



2008-07-18

The Effects of Manageable Corrective Feedback on ESL Writing Accuracy

K James Hartshorn

Brigham Young University - Provo

Follow this and additional works at: <https://scholarsarchive.byu.edu/etd>



Part of the [Educational Psychology Commons](#)

BYU ScholarsArchive Citation

Hartshorn, K James, "The Effects of Manageable Corrective Feedback on ESL Writing Accuracy" (2008). *All Theses and Dissertations*. 1522.

<https://scholarsarchive.byu.edu/etd/1522>

This Dissertation is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.

THE EFFECTS OF MANAGEABLE CORRECTIVE FEEDBACK
ON ESL WRITING ACCURACY

by

K. James Hartshorn

A dissertation submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Instructional Psychology and Technology

Brigham Young University

July, 2008

Copyright © 2008 K. James Hartshorn

All Right Reserved

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a dissertation submitted by

K. James Hartshorn

This dissertation has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date

Paul F. Merrill, Chair

Date

Neil J. Anderson

Date

Norman W. Evans

Date

Diane Strong-Krause

Date

Richard R. Sudweeks

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the dissertation of K. James Hartshorn in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

Paul F. Merrill
Chair, Graduate Committee

Accepted for the Department

Date

Andy S. Gibbons
Department Chair

Accepted for the College

Date

K. Richard Young
Dean, David O. McKay School of Education

ABSTRACT

THE EFFECT OF MANAGEABLE CORRECTIVE FEEDBACK ON ESL WRITING ACCURACY

K. James Hartshorn

Department of Instructional Psychology and Technology

Doctor of Philosophy

The purpose of this study was to test the effect of one approach to writing pedagogy on second-language (L2) writing accuracy. This study used two groups of L2 writers who were learning English as a second language: a control group ($n = 19$) who were taught with traditional process writing methods and a treatment group ($n = 28$) who were taught with an innovative approach to L2 writing pedagogy. The methodology for the treatment group was designed to improve L2 writing accuracy by raising the linguistic awareness of the learners through error correction. Central to the instructional methodology were four essential characteristics of error correction including feedback that was manageable, meaningful, timely, and constant.

Core components of the treatment included having students write a 10-minute composition each day, and having teachers provide students with coded feedback on their daily writing, help students to use a variety of resources to track their progress, and encourage students to apply what they learned in subsequent writing. Fourteen repeated measures tests using a mixed model ANOVA suggest that the treatment improved mechanical accuracy, lexical accuracy, and certain categories of grammatical accuracy. Though the treatment had a negligible effect on rhetorical competence and writing fluency, findings suggest a small to moderate effect favoring the control group in the development of writing complexity.

These findings seem to contradict claims from researchers such as Truscott (2007) who have maintained that error correction is not helpful for improving the grammatical accuracy of L2 writing. The positive results of this study are largely attributed to the innovative methodology for teaching and learning L2 writing that emphasizes linguistic accuracy rather than restricting instruction and learning to other dimensions of writing such as rhetorical competence. The limitations and pedagogical implications of this study are also examined.

ACKNOWLEDGEMENTS

First, I would like to thank my kind and thoughtful committee members—my chair, Paul F. Merrill, and the other committee members, Neil J. Anderson, Norman W. Evans, Diane Strong-Krause, and Richard R. Sudweeks, each of whom is a master teacher who has taught me a great deal in and out of the classroom. Similarly, I would like to express gratitude to all those associated with the English Language Center who provided invaluable support throughout the study.

Second, I would like to thank my family—my wife Joella and our children, Rebekah, Rachel, Adam and Joseph, each of whom made substantial personal sacrifices that made it possible for me to continue my education and complete this work.

Finally, I would like to thank my Heavenly Father whose loving kindness and tender mercies have sustained me and allowed me to accomplish infinitely more than I ever could on my own.

Table of Contents

CHAPTER 1: INTRODUCTION.....	1
Challenges Associated with L2 Writing Pedagogy	2
Pedagogical Priorities of L2 Writing Teachers.....	3
Differences in the Learning Needs of L1 and L2 Writers	6
Limitations in Traditional Approaches to Improving L2 Accuracy	9
Corrective feedback.....	9
Grammar instruction	11
An Alternative Approach to Improving L2 Accuracy	12
CHAPTER 2: LITERATURE REVIEW	15
Writing Instruction.....	15
Process Writing.....	17
Grammar Instruction.....	20
Error Correction	28
Measures of L2 Writing Production	36
Writing accuracy.....	37
Writing fluency	45
Writing complexity.....	46
Rhetorical competence.....	50
Summary.....	52
Research Questions.....	54
CHAPTER 3: METHOD	56
Participants.....	56
The students	56
The teachers	60
The scorers and raters	62
Research Design.....	62
Instruments.....	64
The rhetorical competence rubric	65
The grammar knowledge test.....	65

Reliability Design	66
Scoring	66
Rating	68
Instructional Methods	72
Elicitation Procedures	77
Research Questions Operationalized	80
CHAPTER 4: RESULTS	84
Reliability Estimates	84
Scoring reliability.....	84
Rating reliability	85
FACETS output	93
ANOVA Test Results	97
Linguistic Accuracy Index.....	124
Grammatical Accuracy Index	128
CHAPTER 5: DISCUSSION AND CONCLUSION	131
Discussion.....	131
Limitations	141
Pedagogical Implications	145
Suggestions for Further Research	149
Conclusion	152
References.....	154
Appendix A: Examples of Coded Feedback for Error Correction.....	168
Appendix B: Rhetorical Writing Competence Rubric	169
Appendix C: Partially Nested Design for Estimating Interrater Reliability	170
Appendix D: Error Tally Sheet.....	171
Appendix E: Edit Log	172
Appendix F: Error List.....	172
Appendix G: Institutional Review Board Approval Letter.....	173

