Issue of Focus

Task Awareness and Depth of Processing

The question of whether readers approach a holistic evaluation task differently when they are aware they will be tested for comprehension than when they are not
forewarned is one that must be addressed in any research seeking to determine how much holistic raters understand of what they have read. This is so because questions still persist as to whether holistic readers might somehow read "differently" or more "carefully" if they knew they were to be tested specifically on the content of what they had read.

Recent studies touching upon the difference in reading performance of "naive" readers (e.g., subjects who are unaware they will be tested on the content of a passage) versus "focused" readers (e.g., subjects who are reading a passage for the purpose of remembering it) have shed some light on the problem. These studies include those by Kaczmerek (1980), Baker and Anderson (1982), and Lee (1986a).

Kaczmerek (1980) indirectly addressed the issue of focus in a study that sought to compare teacher assessments of writing proficiency with those made by trained raters. Kaczmerek's study was designed to assess the validity and reliability of two different types of evaluation methods, one of which was subjective and the other of which was objective. In her study, Kaczmerek used trained raters and writing instructors to evaluate compositions written by foreign students enrolled in an American university.

In that phase of the study concerning holistic evaluation of student responses to a traditional essay
question, writing instructors were asked to evaluate essays on a six-point proficiency scale. Trained raters were asked to read the same compositions. Their rating task, however, was different. First, the trained raters were instructed to read each essay for an understanding of the writer's intended meaning. Then, the raters were asked to provide a holistic evaluation of the same essays, using a six-point scale that focused on comprehensibility rather than proficiency.

Kaczmarek was not looking specifically for the differences in recall between naive and focused raters of ESL compositions written by university students. Nevertheless, she was sufficiently struck by the results she obtained to acknowledge that the holistic scores of focused raters may well have been influenced by the very fact that these raters knew they would be rewriting the essays in addition to rating them.

Whether or not their scores were influenced by the fact that they were focused on the purpose of the reading task, however, is by no means clear, as Kaczmarek made no provisions in her experimental design for investigation of that specific question. Nor were her focused readers using the same criteria for holistic evaluation as her naive readers. In addition, as hers was a study of concurrent validity, her focused readers were trained holistic raters, but not experienced ESL composition teachers, as
were her naive readers. Finally, Kaczmarek's procedures did not include having raters, either focused or naive, write recall protocols. Consequently, any conclusions drawn would be tenuous at best, especially insofar as rater comprehension, per se, was not an issue in Kaczmarek's study.

In the Baker and Anderson (1982) study mentioned earlier (see pp.27-28), subjects were also aware that they would be tested on what they had read. Moreover, some were further focused in that they were instructed to be on the alert for inconsistencies in the texts they were to read. Nevertheless, the researchers reported that their highly-focused subjects spent no more time in reading the target passages and were no more likely to notice inconsistencies during the reading process. Baker and Anderson were at somewhat of a loss as to why this was so, although they did conjecture that "One possible reason for this lack of an effect of instructions is that the demand characteristics of the task were such that all subjects processed the text carefully in preparation for the test questions" (1982, p.42).

It is interesting to note that the studies surveyed in previous sections of this literature review were all of a focused variety. That is to say, regardless of whether the task involved writing a summary (Kintsch et al., 1977) or a recall protocol (Meyer, 1982), answering a set of
multiple choice questions (Baker & Anderson, 1982), or holistically evaluating a composition and then rewriting it to more clearly reflect the author's intended meaning (Kaczmarek, 1980), the subjects were all aware that they would be tested in one way or another on what they had read.

That studies of this sort are focused is not surprising. Indeed, to attempt to execute an experiment in which the subjects involved were not sensitized to the test environment would likely not be feasible owing to the large number of subjects required to produce both focused and naive recall protocols on the same text. Moreover, it would be exceedingly difficult to do because of the extreme difficulty in controlling an experimental setting in which such reactive factors as the Hawthorne Effect might cause a subject to approach a reading task differently than usual, even if instructed to read "normally."

A few studies have been executed in which readers were not oriented to read for the purpose of recalling the target passage (e.g., Carrell, 1983, 1984b; Lee & Ballman, in press). However, it is still not clear if holistic raters will, as a matter of course, read in the same way when reading specifically to form a general impression as subjects read when knowingly participating in a recall protocol study.
One study that did seek to provide an answer to the question of how awareness of task influences amount of recall was that of Lee (1986a). In his study, Lee used 320 native English speaking university students enrolled in four different semester level Spanish classes. As his sample was drawn from two different universities, variations in student placement and instructional procedures presented a potentially confounding source of variability. The same textbooks were used in both programs at the levels being tested, however. All testing was carried out in the subjects' regular classrooms during regularly-scheduled class times, although the experimenter was present during testing in only four of the 16 classes tested. All instructions were given in English. The instructions contained within the test packets provided were varied so that an equal number of subjects were subjected to the respective test conditions.

All subjects were asked to read a passage written in Spanish and then provide a written recall protocol of what they had read. An equal number of subjects were required to produce a protocols in Spanish and English. Half of each group was given prereading instructions concerning the fact that they would be required to produce a protocol, while the other half received no prereading instructions beyond being told to read the passage at their own rate.
Protocols were analyzed on the basis of an idea unit matrix of the type recommended by Bransford and Johnson (1972). Recall scores were analyzed by means of a three-factor analysis of variance. On the basis of this analysis, Lee concluded that "prereading directions alone are not sufficient to enhance recall of a passage"(p.208).

It is tempting to generalize the results obtained by Lee (1986a) to a setting in which trained and experienced raters of ESL writing proficiency are engaged in a holistic evaluation task. However, it must be remembered that Lee's subjects were university students performing under experimental conditions in which "subjects assume they will be performing some sort of task with the content of the passage"(Lee, 1986a, p.206). Subjects' assumptions that the reading task involved attention to the content of the passage may have indeed been a confounding variable in Lee's experiment. But in a task involving the assessment of writing proficiency, attention to content becomes crucial to the issue of the role of comprehension in the process of holistic evaluation.

Specifically with regard to holistic evaluation, the question of whether prior knowledge that one will be tested for comprehension will materially influence how closely one attends to detail in the target passage is still unanswered. In other words, will a subject read a passage differently (i.e., more carefully) for
comprehension than he/she reads for the purpose of holistic evaluation? And will a subject be able to read for comprehension while still attending to a holistic evaluation task?

**Shared Goals and Holistic Assessment**

Kintsch and van Dijk (1978) assert that research on comprehension should focus on those cases where readers share similar goals. Johnson's (1973) investigation of the relationship between the perceived importance of ideas in a passage and the likelihood of their being recalled revealed that "ideas perceived as being more important to the passage were more likely to be recalled" (p.107). Whether this is also the case for entire passages read by focused versus unfocused readers has yet to be demonstrated, although Voss et al. (1982) point to recent theories of text processing (e.g., Just & Carpenter, 1980 and Kintsch & van Dijk, 1978) that "emphasize the importance of the reader's goals in determining how processes are executed during reading" (p.379).

The similarity between what a reader reads and what he/she remembers from the text depends, according to Meyer (1981), "on the prior knowledge, purpose, and skill level of the readers" (p.10). Provided that subjects can be assumed to possess essentially equivalent prior knowledge and skill level, the question of purpose (e.g., reading to
assess writing proficiency only versus reading for comprehension in addition to assessment of proficiency) is one that can be directly addressed.

Research Hypotheses

In light of the empirical data provided by the literature review, the general research question concerning the extent to which a trained holistic evaluator of L2 writing proficiency comprehends what he/she has read that was posed at the conclusion of Chapter 1 can be restated as two specific research hypotheses:

Research Hypothesis #1: For experienced teachers of ESL composition engaged in a holistic evaluation task:

Is there a significant difference between the amount of content they recall after reading a text generated by a more proficient non-native speaking author and the amount they recall after reading a text generated by a less proficient non-native speaking author?

In order to investigate this hypothesis, the following question will be posed:

1. Will holistic readers oriented toward assessment of writing proficiency recall significantly more of a higher-rated text than of a lower-rater text?
Research Hypothesis #2: With respect to holistic evaluators of ESL compositions who are specifically oriented toward reading for comprehension in addition to reading for the purpose of assessing writing proficiency, what is the impact of such an orientation upon recall of the content of an L2 text?

In order to investigate this hypothesis, the following questions will be posed:

2. Will focused readers recall significantly more of a higher-rated text than of a lower-rated text?

3. Will focused readers of a higher-rated text recall significantly more than their naive counterparts?

4. Will focused readers of a lower-rated text recall significantly more than their naive counterparts?

5. Will focused readers recall significantly more of both higher-rated and lower-rated texts than their naive counterparts?

6. Will readers of a higher-rated text, regardless of orientation, recall significantly more than readers of a lower-rated text?
CHAPTER 3

PROCEDURES

Population

Recent and current composition instructors in the ESL Department of the Ohio State University.

Sample Selection

Twelve subjects were recruited from among the population of experienced ESL composition instructors.

Research Design

The study employed a repeated measurements design within a 2-way analysis of variance (ANOVA) matrix to test for significant interactions and main effects (see Figure 1). Follow-up procedures included a series of one-way analyses of variance to investigate individual cell means. Results of subject-generated written recall protocols were compared across levels of Lower versus Higher ratings of assessed proficiency of student compositions (a rating of relative text quality) and Naive versus Focused readers.
<table>
<thead>
<tr>
<th>Reader Orientation</th>
<th>Levels of Assessed Proficiency of Authors of ESL Compositions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower ( (A_1) )</td>
<td>Higher ( (A_2) )</td>
</tr>
<tr>
<td>Naive ( (B_1) )</td>
<td>( (A_1B_1) )</td>
</tr>
<tr>
<td></td>
<td>( (A_2B_1) )</td>
</tr>
<tr>
<td>Focused ( (B_2) )</td>
<td>( (A_1B_2) )</td>
</tr>
<tr>
<td></td>
<td>( (A_2B_2) )</td>
</tr>
</tbody>
</table>

Figure 1.

2 X 2 factorial design
Interrater reliability between subjects' mean ratings of ESL compositions and ratings established for those compositions during the original Diagnostic evaluation process were computed using the Pearson Product-Moment correlation procedure.

**Variables**

**Independent**

1. Levels of assessed proficiency, as reflected by qualitative evaluation of ESL compositions (Lower, or Level 107 versus Higher, or Level 108);

2. Holistic rater awareness of reading task (Naïve versus Focused).

**Dependent**

1. Mean scores of recall protocols of ESL compositions generated by holistic raters.

**Experimental Procedures**

**Overview of the Plan**

Sample compositions written by non-native speaking (NNS) graduate students enrolled at the Ohio State University were supplied to the researcher by instructors of ESL Composition at the 106, 107, and 108 levels of the ESL Composition course sequence. These compositions were used, first, to develop and test interrater reliability for a holistic evaluation task and, second, to test the
reading comprehension of both "focused" and "naive" raters engaged in the aforementioned evaluation task.

The researcher divided the sample compositions into four groups (e.g., 106, 107, 108, and Q) and randomly selected one from 107 and 108 to serve as a basis for scoring templates of the kind used by Meyer (1973) and Pederson (1985) in text analysis procedures. These templates, referred to in this study as Protocol Scoring Matrices (PSMs), were constructed in the following manner:

1. After the researcher selected the 107 (Lower) and 108 (Higher) student compositions, he and two other experienced ESL composition instructors individually rewrote both student protocols to reflect what he/she thought the authors meant. The researcher's colleagues were familiar with the procedure, but were still given a detailed set of instructions to refresh their memories.

2. Upon completion of the rewrite task, researcher and colleagues met to discuss and amend the revised protocols. Provision was made for the researcher to seek clarification from the author(s) of the student text(s), should any unresolved ambiguities or differences in interpretation have remained. As the samples in question contained no such problem areas, this step was not taken. The eventual result of this process was a "negotiated" protocol for each student text.
3. Once the "negotiated" texts were produced, the researcher and a colleague experienced in the procedure independently identified the idea units contained in each text. Following the Meyer (1974) procedure, this study defined idea units as "primarily consisting of noun, verb, and prepositional phrases" (Bernhardt, 1983b, p.29). In addition, certain logical connectors were considered as distinct idea units, inasmuch as the composition program in which the subjects participated stresses the crucial role of such logical connectors in producing text coherence and cohesiveness.

4. Lists of idea units for each text were compared and served as the basis on which a weighted propositional count was compiled. This weighted propositional count, in turn, served as the basis upon which a PSM was constructed.

5. Following the method used by Pederson (1985), independent readers read the texts, compared the texts' contents with the appropriate list of idea units, and weighted each unit in terms of relative importance to the text.