List of Tables

Table 1: Experimental Groups by Native Language and Gender	57
Table 2: Descriptive Statistics for Gender and Accuracy Scores	58
Table 3: ANOVA Summary Table for Gender and Accuracy Scores	58
Table 4: Control Group Students by Term, Teacher and Teacher's Experience	60
Table 5: Treatment Group Students by Term, Teacher and Teacher's Experience.....	61
Table 6: Pretest, posttest nonequivalent control group design	62
Table 7: Dependent Variables and Their Methods of Measurement	64
Table 8: Stratification for the Second Scorer's Random Sampling.....	68
Table 9: Pearson Correlation Coefficients by Accuracy Type between S1 and S2.....	85
Table 10: Summary of FACETS Output for Student Essays.....	89
Table 11: Summary of FACETS Output for Raters 1, 2 and 3.....	90
Table 12: Summary of FACETS Output for the Rhetorical Competence Rubric	92
Table 13: Descriptive Statistics for Accuracy Scores.....	98
Table 14: Mixed ANOVA Summary Table for Accuracy Scores	98
Table 15: Simple Main Effects for Pretest and Posttest Accuracy Scores	99
Table 16: Descriptive Statistics for Rhetorical Competence Ratings.....	101
Table 17: Mixed ANOVA Summary Table for Rhetorical Competence Ratings	101
Table 18: Descriptive Statistics for Writing Fluency Scores.....	102
Table 19: Mixed ANOVA Summary Table for Writing Fluency Scores.....	103
Table 20: Simple Main Effects for Pretest and Posttest Fluency Scores.....	104
Table 21: Descriptive Statistics for Writing Complexity Scores.....	105
Table 22: Mixed ANOVA Summary Table for Writing Complexity Scores	105
Table 23: Simple Main Effects for Pretest and Posttest Complexity Scores.....	107
Table 24: Summary of Bivariate Regression Analysis	108
Table 25: Descriptive Statistics for Sentence Structure Accuracy Scores.....	111
Table 26: Mixed ANOVA Summary Table for Sentence Structure Accuracy Scores ...	111
Table 27: Descriptive Statistics for Determiner Accuracy Scores.....	112
Table 28: Mixed ANOVA Summary Table for Determiner Accuracy Scores.....	113
Table 29: Simple Main Effects for Pretest and Posttest Determiner Accuracy Scores ..	114
Table 30: Descriptive Statistics for Verb Accuracy Scores.....	115

Table 31: Mixed ANOVA Summary Table for Verb Accuracy Scores	115
Table 32: Simple Main Effects for Pretest and Posttest Verb Accuracy Scores	116
Table 33: Descriptive Statistics for Numeric Accuracy Scores.....	117
Table 34: Mixed ANOVA Summary Table for Numeric Accuracy Scores	117
Table 35: Descriptive Statistics for Semantic Accuracy Scores.....	118
Table 36: Mixed ANOVA Summary Table for Semantic Accuracy Scores	118
Table 37: Simple Main Effects for Pretest and Posttest Semantic Accuracy Scores.....	119
Table 38: Descriptive Statistics for Lexical Accuracy Scores.....	120
Table 39: Mixed ANOVA Summary Table for Lexical Accuracy Scores	121
Table 40: Simple Main Effects for Pretest and Posttest Lexical Accuracy Scores	122
Table 41: Descriptive Statistics for Mechanical Accuracy Scores	122
Table 42: Mixed ANOVA Summary Table for Mechanical Accuracy Scores	123
Table 43: Simple Main Effects for Pretest and Posttest Mechanical Accuracy Scores..	124
Table 44: Descriptive Statistics for the Linguistic Accuracy Index	125
Table 45: Mixed ANOVA Summary Table for the Linguistic Accuracy Index.....	126
Table 46: Simple Main Effects for Pretest and Posttest Linguistic Accuracy Index.....	127
Table 47: Descriptive Statistics for the Grammatical Accuracy Index.....	129
Table 48: Mixed ANOVA Summary Table for the Grammatical Accuracy Index	129
Table 49: Simple Main Effects for the Pretest and Posttest GAI	130
Table 50: A Summary of Findings Used to Answer the Primary Research Question....	132
Table 51: A Summary of Findings Used to Answer the Phase II Research Questions ..	134
Table 52: A Summary of Findings for a Posteriori Test.....	137

List of Figures

Figure 1. Effort and Skill Mastery Plotted for L1 and L2 Writers	7
Figure 2. Error Families and Error Types Used to Analyze Writing Accuracy	41
Figure 3. Indirect Coding Symbols Used to Mark L2 Student Writing.....	74
Figure 4. Illustration of Components of Identification Coding	80
Figure 5. Vertical Plot of Student Essays, Raters, and Rubric levels in Logits.....	87
Figure 6. Probability Curves for Rhetorical Competence Ratings	92
Figure 7. Pretest and Posttest Means for Accuracy Scores.....	99
Figure 8. Pretest and Posttest Means for Fluency Scores	104
Figure 9. Pretest and Posttest Means for Complexity Scores	106
Figure 10. L2 Writer Performance Plotted by Grammar Knowledge and Accuracy.....	109
Figure 11. Pretest and Posttest Means for Determiner Accuracy Scores	113
Figure 12. Pretest and Posttest Means for verb Accuracy Scores.....	116
Figure 13. Pretest and Posttest Means for Semantic Accuracy Scores.....	119
Figure 14. Pretest and Posttest Means for Lexical Accuracy Scores.....	121
Figure 15. Pretest and Posttest Means for Mechanical Accuracy Scores	123
Figure 16. Pretest and Posttest Means for the Linguistic Accuracy Index	127
Figure 17. Pretest and Posttest Means for the Grammatical Accuracy Index	130
Figure 18. Characteristics of Feedback Designed to Improve L2 Writing Accuracy.....	147

CHAPTER 1: INTRODUCTION

Writing ability is one of the most salient outcomes of learning in higher education. Formal writing appropriately occupies a unique place in professional-level communication for at least two reasons. First, unlike oral communication, formal writing tasks do not allow for an ongoing negotiation of meaning through interlocution. Therefore, the intended meaning must be expressed accurately to the reader. Second, the written medium is often reserved by society when important ideas need to be formalized, standardized or made more permanent. Thus, formal writing carries with it certain expectations of clarity, precision, quality and durability.