6. The resulting templates were used to score rater-generated written recall protocols that were designed to measure the reading comprehension of raters involved in holistic evaluation of texts written by NNS students.
7. Remaining sample compositions were used either as practice papers or calibration samples (Proficiency Models) during a training session in which interrater reliability among participants was established.

8. After establishing an acceptable level of interrater reliability, raters were asked to holistically evaluate the two compositions for which PSMs had been prepared. After they had rated the first composition, raters were asked to generate a written recall protocol of its content as they remembered it. This was the Naive reading. The Focused reading followed immediately upon conclusion of the Naive, and was identical with the exception that the subjects were told that after they had finished reading and rating the second composition, they would again be asked to generate a written recall protocol of what they had just read.

9. Recall protocols were then scored against the PSMs to determine a reading comprehension score for each rater on each target composition, one as a "naive" reader and the other as a "focused" reader.

10. Finally, the data obtained from the recall protocols were analyzed with respect to the variables of level of assessed proficiency and reader orientation.
Overview of the Program

NNS students admitted to the Ohio State University are obliged to produce a writing sample to determine their level of proficiency in written English. NNS students placed into the University's ESL Composition course sequence are required to provide a second writing sample, called a Diagnostic, to double-check the results of initial placement procedures. This writing activity involves responding to a pre-selected topic within one 48 minute class period. This question is of an open-ended nature, and so allows for some flexibility in response, while leading the student to respond within certain parameters of rhetorical organization and tense usage.

This "second chance" mechanism can result in a student moving up in or out of the course sequence. In theory, a student may also be moved into a lower level in the instructional sequence if his/her Diagnostic score is markedly lower than expected. In practice, however, this rarely occurs. Most of the time, students initially placed into one level of the sequence remain in that level.
Instrumentation

Measurement of Reading Comprehension

Procedures for Collecting Writing Samples

After the Diagnostic writing task was administered, compositions were evaluated, and final placement determinations were made, the following procedure was employed:

STEP 1 - Instructors in all sections of the ESL Composition sequence (e.g., levels 106, 107, and 108) randomly selected three compositions from among those written by the graduate students in their classes. In addition, instructors also passed along all compositions scored as "Q" (Qualified, or possessing adequate writing proficiency to move directly into the regular University curriculum) (see APPENDIX C).

STEP 2 - These instructors passed on the compositions they had selected to the researcher, which he in turn photocopied and returned to them.

STEP 3 - The researcher combined writing samples from individual sections into four groups according to sequence level (e.g., 106, 107, 108, Q). From among these groups
of compositions, the researcher selected one from the level 107 group and one from the 108 group (see APPENDIX D). Selection of these two texts was made using two criteria:

1. Topic - Each composition addressed a different topic. This was done so as to avoid the possibility of confusion and carryover of features from one sample to the other.

2. Length - Each text was at least 250 words long. Compositions of this length were chosen to assure that subjects could not memorize substantial portions of the texts given the relatively short time anticipated for reading (Bernhardt, 1986).

These two compositions, respectively referred to as the Lower (level 107) and Higher (level 108) sample, served as the basis upon which the PSMs were constructed.

STEP 4 - From among the remaining sample compositions, the researcher, in conjunction with the Program Director, selected four compositions that were deemed to be representative of the writing proficiency of NNS students placed into levels 106, 107, 108, and Q. These four compositions served as rangefinder samples during the calibration session (see APPENDIX E).
STEP 5 - The remaining sample compositions were randomly ordered to use for practice and discussion (see APPENDIX F) and to establish interrater reliability (see APPENDIX G) during the calibration session.

Procedures for Developing Protocol Scoring Matrices (PSMs)

Upon completion of STEP 3 (see above, pp.81-82), the researcher proceeded as follows to develop PSMs for both the Lower and Higher texts:

STEP 1 - After reading the passage thoroughly, the researcher rewrote the passage to reflect what he thought its author meant.

STEP 2 - The above procedure was repeated by two colleagues experienced in teaching ESL Composition, each working independently of the other (see APPENDIX H).

STEP 3 - The researcher and his colleagues met to compare and discuss their rewritten versions of the original text.
STEP 4 - Had the researcher and his colleagues failed to reach consensus on any portion of the text, the researcher would have met with the author of the text to discuss, clarify, and amend the passages in question. As both texts were relatively transparent, the researcher and his colleagues had no trouble reaching a consensus. Consequently, this step was omitted.

STEP 5 - The amended passage, as rewritten by the researcher and his specialist colleagues, served as the basis upon which the PSM was constructed (see APPENDIX I). The template generated from this "negotiated" document served as the standard against which evaluators' recall protocols were measured.

STEP 6 - To create the PSM, itself, the researcher used a weighted propositional count of the type described by Meyer (1973) and used by Pederson (1985). First, the researcher developed a listing of all possible propositions (idea units) present in the target text (see APPENDIX J). Next, a colleague experienced in the procedure matched the list to the text and made suggestions for additions and/or deletions to the list of propositions. Then, three other colleagues were given a copy of the "negotiated" text to use as a guide
and asked to assign each proposition a rating on a seven-point scale (1 = of negligible importance to the meaning of the text, 7 = extremely important to the meaning of the text) (see APPENDIX K). The assigned value of each proposition equaled the mean rating.

**Measurement of Writing Proficiency**

After the PSMs had been completed, the researcher typed all sample compositions exactly as they were written and duplicated a sufficient quantity to be made available to all raters participating in the training session.

The researcher then arranged for study participants (selected instructors in the 106, 107, 108 sequence) to attend the session. The training/calibration session refined the evaluation criteria to be used by holistic raters. These criteria were established by means of so-called "rangefinder papers" (Proficiency Models), discussion of salient features of "typical" papers at each respective level, and use of guidelines during the rating process (see APPENDIX L).
Procedures for Training Holistic Raters

STEP 1 - Prior to the session, subjects were provided with copies of calibration samples and a list of holistic scoring criteria. Both calibration samples and holistic rating criteria had previously been submitted to the Program Director for his approval to assure that they accurately reflected the ESL Composition program's placement standards. Subjects were asked to familiarize themselves with both calibration samples and holistic criteria.

STEP 2 - Researcher reviewed the process of holistic evaluation and respond to questions concerning samples and criteria.

STEP 3 - After responding to questions and clarifying minor procedural points concerning the rating criteria and general qualities and characteristics of proficiency, the researcher had subjects practice holistically evaluating randomly-presented sample compositions. These compositions were chosen from among previously-rated Diagnostic essays to provide a standard against which to measure the interrater reliability of the group. Placement level (106, 107, 108, Q) for each practice
essay was verified once by the researcher and then once more by the Program Director. Scores were recorded as 106, 107, 108, or Q to reflect what the raters felt was an appropriate level of placement.

Procedures for Establishing Interrater Reliability

STEP 4 - Once raters were calibrated, the evaluation session continued as before, with the exception that subjects kept a log of their holistic ratings (see APPENDIX M). Data from this phase was collected and interrater reliability was computed. Group means of holistic ratings were calculated. These group means on the nine essays were found to correlate with the pre-established ratings of the essays at a level of .9306.

Procedures for Eliciting "Naïve" and "Focused" Recall Protocols

STEP 5 - Once subjects received adequate training and evidenced an acceptable level of interrater reliability on nine NNS compositions, half were given the Lower text and half the Higher text and were asked to rate them in the same way they had rated the previous
texts. This procedure was carried out in such a way that the raters were not aware that this text-rating task was in any way different from the previous nine. This was done to assure that the raters approached their "naive" rating task in the same way that they had approached their earlier rating tasks.

STEP 6 - After finishing the "naive" reading, the subjects were given a coded answer booklet (see APPENDIX N) and instructed to write down everything they recalled about the texts they had just read (see APPENDIX O). Specific instructions followed the guidelines established by Johnson (1970), with paraphrasing allowed when necessary (Frederiksen, 1977; Mandler, 1978). These recall protocols furnished the researcher with his "naive" sample.

STEP 7 - Upon completion of the "naive" task, subjects were administered the second task. Procedurally, this involved giving the Higher text to the raters who had first read the Lower text and vice versa (see Figure 2). This time, however, subjects were informed that in addition to giving a holistic rating to the text, they would again be asked to generate a written recall protocol. These protocols comprised the "focused" sample.
<table>
<thead>
<tr>
<th>Subjects</th>
<th>Naive</th>
<th>Reader Orientation*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>L</td>
<td>Focused</td>
</tr>
<tr>
<td>2</td>
<td>H</td>
<td>H</td>
</tr>
<tr>
<td>3</td>
<td>L</td>
<td>H</td>
</tr>
<tr>
<td>4</td>
<td>H</td>
<td>L</td>
</tr>
<tr>
<td>5</td>
<td>L</td>
<td>H</td>
</tr>
<tr>
<td>6</td>
<td>H</td>
<td>L</td>
</tr>
<tr>
<td>7</td>
<td>L</td>
<td>H</td>
</tr>
<tr>
<td>8</td>
<td>H</td>
<td>L</td>
</tr>
<tr>
<td>9</td>
<td>L</td>
<td>H</td>
</tr>
<tr>
<td>10</td>
<td>H</td>
<td>L</td>
</tr>
<tr>
<td>11</td>
<td>L</td>
<td>H</td>
</tr>
<tr>
<td>12</td>
<td>H</td>
<td>L</td>
</tr>
</tbody>
</table>

* L = Lower-rated (107) text
  H = Higher-rated (108) text

Figure 2.

Order of administration of Naive and Focused reading
tasks with respect to target texts
Data Analysis

Upon conclusion of the experiment, the researcher evaluated the recall protocols using the PSMs. Protocols were matched with the appropriate PSM to determine the number of correctly recalled idea units per protocol. Next, idea units correctly recalled were converted into a weighted propositional score for each protocol. The sum of the weighted scores of all the propositions included in each respective recall protocol comprised the raw recall score for that particular recall protocol.

Following the format adopted by Pederson (1985), neither incorrect nor inaccurate information was counted against the score, as per the recommendation of Murphy and Puff (1982). Moreover, idea units correctly identified in the recall protocols were counted as correct, regardless of the order in which they appeared.

As has been pointed out by Bernhardt (1985), scoring of recall protocols has been under considerable discussion in the past few years. From the earliest schemes, which awarded one point for each "meaningful phrase" appearing in a protocol (cf. Meyer, 1973, 1974), scoring has evolved to the point where scores are now generally weighted to reflect the relative importance of idea units to understanding the passage as a whole. This study used the seven-point weighted propositional count employed by Pederson (1985).
Moreover, since the purpose of the protocol scoring system was to provide a quantitative measure of comprehension, the point that "quantitative measures are product measures and actually lend limited insight into how information is encoded" (Bernhardt, 1985, p.8) is moot. Because insights into the comprehension process, per se, were not the focus of this study, it used only the first step in the kind of content-structure analysis of a text described by Connor (1984); namely, a clause analysis to identify individual or simple propositions contained within the text.

Finally, of the seven guidelines for scoring suggested by Meyer (1981), only the first two (i.e., credit for "any reasonable paraphrase in terms of general meaning," and joint agreement by raters on "limits for acceptable paraphrases" (p.289)) were used. This was so because the study did not concern itself with such details of the content structure as role relationships, lexical predicates, and rhetorical predicates.

In order to establish the rater reliability of the researcher, a second rater was also employed. This rater was trained in both how to use the rating scale and how to score a recall protocol. In addition, she was allowed to familiarize herself with the content of the original sample compositions prior to scoring them. After training, the second rater and the researcher practiced
the procedure using compositions included in the original corpus but not used in the experiment. Midway through the practice session, researcher and co-rater discussed discrepancies in their protocol scores. Practice then resumed. At the conclusion of the protocol session, interrater reliability was computed to be .9150.

This co-rater then scored the recall protocols on the targeted ESL compositions, following the same procedures as the researcher. Finally, interrater reliability was again computed. Interrater reliability for the actual scoring of the recall protocols used in this study was .9695.

Disagreements concerning the presence or absence of idea units in the protocols were discussed. Ultimately, it was agreed that no changes were necessary, so no idea units were added or deleted from the raw recall scores.

Once the raw recall scores were obtained, they were converted into weighted recall scores. These totals, in turn, were converted into percentages by dividing them into the sum total of weighted idea units in the target composition. In this way, the Higher and Lower compositions were made amenable to comparison.
Summary of Null Hypotheses

Ho1: There will be no significant difference between the amount of content recalled by naive readers of the higher-rated text and the lower-rated text.

Ho2: There will be no significant difference between the amount of content recalled by focused readers of the higher-rated text and the lower-rated text.

Ho3: There will be no significant difference between the amount of content recalled by focused readers of the higher-rated text and naive readers of the higher-rated text.

Ho4: There will be no significant difference between the amount of content recalled by the focused readers of the lower-rated text and the naive readers of the lower-rated text.

Ho5: There will be no significant difference between the total amount of content recalled by focused readers and the total amount of content recalled by naive readers.
Ho6: There will be no significant difference between the total amount of content recalled by readers of the higher-rated text and the total amount of content recalled by readers of the lower-rated text.
CHAPTER 4
RESULTS AND DISCUSSION

Introduction

This study investigated the effects of both reader orientation and level of assessed proficiency on the amount of content recalled by holistic raters of compositions written by university-level non-native speakers of English. The experiment was divided into three phases; a preliminary one in which raters familiarized themselves with holistic scoring criteria, a second one in which interrater reliability was established, and a main phase in which written recall protocols were solicited to measure rater comprehension.

In the first phase, twelve experienced raters of ESL writing proficiency at the university level discussed holistic scoring criteria and practiced evaluating four ESL compositions similar to the ones used in the remainder of the study. In the second phase, the evaluators were asked to holistically rate nine compositions authored by non-native speakers of English for the purpose of
assessing each author's level of proficiency in written English.

These nine compositions, which were presented one at a time in random order, served a twofold purpose. First, they allowed the subjects additional opportunity for practice and feedback, so that a high level of interrater reliability could be established. Second, the assignment of several compositions to read and rate holistically created a reading environment similar to that of a typical grading session in which readers holistically evaluate a number of student essays in a relatively short period of time.

Once interrater reliability and rating environment had been established, the next part of the study commenced. In it, copies of one of two essays were distributed to the subjects on an alternating basis. Half received the essay pre-judged as "Lower" (written by a non-native speaker placed into Level 107 of the ESL Composition course sequence at the Ohio State University), while half received the "Higher" essay, which was produced by a non-native speaker placed into Level 108, the more advanced level of the program. Distribution of the two sets of essays was accomplished in such a way that the subjects were unaware that two different essays were being rated by the group instead of the usual one. Thus, to the subjects, this tenth round
of rating tasks was in no way different from the previous nine.

Upon completion of holistic ratings in this round, the researcher distributed coded test booklets and instructed each subject to generate a written recall protocol of what he/she had just read. This constituted the Naive portion of the reader orientation treatment. Next, the order of distribution for the Lower and Higher essays was reversed and subjects were instructed to read and holistically rate them. In this round, however, subjects were advised that they would be writing another recall protocol of the content of this last composition. This constituted the Focused portion of the reader orientation, in that the readers' perceived task was now to read for the purpose of recalling the contents of the text in addition to holistically evaluating it.

A 2 X 2 factorial design was chosen for initial analysis of the data, inasmuch as both independent variables consisted of two levels, statistically significant main effects were expected, and statistically significant interactions were thought to be possible. Thereafter, a series of one-way analyses of variance were computed to probe the relationships between individual cell means.

The first independent variable, level of assessed proficiency, consisted of one essay generated by a less
proficient non-native speaker (Lower) and one generated by a more highly proficient non-native speaker (Higher). The second independent variable, reader orientation, consisted of two levels of task awareness; e.g., Naive and Focused.

As was indicated above, naive readers were conditioned by the preliminary phases of this study (reading ESL compositions and holistically rating them on a four-point scale) to approach the target composition as they had the previous evaluation tasks. Naive readers had no warning that they would be asked to generate a recall protocol based on the target composition, either Higher or Lower, they had just holistically rated. All subjects were Naive until after writing their first recall protocol. Thereafter, all subjects became Focused, both by virtue of having just written a recall protocol and because they were specifically informed that the next task would involve generating a recall protocol in addition to holistically rating an essay.

The dependent variable in this study was content recalled. It was assumed for the purpose of this study that the amount of content recalled in a written protocol would provide a measurement of the degree to which the subjects comprehended what they read. This variable was measured in terms of the percentage of idea units correctly recalled by subjects on the two written protocols, one following the Naive round and the other
following the Focused round. These idea units were weighted according to their relative importance to the respective target compositions.

An analysis of variance appropriate to a two-way factorial design was computed using the SAS (Statistical Analysis System) at the Ohio State University on an Amdahl V-7B computer, which employs an IBM MVS system of operation. In order to determine the nature of the single statistically significant main effect that emerged, a one-way analysis of variance of the "Lower" level of Reader Orientation was conducted. In addition, a one-way analysis of variance was computed for both the "Naive" and the "Focused". levels of Assessed Proficiency. Finally, a one-way analysis of variance was computed for the "Higher" level of Reader Orientation. In this way, data relevant to all six research hypotheses was obtained.

**Results**

**Experimental Data on the Recall Dependent Variable**

Table 1 provides the means and standard deviations for the two independent variables, reader orientation and level of assessed proficiency. Examination of overall means on the table reveals that those subjects whose orientation was focused during the two recall tasks scored substantially higher than those whose orientation was
Table 1

Means and Standard Deviations of Recall Scores by Reader Orientation and Proficiency Assessment

<table>
<thead>
<tr>
<th>Reader Orientation</th>
<th>Assessed Proficiency</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower</td>
<td>Higher</td>
<td>Overall</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>x</td>
<td>σ</td>
<td>n</td>
</tr>
<tr>
<td>Naive</td>
<td>6</td>
<td>27.28</td>
<td>9.09</td>
<td>6</td>
</tr>
<tr>
<td>Focused</td>
<td>6</td>
<td>53.63</td>
<td>14.07</td>
<td>6</td>
</tr>
<tr>
<td>Overall</td>
<td>12</td>
<td>40.45</td>
<td>17.80</td>
<td>12</td>
</tr>
</tbody>
</table>
naïve (\(\bar{x_f} = 52.07, \bar{x_n} = 35.72\)). With respect to the variable of assessed proficiency, examination of the table indicates that subjects who read the higher-rated text recalled more overall than subjects who read the lower-rated text. The mean difference, however, was considerably less than for the Naïve versus Focused grouping (\(\bar{x_h} = 47.34, \bar{x_l} = 40.45\)).

Table 2 provides a summary of a factorial analysis of variance for a two-factor design. Examination of Table 2 reveals no statistically significant interaction between Variable A (Assessed Proficiency) and Variable B (Reader Orientation). Accordingly, we may proceed directly to an investigation of main effects.

As was stated above, main effects for Variables A and B both showed a difference between means. Table 2, however, reveals that while there was no statistically significant difference between A levels of main effects, there was a statistically significant difference between B levels of main effects \(F(1,20) = 8.56, p < .01\). In other words, there was no significant difference between amount of content recalled by readers of "Lower" compositions and readers of "Higher" compositions, but "Focused" readers recalled significantly more overall of A-level (Lower + Higher) compositions than "Naïve" readers.
Table 2

### Analysis of Variance of Reading Comprehension by Reader Orientation and Proficiency Assessment

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proficiency Assessment</td>
<td>1</td>
<td>284.1440</td>
<td>284.1440</td>
<td>1.52</td>
<td>0.2324</td>
</tr>
<tr>
<td>(A)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reader Orientation</td>
<td>1</td>
<td>1602.9541</td>
<td>1602.9541</td>
<td>8.56</td>
<td>0.0084*</td>
</tr>
<tr>
<td>(B)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proficiency by</td>
<td>1</td>
<td>601.6010</td>
<td>601.6010</td>
<td>3.21</td>
<td>0.0883</td>
</tr>
<tr>
<td>Orientation (AB)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error (S/AB)</td>
<td>20</td>
<td>3746.8225</td>
<td>187.3411</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>6235.5218</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p < .01
Next, we turn to an examination of individual cell means. Table 3 illustrates means and standard deviations for individual cells. Perusal of individual cell means within Levels A and B strongly suggests that further investigation be concentrated on Levels A₁ and A₂ across Level B₁ (Levels of Assessed Proficiency across Naive Reader Orientation), as well as upon Levels B₁ and B₂ across Level A₁ (Reader Orientations across "Lower" Proficiency Assessment).

Table 4 provides a summary of a one-way analysis of variance for a one-factor design involving Levels A₁ and A₂ across Level B₁. Examination of Table 4 indicates a statistically significant difference between the A₁B₁ (Lower/Naive) cell mean and the A₂B₁ (Higher/Naive) cell mean (F(1,10) = 5.12, p < .05). That is to say, Naive readers recalled significantly more content of the "Higher" composition than Naive readers recalled of the content of the "Lower" composition.

Table 5 provides a summary of a one-way analysis of variance for a one-factor design involving Levels B₁ and B₂ across Level A₁. Examination of the table shows that a statistically significant difference exists between the A₁B₁ (Lower/Naive) and A₁B₂ (Lower/Focused) cell means. In other words, Focused readers recalled significantly more of the "Lower" composition than did their Naive counterparts (F(1,10) = 14.84, p < .005).
<table>
<thead>
<tr>
<th>Variable B: Reader Orientation</th>
<th>Variable A: Levels of Assessed Proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive $B_1$</td>
<td>$A_1$ 107 (Lower)</td>
</tr>
<tr>
<td></td>
<td>$\bar{x} = 27.28$</td>
</tr>
<tr>
<td></td>
<td>$\sigma = 9.09$</td>
</tr>
<tr>
<td></td>
<td>$A_2$ 108 (Higher)</td>
</tr>
<tr>
<td></td>
<td>$\bar{x} = 44.17$</td>
</tr>
<tr>
<td></td>
<td>$\sigma = 15.86$</td>
</tr>
<tr>
<td>Focused $B_2$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\bar{x} = 53.63$</td>
</tr>
<tr>
<td></td>
<td>$\sigma = 14.07$</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\bar{x} = 50.50$</td>
</tr>
<tr>
<td></td>
<td>$\sigma = 14.72$</td>
</tr>
</tbody>
</table>
Table 4

Analysis of Variance of Reading Comprehension (Lower/Naive by Higher/Naive)

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model ($A_1B_1/A_2B_1$)</td>
<td>1</td>
<td>856.3239</td>
<td>856.3230</td>
<td>5.12</td>
<td>.0471*</td>
</tr>
<tr>
<td>Error</td>
<td>10</td>
<td>1672.4775</td>
<td>167.2477</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>2528.8006</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*P < .05
Table 5

Analysis of Variance of Reading Comprehension
(Lower/Naive by Lower/Focused)

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>2084.2852</td>
<td>2084.2852</td>
<td>14.84</td>
<td>.0032*</td>
</tr>
<tr>
<td>(A_1B_1/A_1B_2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>10</td>
<td>1404.4382</td>
<td>140.4438</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>3488.7234</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p < .005
Results for Null Hypotheses

Following are the specific results for each null hypothesis tested in this experiment.

**Ho1:** There will be no significant difference between the amount of content recalled by naive readers of the higher-rated text and the lower-rated text. This hypothesis must be rejected on the basis of the data. The one-way analysis of variance shown in Table 4 yielded a significant F-ratio ($F(1,10) = 5.12$, $p < .05$). Table 1 shows that Naive readers of the higher-rated compositions produced a mean recall score of 44.17 (S.D. = 15.86), compared to a mean recall score of 27.28 (S.D. = 9.09) obtained by Naive readers of the lower-rated composition.

**Ho2:** There will be no significant difference between the amount of content recalled by focused readers of the higher-rated text and the lower-rated text. On the basis of the data from this study, this hypothesis must be retained. A one-way analysis of variance (see Table 6) yielded an F-ratio that was not significant. Table 1 shows that Focused readers of the higher-rated text yielded a mean recall score of 50.50 (S.D. = 14.72), while Focused readers of the lower-rated text yielded a slightly higher mean recall score of 53.63 (S.D. = 14.07).
Table 6

Analysis of Variance of Reading Comprehension
(Focused/Higher by Focused/Lower)

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model (A_1B_2/A_2B_2)</td>
<td>1</td>
<td>29.4220</td>
<td>29.4220</td>
<td>0.14</td>
<td>.7143</td>
</tr>
<tr>
<td>Error</td>
<td>10</td>
<td>2074.3450</td>
<td>207.4345</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>2103.7670</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Ho3: There will be no significant difference between the amount of content recalled by focused readers of the higher-rated text and naive readers of the higher-rated text. This hypothesis must be retained. A one-way analysis of variance (see Table 7) produced an F-ratio that was not significant at the .05 level. Table 1 shows the mean recall scores of Focused and Naive readers to be ($\bar{x}_f = 50.50, S.D. = 14.72$) and ($\bar{x}_n = 44.17, S.D. = 15.86$), respectively.