Notwithstanding the elevated role of writing instruction in higher education, a majority of Second Language (L2) learners continue to be challenged by it throughout periods of intensive study as well as long after they have been accepted into the university. Extensive observation of those learning English as a Second Language (ESL) suggests that writing difficulties are particularly evident in learners' abilities to produce writing that is linguistically accurate. Even after ESL students learn to produce writing that is fairly substantive, well organized and cohesive, many still struggle to extricate themselves from the linguistic gulf that separates them from their native-speaking peers. Though occasionally inaccurate writing may merely be an annoyance, it often obstructs the reader's ability to understand what is written and may affect the reader's perception of the writer or the writer's language ability (Ferris, 2006; Ferris & Hedgcock, 1998; Horowitz, 1986; James, 1998; Johns, 1995).

With these important contextual factors in mind, this chapter provides a brief rationale for testing the efficacy of an innovative approach to L2 writing pedagogy that

has been designed to improve the linguistic accuracy of L2 writers. In doing so, this introduction touches on some of the challenges associated with developing an effective L2 writing curriculum. It also includes a simple discussion of how different L2 writing teachers have different pedagogical priorities and how the pedagogical needs of L2 writers are different from those of First Language (L1) writers. In addition, this chapter also points out that approaches to corrective feedback vary from one context to another and that there are a number of problems with traditional approaches to corrective feedback and grammar instruction.

Challenges Associated with L2 Writing Pedagogy

In order to understand the need for an alternative approach to L2 writing pedagogy, we must first understand some of the major challenges associated with developing an effective L2 writing curriculum. Despite the need for ESL learners to improve their ability to write accurately, linguistic accuracy is rarely the only objective in writing instruction. Teaching L2 writing is rather complex because of the many dimensions of writing that need attention. For example, consider the accuracy and substance of what is written; the originality of the ideas that are expressed; the organization, sequencing and flow of those ideas; the attention to the purpose of the writing, including the tone and the various needs of the audience; the use of appropriate devices and conventions associated with various genres of writing; the accurate use of citations and references and so on. These and many other important dimensions of writing may compete for the attention of the teacher and student throughout the learning process.

Though many aspects of writing development may demand attention in the writing class, not all seem to be learned or applied equally well by ESL writers. Therefore, it may be useful to distinguish the most challenging aspects of writing from those that may be less problematic. One way to do this may be to separate the linguistic or language-based aspects of writing from those dimensions that are based primarily on rhetorical conventions. While the linguistic aspects of writing might include features such as grammar, word choice, spelling and punctuation, the rhetorical conventions might involve the organization, presentation, development and flow of ideas.

Though there may be some minor overlap among these different dimensions of writing, this distinction is helpful because it allows us to see important differences in how these aspects of writing may be learned or applied by ESL writers. Though rhetorical conventions are primarily conceptual and seem to be learned and applied through conscious cognitive processes, the linguistic aspects of writing appear to be much less conscious and may take much longer to learn. Nevertheless, both seem to be important in developing competence in L2 writing. Just as the structural integrity and beauty of a building made of bricks and mortar would be severely compromised without either the bricks or the mortar, so good writing requires the appropriate rhetorical conventions as well as linguistic accuracy.

Pedagogical Priorities of L2 Writing Teachers

In addition to understanding the unique challenges associated with developing an effective L2 writing curriculum, we also need to understand how different L2 writing teachers emphasize different priorities in their instruction. For example, though most would agree that linguistic accuracy and rhetorical appropriateness are both essential to

quality writing, observation suggests that L2 writing teachers rarely focus their efforts equally on both of these dimensions of writing. For instance, it is interesting to note that historically grammatical accuracy was emphasized in second language learning during most of the last millennium. Then, seeing the limitations of such a narrow focus, writing teachers and theorists seem to have nudged the pedagogical pendulum closer toward rhetorical conventions in the second half of the twentieth century (Matsuda, 2001). Subsequently, many theorists and practitioners became critical of second language writing programs that saw writing simply as part of the learner's language development and that focused on the reduction of grammar errors (Dvorak, 1986; Susser, 1994). Kern and Schultz (1992), for example, lamented over those programs that emphasize "surface feature accuracy rather than on the development, organization, and effective expression of the students' own thoughts or ideas" (p. 2).

While it is appropriate to note the obvious limitations of writing instruction that focuses exclusively on linguistic accuracy, L2 writing teachers who simply adopt L1 instructional methods may lack the theoretical foundation to help their students to improve their linguistic accuracy. Hinkel (2004), for example, observed that the writing process and the rhetorical aspects of writing have been improperly emphasized over the linguistic skills ESL writers need to succeed in regular university classes. She also laments that many L2 writing practices have been adopted from L1 methods. She suggests that becoming a competent L2 writer is a very different process from becoming a competent L1 writer and that a writing process originally designed for L1 writing pedagogy is inadequate for effectively teaching L2 writing.

If linguistic accuracy is such an important component of L2 writing development, some might ask why many L2 writing teachers seem to favor rhetorical conventions at the expense of linguistic accuracy. Though the answers to this question may be complex and may vary from one teacher to another, interaction with colleagues and extensive personal observation suggest the following possible reasons:

1. Some teachers may not feel confident enough to teach the linguistic aspects of writing and may end up avoiding them intentionally or perhaps inadvertently.
2. Some teachers may feel that rhetorical conventions are easier to teach so they spend more time on them, rather than appropriately dividing their time.
3. Some teachers may feel that they lack the needed time to spend on linguistic aspects of writing after focusing on what seems to be more important features.
4. Some teachers may simply be caught in an L1 process to teaching that favors rhetorical conventions without adequately addressing accuracy.
5. Some teachers may feel that the real skill of writing is found in the use of the rhetorical conventions and that writers can get the linguistic help they need from others such as tutors or their grammar teachers.
6. Some teachers may believe that teaching rhetorical conventions makes a greater difference in the quality of student writing than would result from focusing on linguistic accuracy. Indeed, many teachers have lamented that in their personal experience, focusing on linguistic accuracy has done very little, if anything, to improve the accuracy of student writing.

For these and perhaps a number of other reasons, many L2 writing teachers have allowed a focus on linguistic accuracy to be crowded out of the curriculum by other pedagogical problems and priorities.

Differences in the Learning Needs of L1 and L2 Writers

While we have noted above that needs of L1 and L2 writers vary, it may be useful to consider more specifically how some of those needs may differ. Extensive observation of the learning of L1 and L2 writers provides some possibilities illustrated in Figure 1. It represents an attempt to graphically illustrate theoretical similarities and differences experienced by native speakers and non-native speakers on the path to becoming competent writers in English. This figure plots the effort of each writer on the horizontal axis. In addition to personal motivation and exertion, this notion of *effort* might be affected by the writer's access to quality learning resources, teachers and opportunities to practice and receive timely feedback. Skill mastery, or rhetorical writing competence and linguistic writing competence, appear on the vertical axes. Rhetorical writing competence is illustrated by a solid line while linguistic writing competence is depicted by a dotted line. Since competence in writing is perhaps more a matter of degree than an achieved state, these theoretical lines should be considered asymptotic, drawing increasingly closer to a state of complete mastery as competence increases but without actually reaching it.

Of particular note in Figure 1 is the similarity between the effort required of L1 and L2 writers to achieve a certain measure of rhetorical writing competence. This may be because this dimension of writing is based on cognitive mastery of concepts that appear to be equally accessible to native and non-natives alike. Three additional notions depicted in Figure 1 are worth mentioning. First, the figure suggests that though

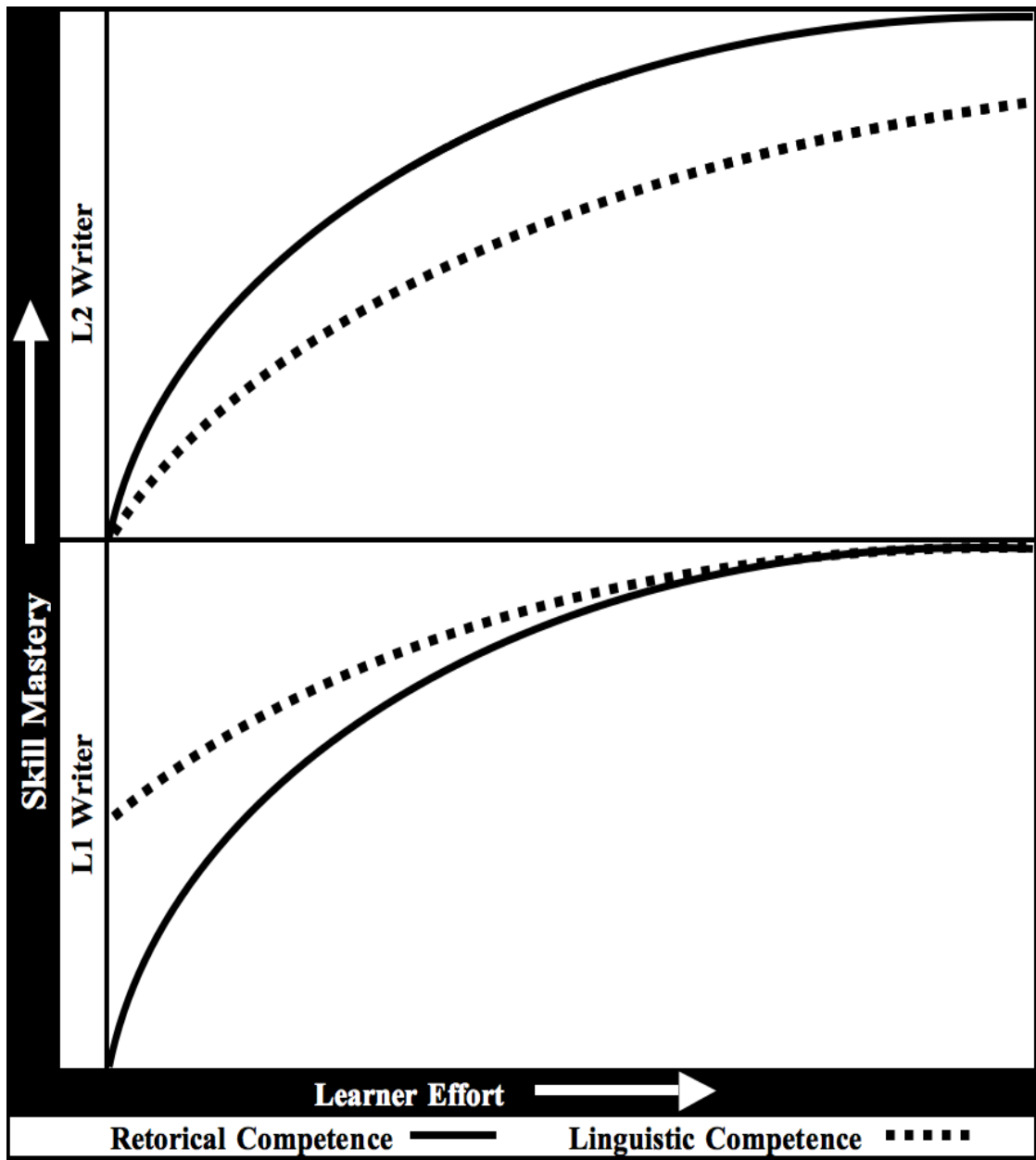


Figure 1. Effort and Skill Mastery Plotted for L1 and L2 Writers

rhetorical writing competence may be more difficult to achieve than linguistic writing competence for native writers, linguistic writing competence appears much more difficult than rhetorical writing competence for non-natives. Second, at the outset of their learning and with equal amounts of effort devoted to the development of rhetorical and linguistic competences, L2 writing students are likely to experience more rapid mastery of rhetorical skills than of linguistic skills.

Third, with equal amounts of effort devoted to the development of both competences, L2 writing students are likely to take much longer to achieve acceptable or more native-like levels of linguistic writing competence than rhetorical writing competence. Based on this simple description of prospective learning requirements for L1 and L2 writers, it is not surprising to see L1 writing models that focus more on rhetorical aspects of writing. This may be because deficiencies in rhetorical aspects of writing will be the most visible compared to the many linguistic conventions which will have been mastered before the native even begins to learn to write. Moreover, it should also be evident that models of L2 writing pedagogy would need to be different from L1 models if the unique needs of L2 learners are to be met effectively.