Ho4: There will be no significant difference between the amount of content recalled by focused readers of the lower-rated text and naive readers of the lower-rated text. This hypothesis must be rejected on the basis of the data from this study. The one-way analysis of variance shown in Table 5 yielded an F-ratio of ($F(1,10) = 14.84, p < .005$). Table 1 reveals that Focused readers of the lower-rated text showed a mean recall score of 53.63 ($S.D. = 14.07$), which was significantly higher than the mean recall score produced by their Naive counterparts ($\bar{x} = 27.28, S.D. = 9.09$).

Ho5: There will be no significant difference between the total amount of content recalled by focused readers and the total amount of content recalled by naive readers.
<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model (A₂B₁/A₂B₂)</td>
<td>1</td>
<td>120.2700</td>
<td>120.2700</td>
<td>0.51</td>
<td>.4900</td>
</tr>
<tr>
<td>Error</td>
<td>10</td>
<td>2342.3842</td>
<td>234.2384</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>2462.6542</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
On the basis of the data, this hypothesis must be rejected. The two-way analysis of variance summarized in Table 2 yielded an F-ratio that was statistically significant \( F(1,20) = 8.56, p < .01 \). Table 1 reveals that Focused readers produced a mean recall score of 52.07 (S.D. = 13.82), while Naive readers produced a mean recall score of 35.72 (S.D. = 15.16).

H06: There will be no significant difference between the total amount of content recalled by readers of the higher-rated text and the total amount of content recalled by readers of the lower-rated text. This hypothesis must be retained on the basis of the data. The two-way analysis of variance summarized in Table 2 yielded an F-ratio of \( F(1,20) = 1.52 \), which did not approach statistical significance. Table 1 shows that readers of the higher-rated text produced a mean recall score of 47.34 (S.D. = 14.96), while readers of the lower-rated text generated a mean recall score of 40.45 (S.D. = 17.80).

**Discussion**

Although the data in this study yielded no significant interactions and only the main effect for reader orientation achieved statistical significance, analysis of individual cell means revealed information of
direct relevance to the hypothesis that formed the basis of this experiment. In particular, comparisons of cell means across levels of reader orientation and assessed proficiency are of special importance to the issue of the impact of comprehension upon holistic evaluation of L2 writing proficiency.

The results of this study indicated that Null Hypotheses 1, 4, and 5 were rejected. In the case of Null Hypothesis 1, it was found that naive readers recalled significantly more of the higher-rated text than of the lower-rated text. That is to say, raters engaged in the holistic evaluation of compositions generated by non-native speakers of English comprehended significantly more of the text written by the more proficient author than of the one produced by the less proficient author.

Null Hypothesis 4 was rejected when it was calculated that focused readers recalled significantly more of the lower-rated text than did naive readers. In other words, reader orientation, or more specifically task awareness concerning the generation of a written recall protocol, appeared to be an important factor in how well subjects comprehended the lower-rated text.

Null Hypothesis 5, in which amount of recall was considered across levels of reader orientation, was also rejected. Focused readers as a group comprehended
significantly more of the content of the two target texts than did naive readers.

On the basis of the data analyzed in this experiment, Null Hypotheses 2, 3, and 6 were retained. Null Hypothesis 2 stated that focused readers would not recall significantly more of the higher-rated text than the lower-rated text. In fact, focused readers of the lower-rated text actually recalled slightly more than did focused readers of the higher-rated text ($\bar{x}_{f/lo} = 53.64$, $\bar{x}_{f/hi} = 50.51$).

Null Hypothesis 3 posited that focused readers of the higher-rated text would not recall significantly more of its content than their naive counterparts. While the focused readers recalled more than the naive readers ($\bar{x}_{f/hi} = 50.51$, $\bar{x}_{n/hi} = 44.18$), the difference was not statistically significant at the .05 level of probability.

In Null Hypothesis 6, it was predicted that there would be no significant difference between the total amount of content recalled by readers of the higher-rated text and content recalled by readers of the lower-rated text. That is, naive and focused readers of the "Higher" composition taken as a group would not recall more than their counterparts who read the "Lower" composition. In this case, the "Higher" group recalled more, but the difference was not significant at the .05 level ($\bar{x}_{hi} = 47.34$, $\bar{x}_{lo} = 40.46$).
In sum, naive readers recalled significantly more of the higher-rated text than of the lower-rated text, while focused readers scored significantly higher on overall recall than naive readers, with focused readers recalling significantly more of the lower-rated text than their naive counterparts. On the other hand, no significant difference in recall was found to exist between overall levels of assessed proficiency, naive versus focused readers of the higher-rated text, or focused readers of higher and lower-rated texts.
CHAPTER 5

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

Summary of the Study

The primary purpose of this study was to investigate the construct validity of holistic evaluation of compositions as a means of assessing the writing proficiency of non-native speaking graduate students at the university level. Proceeding from the assumption that any assessment of language proficiency must include a comprehension component, this study focused upon the effects of writing quality on comprehension for experienced ESL composition instructors engaged in the task of holistically evaluating expository texts generated by non-native speaker/writers of English.

The data collection phase of the experiment consisted of three parts. During the first part, subjects rated a series of ESL compositions to practice the holistic rating procedure, using holistic criteria and rangefinder papers that had been distributed one week earlier as guidelines upon which to base their assessments of the quality of the compositions. Upon completion of the rating of each
essay, scores were compared and subjects discussed those qualities and characteristics which had influenced their evaluations. This part of the session lasted approximately 30 minutes.

After the subjects were familiarized with the guidelines and procedures, the second part of the session commenced. In this part, subjects rated a series of nine ESL compositions on a four-point scale of demonstrated writing quality reflective of the four levels of proficiency into which NNS students are categorized for the purpose of placement into the Ohio State University's ESL Composition course sequence.

These compositions, chosen at random from among a corpus of Diagnostic writing tasks administered to foreign students at the beginning of the Autumn, 1986 quarter, were again holistically rated by the researcher and the ESL Program Director. From among those compositions on which total agreement was reached, the nine compositions used in this part of the study were selected.

This part of the experiment served a twofold purpose. First, rating results obtained were used to compute interrater reliability with respect to the "standard" scores assigned by evaluator, researcher, and Program Director to each of the nine compositions. Interrater reliability, as determined by computation of a
Pearson Product-Moment correlation, was .9306. Thus, subjects demonstrated a high level of rating reliability.

In addition, this phase served to reproduce the environment typically found in a holistic grading session. Thus, the researcher was able to achieve a testing atmosphere in which subjects would enter the final part of the experiment unaware of the experiment's purpose.

Once a typical holistic grading environment was simulated to the researcher's satisfaction, the final part of the experiment began. In it, subjects were directed to read and rate one of the two texts in two conditions of task awareness and then generate a written recall protocol based on the content of the text they had just read. The procedure was implemented in such a way that each subject recalled both texts, with each text being rated and recalled by an equal number of subjects for each condition of task awareness.

The dependent variable of this study was comprehension. This variable was operationalized as the amount of content correctly recalled on a written recall protocol. The amount of content recalled was computed on the basis of a weighted propositional count of the idea units contained within the two respective texts read, rated, and recalled.

The variable of primary interest, level of assessed proficiency, consisted of two student-generated
compositions pre-judged to be qualitatively different from one another. These two texts, which were selected and rated using the same procedure as was employed for the nine compositions used earlier in the experiment, were rated as Higher and Lower. The Lower text represented a typical sample of writing done by a Level 107 (Intermediate) student in the Ohio State University's ESL Composition program, while the Higher text was typical of work done by a Level 108 (Advanced) ESL student. As both texts were assessed an identical rating by the three independent readers employed, the researcher could proceed with a high level of confidence that the texts differed sufficiently in quality to allow one to be designated as Higher and one Lower during a holistic reading task.

The texts, themselves, addressed different topics. The Lower text concerned the question, "Which is the best preparation for life: learning to be cooperative or competitive?" while the Higher text responded to the question, "Which is better: act quickly and decisively or wait and think it over carefully?" Compositions addressing different topics were selected so as to minimize the possibility of the specifics contained in one being confused with specifics contained in the other.

Both texts were comparable in length, with the Lower one containing 272 words and the Higher one containing
329 words. The texts were judged to be sufficiently long to prevent subjects memorizing a substantial portion of either (Bernhardt, 1986b).

In addition to the level of assessed proficiency, one further independent variable was investigated: reader orientation. This variable consisted of two task conditions under which the texts were read, rated, and recalled. The first condition was termed Naive. In it, subjects were not oriented to the written recall task prior to reading a text. The second condition was termed Focused. In this condition, subjects were informed before reading a text that they would be expected to generate a written recall protocol in addition to rating the text holistically.

Procedurally, the Naive and Focused tasks were administered using a variation of a repeated measurements design. Specifically, for the Naive task, half of the raters were given the Lower text to rate, while the other half were given the Higher text. Upon completion of the Naive task, the order of text distribution was reversed, so that for the Focused task, subjects read the other of the two texts. By using this procedure, the researcher was able to reinforce the Focused level of task awareness while at the same time maximizing the number of recall protocols generated per condition.
Data analysis was performed by means of a 2 X 2 factorial analysis of variance, followed by a series of one-way analyses of variance. A 2 X 2 factorial design was chosen for the initial data analysis because of its capacity to examine both the main effects and interactions of reader orientation to task and level of assessed proficiency as they related to comprehension. Followup procedures, consisting of a series of one-way analyses of variance, allowed the researcher access to differences in individual cell means, which in turn provided data of relevance to several of the research hypotheses posed by the study.

**Summary of Findings**

Interaction between level of assessed proficiency and reader orientation was not statistically significant. Accordingly, the researcher proceeded to examine the two main effects.

Analysis of the experimental data revealed a statistically significant main effect for reader orientation ($F(1,20) = 8.56, p<.01$). Level of assessed proficiency, however, was not a statistically significant main effect. In other words, Focused subjects recalled significantly more than Naive subjects of what they had read overall, but readers of the higher-quality text did
not recall significantly more than readers of the lower-quality text.

Examination of individual cell means in a series of pair-wise comparisons revealed two significant differences between means. The first, of assessed proficiency across the Naive level of reader orientation, indicated that Naive subjects recalled more of the higher-rated text than of the lower-rated text. The second significant difference in means occurred between reader orientation and the Lower level of assessed proficiency. This difference pointed to the conclusion that Focused readers of the lower-rated text recalled significantly more than their Naive counterparts. The remaining two pair-wise comparisons in the 2 X 2 matrix revealed no statistically significant differences between means.

**Conclusions**

Below is a summary of the findings of this study as they relate to the research questions set forth at the end of Chapter 2. Each of the six research hypotheses posed by this study is delineated, pertinent results are systematically presented, and conclusions are discussed.
Question I: Will naive holistic readers recall significantly more of a higher rated text than of a lower rated text?

This research question was the primary focus of the study, in that it addressed the issue of the construct validity of a meaning-based model of holistic assessment of second language writing proficiency. The researcher hypothesized that if holistic evaluators of writing proficiency attend to meaning during a rating task, they should recall significantly more of a text generated by a more-proficient NNS author than of a text generated by a less-proficient NNS author.

The data from the experiment indicate that holistic raters reading for the purpose of assessing writing proficiency comprehended significantly more of the better written (i.e., Higher) text, when comprehension was operationalized as the amount of content correctly recalled on a written protocol. Specifically, when performing in the Naive condition, subjects reading the Higher text recalled significantly more than Naive subjects reading the Lower text (F(1,10) = 5.12, p < .05). Based on the theoretical literature pertaining to the organization of content in memory, these experimental results lead to the conclusion that comprehension is facilitated by the level of quality of the text, even when the text is read for the purpose of holistically
assessing writing proficiency rather than for the purpose of comprehension, per se.

On the basis of the results of this part of the experiment, a complex view of the manner in which a reader approaches a holistic evaluation task emerges. In particular, those elements of the holistic reading process that emphasize rapid reading to form a general qualitative impression and the use of rangefinder samples to establish qualitative scoring parameters seem to promote what deBeaugrande (1978) terms a "sampling process." In this procedure, the sample is restricted by a process that accepts certain portions of presented input while rejecting others. These elements of holistic scoring, however, do not appear to deter — and in fact may even facilitate — the process by which a holistic reader apprehends meaning. Thus, the holistic reader's depth of processing does not appear to be materially affected by the time constraints and use of pre-established rating parameters incorporated into the holistic rating procedure.

Indeed, the term "time constraints," as it is applied to holistic reading, seems not to be the limiting factor that it is in such specialized reading tasks as speed reading, skimming, and scanning, where research has revealed an inverse relationship between the rate at which a text is read and how well it is comprehended when
rate of speed is artificially controlled. Instead, in the case of holistic reading, "time constraints" refers more to internal factors influencing the time spent reading a text than to external factors such as the scrolling of text across the screen of a computer monitor.

In short, the statistically significant recall difference in favor of Naive readers of the higher-rated text over Naive readers of the lower-rated text points to the following interpretation: Naive holistic readers comprehended more of the text generated by a more-highly proficient NNS author than of the text generated by a less-proficient NNS author. Given that all experimental conditions save level of assessed proficiency of texts were held constant, we may conclude that comprehension plays a significant role in the holistic assessment of the writing proficiency of NNS's by experienced instructors of ESL composition.

Question II: Will focused readers recall significantly more of a higher rated text than of a lower rated text?

While it was expected that Focused subjects in this phase of the experiment would perform in a comparable manner as the Naive subjects considered in Question I, results indicated that they did not. Contrary to the
consensus of findings contained within the considerable research base on the effect of organization of text structure on recall, in this experiment, Focused readers of the higher-rated text did not recall significantly more of its content than they recalled of the content of the lower-rated text. In fact, descriptive statistics reveal that the Focused/Lower raters recalled slightly more than the Focused/Higher raters ($\bar{x}_l = 53.63$, S.D. = 14.07; $\bar{x}_h = 50.50$, S.D. = 14.72).

The reason for this apparent anomaly is not clear, although one possibility for the lack of significant difference between mean recall scores of the Focused/Lower and Focused/Higher groups is that there exists a ceiling effect on how well a message can be recalled by subjects who are oriented toward reading for comprehension. Such a ceiling effect might serve to depress the total recall scores of Focused readers, thereby obscuring potentially significant differences between recall of higher-rated texts and lower-rated texts. The research literature pertaining to recall protocols, while extensive, offers no clues to the resolution of this problem because empirical studies have, to date, either concentrated almost exclusively on the recall of non-native speakers, employed artificially-altered versions of one text, or analyzed the recall product in order to identify and examine
rhetorical relationships. Moreover, no recall study has yet sought to establish an upward limit on recall protocol scores proper.

Other possible explanations for this result include the nature of the texts, themselves, and the characteristics of the raters whose recall scores comprise the sample means. In the former case, it must be kept in mind that the texts addressed different topics. It is conceivable that the topic addressed in the Lower text was, either inherently or in the way it was developed, somehow more "memorable" or otherwise distinctive in the minds of the raters who read it than the topic of the Higher text. In the latter case, it is possible that the Focused/Lower group performed better than the Focused/Higher group on the recall protocol task. Such an inherently biased cell sample could well have been drawn from the small sample population of subjects for the study.

While either or both of these scenarios are possible, it should be recalled that this experiment employed a variation of the repeated measurements design. Thus, the presence of any such topic-related or subject-related confounding variable should be reflected in a significant interaction effect between levels of assessed proficiency and reader orientation. As was noted in Chapter 3,
however, no such significant interaction effect was achieved.

One final possibility is that the raters' depth of processing was enhanced when they were oriented toward reading both texts for the purpose of recalling them. That, coupled with the possibility that qualitative differences between the Higher and Lower texts were not sufficiently marked to play a significant role in how thoroughly the Lower text could be processed in comparison to the Higher text, could have resulted in both the Higher and Lower texts being processed equally well.

**Question III:** Will focused readers of a higher rated text recall significantly more than their naive counterparts?

The data obtained in this phase of the experiment ran counter to the considerable amount of empirical evidence concerning the impact of task awareness on comprehension. Instead, the expectation that Focused readers would recall significantly more than Naive readers was not borne out. The results pertaining to Question III are particularly surprising in light of the fact that results obtained in Questions IV and VI tended to support the position that attention to task was a
significant factor in reading comprehension (see below for details).

The descriptive data from this phase of the study indicate that the Focused readers of the higher-rated text recalled more of its content than did the Naive readers ($\bar{x}_{fh} = 50.50$, S.D. = 14.72; $\bar{x}_{nh} = 44.17$, S.D. = 15.86). A one-way analysis of variance of the two cell means, however, showed the difference, while in the expected direction, was not statistically significant.

The reason for this result is not clear, although it is possible that the fact that Focused readers were oriented toward both an assessment task and a recall task, while Naive readers were oriented only toward an assessment task, may have contributed to the obtained results. It is possible that by being oriented toward two tasks, Focused readers were limited in the amount of processing capacity they could direct toward recall.

Another possibility is the ceiling effect discussed in Question II, in this instance influencing how well a passage can be recalled, regardless of reader orientation. It is plausible that given the length of the passage, perhaps in conjunction with such other variables as the above-mentioned nature of the reader orientation and subject fatigue, the mean recall score obtained by Focused readers approached the maximum that could be expected under the circumstances. As there is no
empirical evidence concerning the impact on recall exerted by the orientation of native-speaking readers toward multiple tasks, however, all that can be offered at this point is conjecture.

**Question IV:** Will focused readers of a lower rated text recall significantly more of its content than their naive counterparts?

The results of this phase of the experiment support the research base in the organization of content in memory, which indicates that task awareness has a facilitating effect on recall. As expected, Focused subjects recalled more than Naive subjects when reading a text generated by a less-proficient NNS author ($F(1,10) = 14.84$, $p < .005$).

The descriptive data relating to this question are interesting as well. Of all the differences between means yielding statistical significance, the one between Naive/Lower ($\bar{x} = 27.28$, S.D. = 9.09) and Focused/Lower ($\bar{x} = 53.63$, S.D. = 14.07) was the widest.

Taken by itself, this result is hardly unexpected. After all, the facilitating effect of both task awareness and reader's goals on depth of processing are relatively well-documented. What is puzzling, however, is the fact that the statistically significant difference in recall
between Focused/Lower and Naive/Lower readers was not duplicated by the Focused and Naive readers of the Higher texts (see Question II). Apart from the possible factors mentioned in conjunction with the discussion of Questions II and III, the researcher is unable to offer a plausible explanation as to why this is so.

**Question V:** Will focused readers recall significantly more of both the higher and lower rated texts than their naive counterparts?

Question II asked if Focused readers recalled more when processing a higher-quality text. Question IV posed the same question for Naive and Focused readers of a lower-quality text. This research question sought to determine if the total amount of content recalled across levels of assessed proficiency was significantly higher for Focused readers than for Naive readers.

Results of a two-way analysis of variance revealed a significant main effect for Reader Orientation across levels of Assessed Proficiency ($F(1,20) = 8.56, p < .01$). Descriptive statistics showed that overall, Focused subjects recalled more of both texts than Naive subjects ($\bar{x}_F = 52.07$, S.D. = 13.82; $\bar{x}_N = 35.72$, S.D. = 15.16).

The results of the experiment point to a conclusion that is generally supported by empirical research in the
field of organization of content in memory as it relates to depth of processing; namely, that awareness of task—specifically, recent exposure to a recall task that was undertaken without benefit of prior orientation, coupled with explicit warning that the next reading task would involve recalling the content of a text—significantly facilitates recall of textual content. That is to say, in the context of this experiment, Focused readers recalled significantly more of the total content contained within the Higher and Lower texts than Naive readers.

It is noteworthy that statistical significance at the .01 level was achieved despite the lack of significance between mean recall scores of Focused and Naive readers of the Higher text. This result, coupled with the fact that Focused subjects at all three levels recalled more (albeit to a degree that failed to reach statistical significance at the Higher text level) than did Naive subjects, leads the researcher to conclude that reader orientation toward task demands facilitates recall of texts written by non-native speakers of English.
Question VI: Will readers of higher rated texts, regardless of orientation, recall significantly more than readers of lower rated texts?

While Question V was concerned with the impact of the main effect of Reader Orientation upon recall of prose written by NNSs, this question dealt with the second main effect considered in the data analysis, Levels of Assessed Proficiency. The results of this phase of the experiment do not support the data base in both L1 and L2 research which indicates that readers of better-written texts should recall significantly more than readers of less well-written texts, regardless of type of reader orientation or level of task awareness. Consequently, it must be concluded that despite the superior level of recall evidenced by readers of the Higher text ($\bar{x}_h = 47.34$, S.D. = 14.96; $\bar{x}_l = 40.45$, S.D. = 17.80), when Reader Orientation is not considered there is no significant difference between the amount of content recalled by readers of a text generated by a more proficient NNS author and the amount of content recalled by readers of a text written by a less proficient NNS author.

In summary, the results obtained from the experimental data indicate that levels of assessed
proficiency is a comprehension variable in the case where it is applied to naive readers engaged in the holistic assessment of second language writing. Empirical evidence indicates that Naive holistic evaluators of a text generated by a relatively proficient NNS author recalled significantly more of that text than did their Naive counterparts who read a text authored by a less proficient NNS. Contrary to expectations, however, levels of assessed proficiency does not act as a comprehension variable when the level of reader orientation is Focused. Instead, Focused readers of both Higher-rated and Lower-rated texts were found to recall virtually the same amount of content.

Further results indicated that type of reader orientation is a comprehension variable when applied to the reading of texts judged to be written by less highly proficient NNS authors. Evidence shows that Focused readers of the Lower-rated text recalled significantly more of that text than Naive readers of the same text. No such generalization can be extended to experimental subjects reading the Higher-rated text, however. In that case, there was no significant difference in recall between Focused and Naive readers of the Higher text.

Examination of the two main effects posited in this study, Levels of Assessed Proficiency and Reader Orientation, indicated that only Reader Orientation was a
significant main effect, signaling the fact that, overall, Focused subjects reading both the Higher and Lower texts recalled significantly more of the total content than Naive readers.

**Limitations**

The texts used in this study to measure the reading comprehension of holistic raters of L2 writing proficiency addressed two different topics. Although no clear evidence exists that the use of texts responding to different topics was a confounding variable, it does raise questions concerning how comparable the texts were insofar as they were recalled by the subjects.

For example, the lower-rated text discussed the relative merits of cooperation and competition. The author of this text chose to develop the topic in a conventional way, using as examples of cooperative and competitive societies Japan and the United States, respectively. The author of the higher-rated text, on the other hand, used a much more ideosyncratic approach to developing the topic of the relative merits of quick action and deliberation. The examples this author used to develop his/her argument were not of such a global nature as those used in the other essay.

Because those subjects responding to the lower-rated text were faced with a more conventionally developed set
of supporting examples, it is possible that they could have drawn upon a combination of world knowledge and logical inference to either facilitate recall or construct a plausible—but not necessarily recalled—scenario upon which to base a recall protocol. For example, had a subject recalled that Japan and the United States were used as examples of the two types of society in question, attaching the appropriate "cooperative" or "competitive" label would not have been especially difficult. In much the same way, had the connection between the United States and competition been recalled, Japan and cooperation could easily have been deduced.

No such transparent logical connection between specific examples existed for the "Act Quickly or Wait?" text, however. Readers of this text, instead, had to rely much more on "bottom-up" processing to instantiate the appropriate schema. Hence, their mean recall score may have been affected accordingly, especially when reading in the Naive condition. How—or even if—such differences in interaction between specific textual content and individual schemata affected subjects' recall, however, is a question that could not be answered in this study.

Another limitation of this study lay in the choice of recall measure used. Although the written recall protocol is widely employed, persistent questions remain
concerning its validity as a means by which to measure reading comprehension. In particular, it has been observed that a recall protocol not only measures what has been recalled, per se, but also measures, at least in part, the ability of subjects to write down what they have recalled from the text. Nevertheless, in the researcher's judgment, the use of a written recall protocol provided this experiment with an instrument that is superior to objective measures such as multiple choice, matching, identification, or true/false tests of recall.

Finally, the matter of how subjects perceived the recall task was of some concern to the researcher. Specifically, the researcher noted a wide range in the extent to which examples were developed and details attended to in individual recall protocols. Despite specific instructions on what to include in a recall protocol, the researcher was left to wonder if the paucity of detailed content contained in some protocols, including a few written in the Focused condition, was more a case of subjects trying to present a summary or overview of what they had read than of an accurate barometer of what they had, indeed, recalled. Given the nature of the study, however, it is obvious that subjects could not have been given prior orientation to and training in written recall protocols.
Recommendations for Further Research

With respect to the primary focus of this dissertation, which was to investigate the construct validity of holistic scoring as a means to assess writing proficiency in non-native speakers, several avenues of inquiry are apparent. The first involves replication of this experiment using a larger sample. In so doing, mean recall scores could be obtained that were not so susceptible to influence by an abnormally high or low score.