If Figure 1 captures fundamental differences of what is required for L1 and L2 learners to become competent writers, then it seems that ESL educators need a more complete model for teaching L2 writing that reflects those differences. Among other things, it seems that this model would need to incorporate the relevant rhetorical conventions along with better methods to help students improve their linguistic accuracy. However, as noted above, even when L2 writing teachers strive to help their students improve their linguistic accuracy, it often appears to be an unproductive endeavor.

One great irony in this process seems to be that even when L2 writing teachers laboriously provide corrective feedback, many students continue to struggle with the same linguistic problems in the final drafts of their paper or in subsequent writing tasks. As one teacher deeply entrenched in this labor, Hall (1991) described the dilemma this way:

Error correction does not appear to have much effect on students' written work. ESL writers continue to make the same errors time and time again, no matter how much time, effort and red ink is spilled over their papers by . . . teachers in the cause of grammatical accuracy. It is little wonder that some teachers have begun to question the validity of error correction. (p. 1)

Limitations in Traditional Approaches to Improving L2 Accuracy

Though it seems evident that L2 writers may have different learning needs when compared to L1 writers, what has been less clear for many L2 writing teachers is how they should design their instruction and feedback so it can make a tangible difference for their students. While many teachers struggle to see the positive effects of their corrective feedback, it is important to note that the way teachers provide corrective feedback throughout the writing process may vary greatly. For example, some teachers will identify the error and supply a correction, with the expectation that the students will fix the error in the subsequent draft. Other teachers will identify errors but will expect the students to come up with the corrections on their own based on what they have studied in the class.

Corrective feedback. While there seems to be growing evidence that some methods for providing corrective feedback may be more effective than others (Ferris,

2006), either approach may become less effective when the student papers are several pages in length. This is because the sheer number of errors can be overwhelming for the teacher to identify and equally overwhelming for the student to correct. As a result, neither the teacher nor the student may end up doing their job well. Although both painstakingly go through the motions dictated by the exercise, the intended outcome may not be realized.

Moreover, if the teacher's instructional load is particularly heavy and the papers are relatively long, several days (if not weeks) could pass before the teacher is able to return the papers with the needed feedback. Similarly, several days might pass before the student is able to make the needed corrections before the newest draft is returned to the teacher. In addition, it is not uncommon in this process for a student to fail to provide an acceptable correction or to miss an error marked by the teacher. Thus, the teacher may be confronted with new errors as well as old errors that need to be marked again in the newest draft of the paper. Such an approach tends to place an excessive strain on the teacher and the learner because of the large volume of errors. At the same time, this approach minimizes the number of opportunities to give and receive feedback.

While this process may eventually result in an error-free paper, it seems very unlikely that it will help the L2 writers to learn to write more accurately in future writing tasks. For example, even when students are astute enough to successfully make the needed corrections on a particular draft of a paper, it is not uncommon for the same types of errors to resurface again in subsequent drafts or in new pieces of writing (Truscott, 1996). Unfortunately, it seems that some L2 writing educators are focused on helping

students to produce good writing, rather than the more appropriate aim of producing good writers.

Grammar instruction. In addition to the apparent ineffectiveness of traditional approaches to corrective feedback in L2 writing classes, traditional approaches to grammar instruction in grammar classes seem equally ineffective in helping students to write more accurately. At times, the recurring linguistic problems noted above seem particularly perplexing when we realize, for example, that students are making errors with grammatical structures that they have already studied extensively in their grammar classes. In some cases, this may involve grammar that students have studied for a number of years, including grammar that is taught at some of the lowest proficiency levels. This raises serious questions about how we teach and assess students' grammar production. For example, (a) Why do students continue to use particular grammar structures inaccurately after being taught them in their grammar classes? (b) Why do some students continue to struggle with the accuracy of their writing even after demonstrating high levels of cognitive mastery of the grammar they have studied?

Perhaps at issue here is the different nature of the grammar instruction and assessment on the one hand, and the production required in the writing tasks on the other hand. It seems that the most meaningful applications of learning grammar would be in productive tasks such as speaking and writing, yet many assessment tools used widely involved objective test items rather than production tasks. Unfortunately, in interpreting the results of such tests, many erroneously assume they are an indication of students' productive grammar skills. While many intensive English language schools use multiple-choice grammar tests for placement, achievement and proficiency assessments, personal

observation suggests that such tests may not always correlate well with grammar performance in productive contexts such as writing. If this is true, multiple-choice tests may not be the most valid measure of productive grammar skills and other methods of assessment should be explored.

One argument in behalf of objective test items is that they allow the tester to assess student knowledge about grammatical structures that students are not likely to choose to produce on their own. While this may be one appropriate way to assess student knowledge of such structures, it raises a compelling question about instructional priorities. For example, consider the students who, in actual production tasks, consistently avoid particular structures. This may be because they do not “know” the structures or because they simply do not feel comfortable using them. Yet, they will consistently use a number of other constructions despite the fact that what they actually produce may be laden with errors. Could we conclude that based on their written idiolects that they are more ready to learn the correct form of the constructions that they regularly use than to learn the correct form of the constructions they regularly avoid? If so, perhaps our pedagogical focus at higher proficiency levels should be on those constructions that learners demonstrate a willingness to use.

An Alternative Approach to Improving L2 Accuracy

Now that we have reviewed some of the challenges associated with an L2 writing curriculum, including the different needs of L2 writers and some of the approaches L2 writing teachers, we are prepared to briefly examine an innovative approach to L2 writing pedagogy designed to improve L2 writing accuracy. Though a traditional grammar syllabus and traditional grammar and writing instruction may still have an important

place in a larger curriculum, perhaps a priority at the higher levels of proficiency should be on a dynamic grammar syllabus that focuses on teaching to meet individual needs rather than on providing instruction on a list of grammatical principles that, in the end, the student may choose to avoid in their writing tasks.

Perhaps we need to rethink how we go about organizing the teaching and learning experiences involved in L2 writing. In addition to the inclusion of rhetorical conventions, another course component that L2 writers may need is a method that helps them learn to edit their own writing and reduce their errors. It seems that both teachers and learners would benefit from an approach that would focus on fewer corrections at a given time with more frequent feedback. Such an approach to writing pedagogy has been used by Dr. Norman Evans of the BYU Department of Linguistics and English Language, (Evans, forthcoming) and it has shown some promise in helping students improve their writing accuracy at BYU's English Language Center (ELC).