It would also be advisable to replicate this study using a more homogeneous corpus of texts to be rated and recalled. By so doing, the possible influence of increased attention to textual features on the recall of qualitatively equivalent texts could be investigated. For example, if a sample of texts could be obtained from within one grade level instead of several, we could begin to make some determinations concerning the sensitivity of recall protocols in measuring comprehension of texts for which qualitative assessments are much more subtle.

In the same way, replication using a more heterogeneous corpus of texts could be performed. In this instance, however, such a study would explore the premise that the greater the qualitative difference between texts, the less the reader attends to meaning.
One other variation would involve holistically scoring NNS compositions on a more detailed scale than the four-point version used in this study. In so doing, a more sensitive measurement of the relationship between rater reliability and recall could be obtained.

Further research is also recommended to test the hypothesis of a "ceiling effect" among Focused readers with respect to performance on a recall protocol. If such an effect is found to exist, research into its characteristics and limits would then be appropriate.

Finally, the effects of multiple task demands on a reader's depth of processing should be studied. For example, a study comparing the interrater reliability of Focused and Naive subjects engaged in holistic scoring tasks could be executed to determine if the added task demands imposed upon Focused readers adversely affects scoring reliability.
LIST OF REFERENCES


Appendix A

Diagnostic Essay Questions (Alternate Forms)

Name_____________________________ Diagnostic Theme
Instructor_________________________ Autumn Quarter
Course___________ Time___________

You have the entire class period to write a well-organized, well-developed essay on the topic given below. Take five minutes or so to plan your composition before you begin to write, and allow at least five minutes at the end for making corrections.

Form A

Some people say that the best preparation for life is learning to be cooperative. Others say that learning to be competitive is the best preparation. Discuss these positions, using examples of both. Tell which you agree with and why.

Form B

Many people believe that it is better to act quickly and decisively than to wait and think something over carefully, because you may lose the opportunity by waiting. Do you think that taking quick, decisive action is or is not advisable? Explain your position.
Appendix B

Protocol Scoring Matrices for Higher (FORM A) and Lower (FORM B) Texts

FORM A, Text #2

<table>
<thead>
<tr>
<th>WEIGHTED SCORES</th>
<th>IDEA UNITS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. 3.33</td>
<td>I think (OPINION)</td>
</tr>
<tr>
<td>2. 5</td>
<td>&quot; &quot; the question of quick action versus deliberation is a problem.</td>
</tr>
<tr>
<td>3. 4</td>
<td>The problem is rather controversial.</td>
</tr>
<tr>
<td>4. 3.33</td>
<td>It is hard to say one is definitely better than the other...</td>
</tr>
<tr>
<td>5. 3.67</td>
<td>Since (Because)</td>
</tr>
<tr>
<td>6. 6.33</td>
<td>The choice between alternatives may depend on the importance of the situation.</td>
</tr>
<tr>
<td>7. 6</td>
<td>Sometimes taking quick and decisive action is more desirable.</td>
</tr>
<tr>
<td>8. 4</td>
<td>For instance (EXAMPLE #1)</td>
</tr>
<tr>
<td>9. 5.67</td>
<td>You have a disease.</td>
</tr>
<tr>
<td>10. 5.33</td>
<td>The disease is very serious.</td>
</tr>
<tr>
<td>11. 4.33</td>
<td>Your doctor told you.</td>
</tr>
<tr>
<td>12. 5.33</td>
<td>&quot; &quot; told you THAT YOU NEED AN OPERATION.</td>
</tr>
<tr>
<td>13. 5.33</td>
<td>The operation is dangerous.</td>
</tr>
<tr>
<td>14. 4.67</td>
<td>(BUT)</td>
</tr>
<tr>
<td>15. 6</td>
<td>... You may die without it.</td>
</tr>
<tr>
<td>16. 5.33</td>
<td>In this case, it would be very hard to make a quick decision.</td>
</tr>
<tr>
<td>17. 3.67</td>
<td>But</td>
</tr>
</tbody>
</table>

151
18. 4.33  You should do so...
19. 4  Since... (Because)
20. 4.67  ... the matter is very urgent.
21. 4.33  For another example (EXAMPLE #2).
22. 5  I (the author) go to a restaurant...
23. 3.33  ... with some of my friends...
24. 6.33  ... and we only have 20 minutes to eat
25. 4.67  Because...
26. 5  ... we are going to a movie afterwards.
27. 6  In this situation, I have to make a quick decision.
28. 3.67  The decision is about what to order.
29. 2.33  The decision is FOR MY OWN SAKE.
30. 4  Otherwise (if I don't make a quick decision)...
31. 3.67  I may not be able to finish my meal...
32. 3  ... or
33. 4  ... I may miss part of the movie.
34. 3.33  (ON THE OTHER HAND)
35. 5.67  There are cases when it is much better to think something over carefully...
36. 4.33  " " carefully RATHER THAN
37. 4.67  ... making a hasty choice.
38. 5.33  It is more advisable to be careful...
39. 6.33  " " careful WHEN YOU HAVE TO MAKE A VERY IMPORTANT DECISION IN YOUR LIFE.
40. 4.33  For example (EXAMPLE #1)
41. You should take your time to think it over...
42. "" think it over WHEN YOU DECIDE
43. "" decide WHAT YOU'LL STUDY IN COLLEGE...
44. Since (Because)
45. Your major will be related to your career.
46. For another example (EXAMPLE #2)
47. Take your time...
48. ... when you choose your spouse.
49. You have to live with that kind of decision...
50. "" decision ALL YOUR LIFE...
51. ... and you cannot change (that decision)...
52. ... even though (if) you want to.
53. In conclusion...
54. ... being careful
55. ... and taking enough time
56. ... time IS PREFERABLE
57. ... if
58. ... the issue is not very urgent
59. "" urgent AND (action) DOESN'T NEED TO BE DONE (taken)
60. ... right away (immediately)
61. ... because
62. ... you'll probably make fewer mistakes
<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>63. 4</td>
<td>&quot;&quot; make fewer mistakes AND REPENT LESS</td>
</tr>
<tr>
<td>64. 3.33</td>
<td>... if</td>
</tr>
<tr>
<td>65. 3.67</td>
<td>if YOU DO SO (take your time).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>295.65</td>
</tr>
<tr>
<td>WEIGHTED SCORES</td>
<td>IDEA UNITS</td>
</tr>
<tr>
<td>-----------------</td>
<td>------------</td>
</tr>
<tr>
<td>1. 4.67</td>
<td>It is necessary to consider a few factors.</td>
</tr>
<tr>
<td>2. 4.33</td>
<td>&quot; &quot; a few factors WHICH ARE IMPORTANT.</td>
</tr>
<tr>
<td>3. 3</td>
<td>In order to give an answer to the question...</td>
</tr>
<tr>
<td>4. 4.67</td>
<td>&quot; &quot; of whether it is better to learn to be competitive...</td>
</tr>
<tr>
<td>5. 4.67</td>
<td>... or cooperative.</td>
</tr>
<tr>
<td>6. 6</td>
<td>In which society do the people live?</td>
</tr>
<tr>
<td>7. 6</td>
<td>How is it organized?</td>
</tr>
<tr>
<td>8. 3.67</td>
<td>(AND)</td>
</tr>
<tr>
<td>9. 6.33</td>
<td>The personality of the individual.</td>
</tr>
<tr>
<td>10. 5.67</td>
<td>Individual abilities.</td>
</tr>
<tr>
<td>11. 2.33</td>
<td>Everyone will agree with the statement...</td>
</tr>
<tr>
<td>12. 4.33</td>
<td>&quot; &quot; the statement THAT IT IS IMPORTANT TO LEARN BOTH OF THESE ABILITIES (cooperation and competitiveness).</td>
</tr>
<tr>
<td>13. 2</td>
<td>Everyone will agree...</td>
</tr>
<tr>
<td>14. 3.67</td>
<td>&quot; &quot; agree THAT EVERY SOCIETY HAS DEVELOPED ITS OWN WAY...</td>
</tr>
<tr>
<td>15. 3.33</td>
<td>&quot; &quot; its own way TO SELECT PEOPLE...</td>
</tr>
<tr>
<td>16. 3</td>
<td>&quot; &quot; to select people AT SCHOOL...</td>
</tr>
<tr>
<td>17. 3</td>
<td>&quot; &quot; IN BUSINESS...</td>
</tr>
<tr>
<td>18. 2.33</td>
<td>&quot; &quot; EVERYWHERE.</td>
</tr>
<tr>
<td>19. 5.33</td>
<td>There are differences between the social aims...</td>
</tr>
</tbody>
</table>
20. 5.33 "" aims WHICH INFLUENCE INDIVIDUAL AIMS.

21. 5.67 In America...

22. 6 organization of society is much more competitively oriented...

23. 4.67 "" oriented THAN IN SOME STATES IN THE EASTERN HEMISPHERE.

24. 1.33 As far as I know...

25. 6 In Japan...

26. 6.33 "" the people are much more oriented in a cooperative style of working...

27. 5.33 "" of working AND LIVING,

28. 4.67 but

29. 6.33 they are very competitive as a nation or

30. 3.33 as a company –

31. 5.33 whenever they function as a group.

32. 6 There are reasons for these different kinds of social organizations.

33. 4.67 These reasons can be found in the historical development...

34. 5 "" and in the conditions of living.

35. 5.67 Japan has been isolated from outside influence...

36. 5.67 "" influence FOR MANY YEARS...

37. 4 "" and has had little space for too many people...

38. 5.67 "" people FOR MANY YEARS.

39. 3.67 (CONSEQUENTLY)
41. 6  Japan was forced to develop a very cooperative style of living.

42. 4.67  In contrast...

43. 5.67  The development of the USA was determined by individual activities.

44. 5  So (consequently)

45. 6  Americans developed a more competitive style of social organization.

46. 4.33  It is important to see these historical and social criteria...

47. 4  "" criteria BEFORE ANSWERING THE QUESTION OF WHETHER IT IS BETTER TO LEARN TO BE COMPETITIVE OR COOPERATIVE.

48. 4  No individual person can be seen as isolated from...

49. 3.67  "" from THE SOCIETY...

50. 2.33  "" society WHICH HE/SHE BELONGS TO.

TOTAL

229.34
Appendix C
Instructions Concerning Criteria for Collecting Sample Texts

Dear Colleague:

I would like to ask your help in collecting data for my dissertation. Specifically, I need to examine writing samples obtained from students who are newcomers to the ESL Composition program. What must be done involves the following steps:

1. After the Diagnostic test is administered and graded, randomly select three compositions from among those submitted by your class.

2. Pass on these sample compositions to me as soon as you have graded them. I will photocopy them as I receive them and return them to you as quickly as possible.

3. Do not include compositions scored as "move-ups" or written by students who are not entering the program for the first time.

4. In addition to the three randomly-selected compositions, include any compositions scored as "Q."

Please make sure that the three compositions are selected randomly. A limited amount of followup involving a few students may be required, so include both your name and student's name on each sample. Naturally, subject confidentiality will be respected.

158
To summarize, from each instructor I need three randomly-selected compositions written by students in his/her class whose initial placement into the ESL Composition sequence has been confirmed through the Diagnostic test. I also need any Diagnostic tests scored as "Q." Samples are selected from among those written by newcomers to the program only.

I'd like to thank you in advance for your cooperation in this task. Quite literally, I couldn't do it without you! If you have any questions or comments concerning what this is all about, please let me know.

Again, many thanks!

Mike Janopoulos
Appendix D

Original Texts (Levels 107 & 108)

Composition #1, Level 107, Form B

For giving an answer to the question if it's better to learn to be competitive or cooperative it is necessary to consider a few factors which are important.

- In which society do the people live and how it is organized.

- The personality of the individual and the individual abilities.

Everyone will agree with the statement that it is important to learn both of these abilities, to be competitive and cooperative.

And everyone will agree that every society has developed its own way to select people at school in business everywhere.

There are differences between the social aims with an influence on the individual aims. In America the organization of the society is much more competitive oriented as in some states of the eastern hemisphere. And as far as I know are the people in Japan much more oriented in a cooperative style of working and living but they are very competitive as a nation or as a company - as every kind of a group.
The reasons for these different kinds of social organization can be found in the historical development and the conditions of living.

Japan, for many years isolated from any influence and with little space for too many people was forced to develop a very cooperative style of living. The development of the USA was determined by individual activities. So they developed a more competitive style of social organization.

It is important to see these historical and social criteria before answering the question if it's better to learn a competitive or cooperative style of living. No individual person can be seen isolated from the society to which it belongs.
I think this is a rather controversial problem. It is difficult for me to say that one is definitely better than the other since it may depend on the importance of the matter of the situation.

There are times when taking a quick and decisive action is more desirable. For instance, you got a very serious disease and your doctor told you that you need an operation which is very dangerous. You were also told that you might die if you don't have the operation. In such a case, it would be very hard to make a quick decision, but you should do so since the matter is very urgent. For another example, I go to a restaurant with some of my friends and we are supposed to have only 30 minutes in that restaurant because we are going to a movie afterwards. In that particular situation, I have to make a quick decision for my own sake. Otherwise, I may not be able to finish my meal or miss some part of the movie.

There are many other cases when it is much better to think something over carefully, rather than to make a hasty choice. It is more advisable to be careful especially when you have to make a very important decision in your life. For example, when you decide what you'll study in college, you should take time and think over and over since your major will be related to your career.
For another example, I can say when you choose your spouse. That kind of decision is what you have to live with it all your life and what you can not change easily even though you want to.

In conclusion, I would say that being careful and taking enough time is more preferrable if the issue is not very urgent and doesn't need to be done right away because you'll probably make less mistakes and repent less if you do so.
Appendix E

Calibration Samples

Every human being lives in a society which is consisted with many persons. To live with other persons pursuing each of the individual's happy life and the development of a society, we need to live with other persons. To learn to be cooperative or to be competitive would be the two ways of preparation for life.

Every man and woman has a tendency to live his or her life with a short-term view and a egoism, because he or she lives within the limit of time and space. If we see the life with this view of life, competitive way seems to be superior to cooperative one for a man or woman to get his or her happiness in a society. Let me give you an example:

To be a best athlete in a certain baseball team, every member of the team trains hard. But to overcome all the hard trainings is not easy to each of the members. Sometimes, it seems more easy for an athlete to disturb the competitive athlete than to overcome oneself. It may be a way of competitive view of life. Every members of this team trys to do one's best by one's own way without concerning the team-work. Then the team is getting weaken. If every team of baseball play the game too much comparatively, the play has a tendency to lost fair play.
Then the baseball game will be getting lost the pan.

I do not mean that the competitive view of life is not good. This view of life is necessary. Without this view of life, human beings lost motivation to improve his or her quality of life. I mean that competitive way of life should be alway based on the cooperative way of life, in order to build a more good society.
The best preparation for life is based on the values and perspectives on the life. Sometimes, culture and personality are crucial factor in planning life. Which is more profitable of the two, cooperation and competition, can be determined by this logic.

For example, competition can be emphasized in the society where individualistic values and public choice logic. This society makes rewards to those who win in the competition. I think that industrialized and rationality-oriented societies, such as U.S.A. and France, are in this category. Therefore, in order to be successful, individuals should gain the skill which brings the victory in the competition.

But, social life, if competition-oriented or not, always needs mechanism by which social conflicts are arranged.

Competition is more productive in several realm. So competitive atmosphere is encouraged. But extreme competition without cooperation will cause a breakdown of social system.

In the other hand, cooperation is also important but should not overvalued in sane society.

Cooperation is useful when it can achieve a things which individual can not do. In scientific discovery, individual can do a great achievement. But in building a
107, continued

large museum, individual can not do a great contribution. Cooperation among men is required.

In summary, competition and cooperation are not to be contradictory. They should supplement each other in the healthy, more productive and harmonious society. Therefore, social man has to learn not only competition but also cooperation with others.

A healthy human education requires a sense of balance. It will make good citizens for democratic society.
What's life? Life is a long period in which one enjoys the fruit of achievement and in turns one suffers from the wounds of defeat. The preparation for life ahead is certainly helpful to obtain the joys of life and avoid the pains. Both learning to be cooperative and to be competitive are key issues for **life preparation**.

Most of one's life will be spent working, communicating with people, earning life grades. Team work is always regarded the best way to reach a group goal, in a sense this is also a personal goal because one belongs to the group. For instance, a auto worker in a assembly life has to learn how to cooperate with other colleagues since the production performance is a total of individual output. If there is anyone in the assembly lines trying to do things by his or her own rules and being regardless of group requirements, the group output will tend to decrease, and in a longer period, total performance will surely go down. And the individual will be accused of

But does cooperation mean non-individualism? Being cooperative is not to avoid the presentation of one's talent. In order to be successful in one's life, the fully utilization of talents and abilities is **essential**. One has to be competitive to give himself (or herself) a best chance which will consequently win him or her the fruit of victory. Otherwise, human history would not
progress if there is no competition between people, and life would be less abundant.

One can find a mean way between being cooperative and competitive. It is to play strictly by the group agreement and at the same time, present one's superior sides as possible as one can. A mean way will be the best way for the preparation of life.
In the rapidly changing world, I cannot help recognizing that we are required to act quickly and decisively in many areas. But on the other hand it should be stressed that something important and irreplaceable is lost to us because of our impatience. I think Korea's rapid economic growth is a good example of such cases.

The latter half of the twentieth century has been called the age of economic development. During this period, Korea faced the task of building a new nation on the ruins left behind after the Korean War ended. The postwar economic situation, political uncertainty and various upheavals darkened the face of Korea for a long time. Since the early 1960's, however, that dark image began to change. Political innovation and rapid economic development through industrialization brought about remarkable economic growth, spearheaded by the quick expansion of exports. The start was late, but the achievement through the 60's and the early 70's was outstanding among the developing nations of the world.

But the radical and concentrated push for the economic development has resulted in the emotional instability in the minds of Korean people. If we now ask "Today is there any sense of confidence common to all the people of Korea?", something negative is immediately obvious. In this situation Korea can't create a foundation
Q, continued

for further development. Nowadays the problem of quantity has become a matter of serious concern.

In summary real world needs a person who is responsible and deliberative. And I take an optimistic view of the future world because I believe the society knows what it should esteem and what it should despise.
Appendix F

Practice Texts (Levels 106-Q)

106

I think that learning to be cooperative and learning to be competitive are all the best preparation for life. The society is complex. It is composed of many kinds of people and career. You need to cooperate with others. The society is cruel too. It is full of competition and fight. You can't live well without the ability to compete.

We contact with many people everyday. We can't do everything all by ourself. So we have family, friends, classmates, bosses, and etc. We must learn to live well with them. Without their cooperations I believe we will not be successful. That's why we must learn to be cooperative.

There are many competitions in our lifes too. To pass examinations, get degrees, take jobs... All need to compete with others and even with you. If you want to be strong then you must learn how to compete.

We learn at hom and study at school. From child to adult. Our parents and teachers teach me every. Both to be cooperative and to be competitive. The purpose is just to make the best preparation for our future life.
There are several ways to prepare for life. Among them, two best preparation for life have been dominated. One is learning to be cooperative and another one is learning to be competitive.

First, to be cooperative means that people works together. Under the circumstance they exchange their idea, help others and be helped by others.

In the other hand, to be competitive means that people works personally with rivalism. They always want to be more advanced then others.

In my opinion, the first one is more positive way. As most people can realize, the present world is influenced by individualism which is resulted by the excess competitism. This has been the seeds of not only personal but also international discord.

When people cooperate well each other, the competition of the proper level can naturally occur with the satisfaction. Good ideas can be stimulated under peaceful and comfortable situation.

For example, if the USA and Soviet Union cooperate well, the future world could be peaceful.

Therefore, people should learn how to cooperate each other. It is the best way to develop the industry and the personal relationship.
It seems that it is better to act quickly and decisively than to wait and think something over carefully, for you may lose the opportunity by waiting. However, this question is not so simple enough for you to jump to the conclusion. There are times when you need to take into account every factor that might be involved in the situation you consider; still, there are times when you had better make your decisions and take actions quickly.

As might be expected, it is not so easy to take decisive actions. Since you don't acquire as many possible informations as you need, you are not sure about what outcome of the situation would be if you act quickly. Therefore, you may think that it is best for you to spend some time in collecting informations as needed. The moment you have all the informations necessary to make a decision, you don't wait and take quick actions. Thus, the point is whether you have sufficient reasons to make up your mind.

Also, there exists a situation that you have to act decisively and quickly. That is when you face a opportunity that will disappear if you don't catch it intermediately. This opportunity may be some kind of chance to make a vast fortune or take a promising job. If you take actions, and go ahead of other pepole, you will get whatever surprises you. To make that come true, you have no other choice but act quickly. Therefore, it seems
108, continued

that it depend on what kind of situation you face when you are hesitant about taking further steps. It is absolutely not a simple case as you might think.
How should people get prepared for life? Which is the best preparation for life? Some people say it is learning to be cooperative. Others say it is learning to be competitive. There are arguments now let us analyze the argument carefully.

At first, let us consider 'cooperative.' With the progress of society, more and more work requires the involvement of many people. People should be cooperative in their work. How could a man-made satellite be turned out if people had not cooperated well? What about a software system, an atomic bomb, even a car? So it is important for people to become cooperative.

Now let us turn to competition. Many people want to succeed, but not all the people can succeed. Competition makes people diligent, industrious, and makes the faster development of the society. One has little chance to succeed unless one is competitive. So it is also very important for people to become competitive.

Then which is more important, 'competitive or cooperative'? What is the relation between them? In my opinion, learning to compete is the best preparation for life, for if one learns to be competitive, one certainly wants to be cooperative. If one doesn't cooperate, one cannot succeed.

In a word, learning to be competitive is the best preparation for life, but one must be cooperative, as well.
Appendix G
Reliability Texts

First of all I consider important to point out what I think a best preparation is. In my understanding, a best preparation implies how you will do your life, that is to say, what you are expecting of the life; according to this, the best preparation is the way you want to live.

Some people, for instance, think that the best preparation for life is learning to be cooperative. As an example of these people we have the oriental culture (Chinese, Japanese, etc). They based their culture in cooperatives; in fact that is a kind of philosophy of the life. The way they take the life is easy. I'm not saying that they are lazy people, I just want to pinpoint that their mind is clear and unconflictive. With this phylosophy they've had the sight to progress with no neurosis.

On the other hand, some other cultures say that the best preparation for life is to be competitive. This is the occidental culture (I mean U.S.A., German, Great Britain, etc). Although they have had much progress with competition, they are more neurotic people than the oriental culture. Again, I don't want you to understand that this is bad, on the contrary, I just want to express this disadvantage so that everybody evaluates the two thesis
Let's take an argument. Statistically occidental people have more cardiac attacks than orientals, that means that competitive society is more neurotic; but, on the other hand, most of the technological advances have been developed by the occidentals, that means they have more discipline.

What I mean with all of these statements is that all depend on what you want to be, what role you want to play. In other words, if somebody prefers an easy life full of phylosophic moments and share with his/her people all he/she does, the best preparation is learning to be cooperative.

All the way around, if you prefer to be the best in your society and fight against life, the best preparation is to be competitive. Since these point of view, you are forced to be better every day.
Two manners of decisive action are to be compared, "quickly and decisively" versus "thinking over carefully." Each of them can be seen as opposite extremes of the man decision making behaviour. The main point is to get the goal without wasting time. On one hand, if one acts quickly and decisively he will never lose an opportunity by waiting but he can mislead his ultimate goal by choosing the wrong alternative. On the other hand, one which always thinks over carefully will decide wisely toward his goal, but can lose opportunities loosing time. Between the two extremes exist a range where anyone can choose to make a decision and the speed of the decision action should be related to the type of goal one is pursuing. As example of quick action goal one has a test, a sport, a task, etc. On the other hand as an example of thinking carefully one has life planning, career planning, research develop. The type of action taken to accomplish any decision has to be choose depending on the type of goal that decision is pursuing, i.e. goals in which one can spend time on.
Beeing competitive and cooperative are both very important as far as making an successful career goes. As a student, I think the more important personal characteristic is beeing competitive. You "compete" with the other students to see how well you perform compared to them. Usually this competition is something you just can feel and it can be a great stimulance for reading a little extra for an coming exam. You compare grades and so forth with your fellow students, and often this can leed to a competitive attitude. One might not be consious of this, but I think most students compete among each other even though many don't realize it themselves. At certain points in ones academic career this competition gets to be quite concrete. For instance there are only a certain number of places available in the Engineering college, so these places will be given to the students having the best grade point average among the applicants. Certainly beeing cooperative is important as a student, but if I had to choose one of the two as beeing more important I would choose beeing competitive. When one consider ones professional career, I think a cooperative attitude is more important. Your employer would apreciate a cooperative employe much more than a competitive one. After all, you would be competing with people within your own company, possibly creating an almost hostile atmosphere. The company must be competitive, but one creates a
competitive company by having a cooperative working force and a cooperative attitude among the different groups within the organization. A good relationship between the labor organization and the management, and a cooperative attitude within the groups, will help to keep a fabric or any commercial outfit to run smoothly.
All the thing has strong points and weak points. Complete-good or complete-bad action is not exist in the world because man is not God and the decision is made by man.