Rather than overemphasizing the process of writing and rhetorical conventions at the expense of linguistic accuracy, the core component in this approach is a 10-minute writing completed at the beginning of each class period. Because the writings are small, the teacher is able to provide corrective feedback by the next class period. Moreover, since the writing is a manageable size, more is expected of the learner in terms of processing and applying the feedback. For example, the learner keeps track of errors using a running log of each error he makes in terms of its type and frequency. Over time he becomes well acquainted with his most frequent error types and may be less likely to make a particular error in the future. He also needs to rewrite the essay until it is free of errors. In addition to writing activities found in traditional writing classes, this daily

approach seems to show promise of helping students improve their writing in terms of their linguistic accuracy and editing skills.

Another argument for this type of instructional method has to do with the nature of the learner's continuing education. Though ESL learners will continue to refine their writing skills at the university, their writing experiences and the feedback they receive may be quite different. While L2 writers are likely to continue to learn about various rhetorical conventions associated with different specializations, they are not likely to receive the same kind of specialized feedback about their linguistic accuracy that was possible in their intensive English program. Even in English writing courses at the university, there is likely to be a greater focus on rhetorical conventions than on the linguistic accuracy of the learner's writing. If this is true, the intensive English program needs to strive to help its higher level students to become as linguistically independent as possible, so when they leave for the university, they will be better equipped to recognize their own errors and to edit their own writing.

These arguments provide a rationale for studying an approach to L2 writing that seeks to improve writing accuracy by helping teachers to provide students with corrective error feedback that is both immediate and manageable. However, since this method has not yet been tested with proper controls, it may be difficult to determine its true effect on student writing. Therefore, the aim of this study is to test this approach against a more traditional L2 writing method to determine what effects it may have on L2 student writing. The findings of this study should have important implications for the ongoing refinement of the curriculum at the BYU's ELC and may have broader implications for the fields of L2 writing research and pedagogy.

CHAPTER 2: LITERATURE REVIEW

As stated in Chapter 1, the purpose of this study is to test the effects of one approach to L2 writing pedagogy. This approach aims to improve the linguistic accuracy of L2 writers without diminishing other important dimensions of writing. This is to be done, primarily, by focusing on corrective feedback that is manageable and fairly immediate. To help contextualize this study, we will briefly examine a variety of relevant literature that addresses writing instruction, process writing, grammar instruction, and error correction. We will also examine common methods of measuring L2 writing accuracy, fluency, complexity and rhetorical competence.

Writing Instruction

Corbett (1971) informs us that by the late nineteenth century, various remnants of classic rhetoric could be seen in the writing instruction of native speakers of English. This was most likely due to the work of Whately in 1828 and the writing textbooks such as those written by Hill in 1878 and Genung in 1886 (Berlin, 1984). At this time, writing began to take on a more prominent role with an increased emphasis on the organization of a written work. Rather than attending to the *process* of writing, however, the objective usually was to produce the perfect *product* in the first draft, including the accurate and skillful use of grammar, spelling, punctuation, vocabulary as well as organization (Murray, 1978; Raimes, 1986; Taylor, 1981; Zamel, 1976).

Matsuda (2001) indicates that during this same period, second language writing was largely ignored because the field of applied linguistics was preoccupied with spoken language and “writing was merely defined as an orthographic representation of speech” (p. 17). Though some script was used in the early part of the twentieth century, its

primary purpose was to facilitate the learning of spoken language. Matsuda also explains that though many early attempts to teach writing in ESL contexts drew from L1 approaches to writing instruction, the 1950s saw the beginnings of a “division of labor” between those with expertise in teaching L1 composition and ESL specialists who would take on the task of teaching writing to non-native speakers (p. 18).

Though early ESL training had focused on preparing teachers to teach the spoken language, in the 1960s, second language writing began to emerge as its own discipline that attempted to guide ESL teachers in methods of writing instruction (Ferris & Hedgcock, 1998). Some early approaches to teaching writing in ESL classrooms included exercises where students produced their own original compositions (Erazmas, 1960), but this practice was heavily criticized by those who felt that allowing students to produce their own writing would be harmful because of the many errors they would make (Pincas, 1962). Subsequent approaches included controlled composition, where errors were prevented by carefully controlling student writing, and later guided composition, where errors were avoided through highly structured writing activities (Pincas, 1982; Raimes, 1991). Ultimately, however, these approaches were mostly limited to sentence-level exercises and were too restrictive to help students learn to produce their own original writing (Matsuda, 2001).

Later, others such as Kaplan (1966) and Arapoff (1967) suggested that writing pedagogy must do more than simply acquaint student with sentence-level constructions. Based on the growing assumption that the structure of paragraphs are specific to a particular language and culture and are subject to L1 transfer errors, they recommended that writing teachers expand their focus to paragraph-level discourse. Thus, the notion of

rhetoric, or the principles that guide the organization of writing, so central to L1 writing, began to be emphasized in a great deal of L2 writing as well.

With an increased awareness of the numerous parallels between the writing of native English speakers and advanced ESL writers, Zamel (1976) suggested that ESL writing teachers could benefit from the theories and research that shaped L1 composition. Subsequently, many ESL writing teachers reverted back to the product-centered approach used in L1 writing that encouraged students to analyze and mimic samples of model writing. However, Coe (1987) suggested that though this approach showed students what their writing should look like, it was not successful enough at helping them learn how to apply these idealized patterns of rhetoric and form in their own writing.

Process Writing

Ironically, at a time when many ESL writing teachers began to look to L1 writing approaches for guidance, many L1 writing teachers began to replace product-centered approaches with process writing. Many began to see L1 and L2 writing as a process of discovery that went far beyond the limitations of the product approach (Murray, 1978; Raines, 1985; Taylor, 1981; Zamel, 1983). Subsequently, in an attempt to address the broader needs of learners, many ESL researchers began to advocate the use of process writing for the second language classroom (Roca de Larios, Murphy & Marin, 2002; Scott, 1996; Susser, 1994; Zamel, 1983).