In the same way, quick action and careful action have both strong points and weak points. Assume that you went to see the base game and your friend asked you if it was interesting. If your immediate answer was not given, because you had a difficulty to make a decision, your friend maybe disappointed at your lack of sense.

On the other hand, we could not hear the 5th Symphony of Beethoven without his long & deep consideration about composition

So we cannot say uniformly which way of action is good or not without a consideration of circumstances. Rather, it's better to say "it depends"

The most advisable way is you should be able to think both kinds of action and be able to use them depending on circumstances

In fact, Being moderate is one of the most difficult thing.

Anyway, you always think the situation before, and then act depending on circumstances.
The life of a human being becomes to a technical and progressive society since the Industrial Revolution in the end of 18 century. The impacts of made-by-machine rather than made-by-hand really change almost everywhere in the world. Thus, everyone should change his behavior, action, or even attitude to fit the new, modern and aggressive ways of living. Thus someone may think he should be more competitive than before to get the more resources to improve his living. The monetary value becomes the major beliefs of many people in this era. As a result, men become not to believe each other and to be more selfish.

On the other hand, someones think the progression of the modern life is based on the cooperation of human-being, rather than competition. Everyone's success or innovation comes from the success of someone before him. This phenomenon is especially obvious in the field of science. Many new theories discovery are the inference of some important theory proved before. At the same time, many new productions are developed by "team work". Japan is the best example of "team work". The spirit of "team work" makes Japan being the new modern country of the world.

Both the competition and cooperation make the human lives to be more comfortabale and improvable. However, men should realize that we should have the competitive attitude to the improvement of living and have the
cooperation attitude to help each other and share the best modern life which is never before.
TOPIC: Learning to be cooperative is the best preparation.

Learning to be cooperative is the best preparation for life. Even some people say that learning to be competitive is the best way to face life. I choose a cooperative life. There are several reasons that explain why I have this choice.

First, from the psychological points of view, people are a group-living type. Man loves to be loved and respected. If one person chooses the competitive way, he will be jealous when he finds someone better than him. His mind is full of hatred when he don't satisfy what he got already. He feels unfair. On the contrary, you will see the satisfactory smile on the cooperative people. They will not feel empty in their soul. Instead that they are happy.

Second, from the physical points of view, you will find a competitive king has the highest position in his county. His ambition will never be satisfied. He loves war. His behaviors become disgusted. One the other hand, you will see even a poor fellow. He learns to be cooperative. His attitudes are friendly.

Above all, we knows that if you want a happy life and a successful life. Learning to be cooperative is the best choice in you life.
Everyone has his own value system about how to prepare for his life. On the whole, there are three ways to prepare for one's life. First of all some people regard learning to be cooperative as the best preparation for life. Second, some people look upon learning to be competitive is the best one. Finally, some people consider the harmony between to be competitive and to be cooperative as the best one.

I think the third one is the best preparation for life. As far as cooperation is concerned, to be cooperative is one of the good tools in keeping, maintaining, and deepening human relations. To prepare for life, however, learning only to be cooperative is not enough, because life is not other's but mine. To keep my life more fruitful, I have to raise my capacity to cope with the problems taking place accidently.

Increasing one's capacity is to learn to be competitive. As only learning to be cooperative is not the best preparation for life, so only learning to be competitive is not the best one. Because human is a social being, everyone needs to be cooperative to keep his life happy.

The best way to prepare for life, therefore, is learning to be cooperative as well as to be competitive. In other words, we must pursue the harmony between them in order to manage our life successfully.
It is a generally accepted idea that the point of view toward life depends on the society, or that the ways of thinking about life differ from society to society. Cooperation is emphasized in some societies whereas competitiveness is put first in other societies.

Japan and the United States, as cultural bodies, clearly illustrate the diversity in ways of thinking. The Japanese are usually educated to be cooperative with others as they grow up, and, as a result of that, Japanese companies became strong in the international market, because company is basically a unit of cooperation. It is, however, widely said that the Japanese people do not have their originality as an individual, that is to say, they do things like others do, or they do not decide until they check out what others decide, and so on.

On the other hand, the people in the United States are educated to be individual, independent, and original. Because of this natural trait, they can, or tend to, take their own initiative in the international society, and, of course, they take their own initiative as an individual in daily life as well. Owing to this trait, however, they often have conflicts with others, and the activity of companies are relatively getting down over these couple of decades.

As a painter, I prefer the individually - or competitively - oriented society, because it is definitely
necessary to have my own time to do my job, which should not be a simple imitation of others. Original colors, layout and thought in painting are created on the individual basis. Since painting is going to be my career, it is important for me to be among individually-motivated people. Thus, I agree with the competitive life style.
There are many ways to live happily. Everyone has his own life and lives alone or together.

Which way is the best for life?

Some people say that the preparation for life is learning to be cooperative. Others say that learning to be competitive is the best.

Let's think about the merits and demets of each way and about which way we will choose.

There are many society and countries in the world and live many people. They live alone or together. Most of them live together.

There are a lot of problems and troubles to live with others. We have families. We also have a lot of friends.

It's very important to learn to be cooperate with others. If not, there are so many trouble and wars that we can't live together. The most important thing to live together is mutual understanding.

Modern society has been being developed day by day. If man had lived harmourously without competition, there was no development in our world.

Man naturally has the jealousy against the others. So man always tries to get better. This is the motive power to develope our society.

I discussed the two way to live. The best way for life is learning to be cooperate and competitive harmourously. We must learn the way to be cooperate with
others and then adequate individual competition may be
develop our society. The most important thing to live
with is the effort to understand each other.
Appendix H

Guidelines for Generating Negotiated Texts

Please read this composition until you feel you understand what the author was trying to convey. Then, rewrite the passage in such a way that it is grammatically correct but still maintains the character of the original. You may refer to the original whenever you want. Change original vocabulary, punctuation, and phrasing only when it is necessary to clarify meaning, as you understand it to be.

If you encounter a passage in which meaning is unclear, or for which there is more than one plausible interpretation, underline the passage in the original and record the interpretation you feel is most likely to reflect the author's intent. If you feel that equally-plausible alternative interpretations exist, underline the passage in question in the original and indicate on your protocol the possible alternatives in the following manner:

(Alternative #1)

TEXT - TEXT

(Alternative #2)

Remember, make changes only when meaning is impaired in the original. Resist the impulse to improve on the original stylistically or in terms of choosing more elegant or powerful vocabulary. If it makes sense, don't change it!
Appendix I

Negotiated Texts

COMPOSITION #1 (LEVEL 107, FORM B)

In order to give an answer to the question of whether it's better to learn to be competitive or cooperative it is necessary to consider a few factors which are important:

- In which society do the people live and how it is organized;
- The personality of the individual and the individual abilities.

Everyone will agree with the statement that it is important to learn both of these abilities, to be competitive and cooperative.

And everyone will agree that every society has developed its own way to select people at school and in business everywhere.

There are differences between the social aims that have an influence on the individual aims. In America the organization of the society is much more competitively oriented than in some states in the eastern hemisphere. As far as a know the people in Japan are much more oriented in a cooperative style of working and living but they are very competitive as a nation or as a company – whenever they function as a group.
The reasons for these different kinds of social organization can be found in the historical development and the conditions of living.

Japan, isolated for many years from any influence and with little space for too many people, was forced to develop a very cooperative style of living. In contrast, the development of the USA was determined by individual activities. So Americans developed a more competitive style of social organization.

It is important to see these historical and social criteria before answering the question of whether it's better to learn a competitive or cooperative style of living. No individual person can be seen as isolated from the society to which he/she belongs.
I think the question of quick action versus deliberation is a rather controversial problem. It is difficult for me to say one is definitely better than the other since the choice between alternatives may depend on the importance of the matter of the situation.

There are times when taking quick and decisive action is more desirable. For instance, you got a very serious disease and your doctor told you that you need an operation which is very dangerous. You were also told that you might die if you didn't have the operation. In such a case, it would be very hard to make a quick decision, but you should do so since the matter is very serious. For another example, I go to a restaurant with some of my friends and we are supposed to have only 30 minutes in that restaurant because we are going to a movie afterwards. In that particular situation, I have to make a quick decision about what to order for my own sake. Otherwise, I may not be able to finish my meal or may miss some part of the movie.

There are many other cases when it is much better to think something over carefully, rather than to make a hasty choice. It is more advisable to be careful, especially when you have to make a very important decision in your life. For example, when you decide what you'll study in college, you should take time and think it over and over since your major will be related to your career.
For another example, I can say when you choose your spouse. That kind of decision is what you have to live with all your life and what you can not change even though you want to.

In conclusion, I would say that being careful and taking enough time is preferable if the issue is not very urgent and action doesn't need to be taken right away because you'll probably make fewer mistakes and repent less if you do so.
Appendix J

Criteria for the Identification of Idea Units

Idea Units: Propositions either stated or implied in the text.

Can include:

1. Simple sentences.
2. Subordinate (relative) clauses.
3. Logical relationships (connectors) indicating:
   a.) Comparison (... than, or);
   b.) Causation (because, for since);
   c.) Consequence (as a result, consequently, thus);
   d.) Contrast (but, however, yet);
   e.) Addition (and*);
   f.) Exemplification (for example, for instance);
   g.) Summary (in conclusion).
4. Key modifiers that can be "relativized" (e.g., The ugly man made the girl scream. ➔ The man was ugly. He made the girl scream.).
5. Prepositional phrases giving specific information (e.g., In America, people like sports. = two propositions).

* Linking two or more distinctly different propositions.

196
Appendix K

Valid Propositions and Their Mean Ratings of Importance to the Two Target Texts*

Instructions given to the three evaluators follow.

The score listed before each proposition reflects the mean rating of importance given to each proposition.

STEP #1: Read the attached selections and judge whether or not the following propositions are valid thoughts presented in the text, using the criteria provided as a guide. If for any reason you judge that a given proposition inaccurately reflects the content of the selections or fully duplicates the information presented in another proposition, please indicate this by writing "no" in the blank preceding the item.

Note: When a part of a proposition is in boldface, that portion is the only portion being evaluated. The rest of the sentence is included to provide a meaningful context. Information in lower case parentheses provides another possible wording of immediately preceding information. Information in boldface parentheses is inferred information, strongly implied but not specifically stated.

STEP #2: Using the blank lines at the end of the lists, please include any additional propositions that are presented in the selections but not included on the list.
STEP #3: Finally, using the scale below, rate each proposition that was not marked "no" in STEP #1 according to your estimate of its relative importance to the overall meaning of the text:

7 = extremely important
6 = of great importance
5 = quite important
4 = of average importance
3 = of below-average importance
2 = only slightly important
1 = of negligible importance

* Adapted from Pederson (1985).
Appendix L

Holistic Criteria

106 - The essay is restricted to a very narrow range of language, ideas are underdeveloped or presented in a rudimentary or repetitious way, and serious grammatical problems occur. Error density is high, and may render long stretches of text incomprehensible. Serious contradictions and/or inconsistencies within the thread of discourse may further obscure meaning.

107 - Basically the default category, in which new students who do not obviously belong in either 106 or 108 are placed. Essays reveal an ability to use regularly, but not necessarily faultlessly, a variety of grammatical structures generally appropriate to the topic. Overall, a sense of pattern and development is present to a degree where meaning is only occasionally obscured. Syntax and vocabulary are richer than in 106, but still somewhat limited in comparison to 108.

108 - The essay demonstrates sufficient command of syntax and vocabulary to convey, without serious distortion or excessive simplification, the range of the writer's ideas. Because the paper is generally well-organized and errors are comparatively minor or sporadic in nature, meaning is easy to follow throughout.
Q - The essay is well-organized and ideas are expressed clearly and in appropriate language. A sense of pattern or development is evident throughout. Sentences reflect a command of syntax within the ordinary range of standard written English.
Appendix M

Subject Score Sheet

NAME________

<table>
<thead>
<tr>
<th>SAMPLE #</th>
<th>106</th>
<th>107</th>
<th>108</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix N

Coded Cover Sheet for Recall Protocols

FORM A/B

NAME

SAMPLE # ___
Appendix 0

Instructions for Writing Recall Protocols

If the paper you have just read was on the topic "Competition versus Cooperation," circle 'B' on the booklet you've just received. If the topic was "Quick versus Careful," circle 'A'. Be sure to indicate in the space provided that this is sample #10.

Now I want you to write down everything you remember about the paper you have just read. Wherever possible, use the exact wording and ordering of content contained in the text. However, feel free to paraphrase the contents as you remember them, so long as the original intent of the passage is preserved. Just write down the author's words as you remember them. Do not refer back to the text, itself.