Though the notion of process writing may have been around since the 1950s (Matsuda, 2003), its true momentum seems to have started in the 1970s. Perhaps the most salient feature of this approach to writing pedagogy was its recursive nature as learners worked through a number of phases of their writing. This process was not only designed

to help student produce a better product, but it was designed to help them learn to become better writers. This process has been variously described by different authors who have occasionally used different terms though the underlying approach has been quite similar, if not the same. For example, Flower and Hayes (1981) include the prewriting, writing, revising, and editing phases. Similarly, Zemach (2007) refers to brainstorming, organizing, drafting, reviewing, editing and revising, and rewriting. Throughout this process, teachers provided explicit instruction and feedback for multiple drafts of a work to help learners master the various conventions of writing. Often the finished work was included in some type of portfolio or published for a specific audience (Hoffman, 1998).

In second language learning contexts, most process writing teachers generally followed the instructional pattern described by those such as Murray (1978), Sommers (1982) or Zamel (1985) where the initial drafts focused on content and organization and the later drafts focused on linguistic accuracy. For example, Murray (1978) suggests that in the prewriting stage, emphasis is placed on the generation of ideas. This may include activities such as brainstorming, outlining or free writing. Later, the process includes gathering additional information from sources such as books, other publications or interviews. Murray further explains that in these stages, the writers are not overly concerned with spelling, grammar or word selection.

In the final stages of the process, writers revise their work to refine the content and structure of their writing. At this point, attention is given to ensure that the introduction has an appropriate thesis statement, that the body has well-formed paragraphs and topic sentences and an effective conclusion. The writers refine the overall structure and incorporate appropriate transitions. Finally, the writers edit their work with

special attention to spelling, punctuation and grammatical accuracy (Murray, 1978). Since providing linguistic feedback can be very labor intensive, waiting toward the end of the process seemed wise because it limited the feedback to the content in a composition that had already been refined and that was likely to remain in the final draft.

Though some have claimed that ESL writing instruction is in a “post-process era,” where approaches to discourse strive to deal with varied issues such as the role of power, criticism of objectivity, social and cultural orientations and the irreducibility of the writing process (Kent, 1999), the reality is that today the general tenants of process writing are used fairly extensively in ESL classrooms around the world (Matsuda, 2003). This is particularly true in programs that prepare ESL students for university-level study.

Many students finish this process with a substantive piece of writing that includes satisfactory organization and cohesiveness. In addition to producing a *product* that may be publishable or that may serve as a useful model for the student in the future, the learner may also benefit a great deal from the writing *process* itself. This is particularly true when students are able to learn important skills that can be applied in later writing tasks. They may also benefit from their experiences that are associated more broadly with the writing process such as learning how to use library or internet resources, conduct interviews or engage in other activities that enhance their writing.

However, although process writing continues to be used widely, it is not without some controversy. For instance, those such as Silva (1993) and Hinkel (2004) have raised questions of the appropriateness of L2 writing methods that rely on L1 composition theory. Others have lamented over visible inconsistencies in what constitutes process writing as well as its wide-ranging and diverse applications in practice. For example,

Tobin (1994) observes that process writing “has become an entity . . . apart from its first theorists” (p. 8). Those such as Raimes (1986) and Tobin (2001) have also observed an oversimplification in the perceptions of some writing teachers, resulting in dichotomous views that tend to see process writing as too slack and unstructured or that see product-oriented approaches as preoccupied with grammar and too stifling. In response, Raimes (1986) has recommended that rather than debating over which focus may be best, we should explore “how to include the best of both” (p. 20) in our writing instruction.

Grammar Instruction

Along with our examination of writing instruction and the process writing movement, it will be useful to review some of the developments in formal grammar instruction. Grammar instruction in L2 study has a long history indeed. Howatt (1984) informed us that after the fall of the Roman Empire and the eventual rise of the Romance languages, Latin and Greek were often taught in schools where teachers would focus almost exclusively on grammatical structures. However, the rise in international commercial enterprises in Europe near the end of the eighteenth century precipitated the need for many to study modern languages as well. Howatt went on to explain that the Grammar-Translation Method emerged in response to this new need. It originated in Germany and then spread to England and other parts of Europe.

Using the method, teachers presented a number of grammar rules along with new vocabulary to aid student efforts to translate authentic classical texts. The main objectives were to develop L2 reading and writing skills, and throughout the classroom experience, linguistic accuracy was a major focus. Though other methods and approaches to L2 teaching and learning would appear later, the Grammar-Translation Method was the

avored mode of L2 instruction until the first part of the 1900s. It is interesting to note that even today the method can be seen in foreign language classes at many universities (Richards & Rodgers, 2001).

Despite the long history of grammar instruction and its enduring presence in L2 classrooms, L2 educators continue to struggle to understand its precise role in the teaching and learning processes. In fact, Richard and Rodgers (2001) have observed that this debate over the role of grammar instruction in the classroom has appeared in the professional literature for over a century. To help inform this debate, a number of empirical studies began to emerge in the 1960s and 1970s. Some of these will be highlighted below.

One early study sought to determine the best mode of L2 instruction. Using two groups of college-level students learning German, Scherer and Wertheimer (1964) compared the Grammar-Translation Method (which emphasized grammar in reading, writing and translation contexts) with the Audiolingual Approach (which minimized explicit instruction and focused on pronunciation and memorized phrases). Tests were administered to both groups at the end of the first year and at the end of the second year of language study. Not surprisingly, students who were taught with the grammar-translation method performed better in reading and writing tasks and the students who were taught using the Audiolingual Approach performed better in listening and speaking tasks.

This early study seems to underscore an important idea that would resurface many times over subsequent years. That is, the perceived effect of a particular mode of instruction may depend largely on the specific task that is used to measure that effect.

Though other studies during this period sought to examine the effect of formal language instruction on language development, many of these studies had major design flaws that made them difficult to interpret. For example, one such study was conducted by Upsher (1968), who compared three groups of students attending a summer session of law school at the University of Michigan. Alternate forms of an ESL proficiency test were administered at the beginning and end of the seven-week term. The groups were formed based on their initial performance on the test.

Students who received the lowest scores were placed into one group and received two hours of English language instruction each day in addition to their law classes. The second group received higher scores than the first group and participated in one hour of English language instruction each day. The third group of law students received the highest scores on the test and took no additional courses. At the end of the term, Upsher reported that while improvements were observed for all three groups, the amount of language instruction was not a significant factor in these gains. Though Upsher concluded that formal instruction may not be useful, it seems clear that his nonrandomized method for assigning students into groups makes it very difficult to draw any meaningful conclusions about the potential effects of instruction.

In another study conducted by Mason (1971), students who were required to take university and ESL classes concurrently because of their lower placement test scores were randomly assigned to one of two groups. Students in the first group supplemented their regular university classes with the required ESL classes and students in the second group were allowed to forgo the ESL classes. Mason reported that comparisons of pretest and posttest scores revealed no significant difference between those who had received the

ESL instruction and those who had not. Though Mason's design may have been somewhat of an improvement over Upsher's study, many variables in both of these studies were not controlled well, if at all. In addition, all of these groups spent more time studying in non-ESL classes without explicit language instruction than they did in the ESL classes with the instruction. This and the short duration of these studies could have diluted the potential effect of the explicit instruction.

Though not definitive in their conclusions, a number of subsequent studies have provided at least some evidence of the benefits of explicit instruction. For example, attempting to build on these early studies, Krashen, Jones, Zelinski and Usprich (1978) correlated the performance of 116 ESL students on a placement test with their total number of years of explicit language instruction and their total number of years living in a country where English is spoken as the native language. While no significant correlation was found between the years of residency and performance on the placement test, there was a strong correlation between the number of years of formal ESL instruction and the test scores. They concluded that explicit ESL instruction may be a greater predictor of English language proficiency than length of language exposure or residency in an environment of English speakers.

Wanting to test the effect of explicit grammar knowledge on a production task, Hulstijn and Hulstijn (1984) had learners retell a story to examine the effect of time pressure and focus of attention on the use of two word-order rules. They also conducted interviews with the informants to determine the level of explicit rule knowledge for each. Using a repeated measures design, they observed that learners who had more explicit knowledge of the grammar rules made fewer errors. However, they also noted that for

both groups, focus of attention had a significant effect on performance, but time pressure did not. In addition to suggesting a possible benefit to formal instruction, these results seemed to suggest that context in which learners use language may affect the quality of their performance.

Also interested in the connection between explicit instruction and the accuracy of production, Sorace (1985) studied 17 native English speakers who were learning Italian at two universities in Scotland. He hoped to see the effect of an environment with very little L2 input. Therefore, these locations in monolingual environments were particularly attractive since opportunities to practice outside of the classroom would be minimal. The elicitation instruments included a written test of metalinguistic ability, an oral description task with picture prompts and a simple oral interview. Sorace concluded that despite a lack of opportunities to practice L2 production, learning linguistic structures explicitly resulted in more native-like productions.

Similarly, Scott (1989) used implicit and explicit methods to teach two grammar structures to 34 university students who were learning French. Then she tested the students on their knowledge of the relative clauses and the subjunctive using an oral and written test that required students to fill in a blank. These posttest results demonstrated that the group who had received the explicit instruction made significantly greater progress overall than the group who was taught without the explicit method.

In addition, Green and Hecht (1992) further examined the effect of instruction on grammatical awareness with a much more substantial group of subjects. They used 300 native German speakers who had been studying ESL for three to twelve years. Learners were provided with a number of sentences that included 12 common types of

grammatical errors. Their task was to correct the sentences and then identify the grammar rule behind the error. While a group of 50 native speakers were able to rewrite the sentences correctly 96% of the time, overall the group of ESL learners were only able to rewrite the sentences accurately 78% of the time. Though the most proficient learners were able to correct 97% of the sentences, they were only able to state the grammar rule 46% of the time.

However, those learners who had received the most explicit instruction identified the correct rule 85% of the time. Though in 97% of the cases, learners who could correctly identify the grammar rule could also provide an accurate correction, 43 % of the time learners provided appropriate corrections without reference to explicit knowledge. The results of this study not only seem to highlight a possible benefit to explicit instruction, but they also seem to suggest that there may be a body of implicit language knowledge apart from that which is gained through explicit instruction.

In a study with some direct relevance to the current study, Frantzen (1995) examined the effects of explicit grammar instruction and corrective feedback on grammatical knowledge and the accuracy and fluency of writing. Four intermediate Spanish classes were used, two of which formed the experimental group and two that formed the control group, with a total of 44 students. A grammar test and a writing task were given to all of the students before and after the treatment. Results show that both groups experienced significant improvement on the grammar and written posttests. However, the performance of the treatment group was significantly better than the performance of the control group only in the grammar test. In other words, while

grammar knowledge increased, there was no significant difference between the written accuracy for the two groups.

Though the preceding studies seem to provide at least some evidence that explicit grammar instruction may be beneficial in certain contexts, a number of other studies have struggled to find such evidence. For example, Seliger (1979), building on the work of earlier studies, also sought to determine the extent to which grammar knowledge affects the accuracy of production. To do this, he elicited student responses that required an obligatory indefinite article. Then students were invited to explain the grammatical rules for using indefinite articles. It was assumed that the explanations provided by the learners demonstrated their conscious knowledge of the grammar rules. Interestingly, Seliger claimed, “No relationship was found between performance and having a rule” (p. 366). Considering these results from a cognitive perspective, Seliger concluded that while conscious awareness of grammar rules may serve an important function, such awareness probably does not help learners to monitor language production.

Continuing this line of research, others also failed to find the connection between explicit instruction and language performance. For example, Alderson, Clapham and Steel (1997), who studied university students who were learning French, were unable to find any evidence that students with greater grammatical knowledge of the language are better at using French or that they learn French faster than those who lack the same level of grammatical awareness. Similarly, after having learners complete a test of explicit grammatical knowledge and a production test, Han and Ellis (1998) found comparable results. They indicated that their findings supported the claim of Bialystok (1982), who argued that different language tasks utilize different kinds of L2 knowledge and

knowledge about grammar rules has little bearing on L2 proficiency. This notion also seems in harmony with the findings of Macrory and Stone (2000), who explored differences between knowing the grammar rules for using the perfect tense in French and actual production. They observed that in grammar tests students used the structure, but in oral and written production tasks, it was often left out entirely.

In a more recent study, Macaro and Masterman (2006) also sought to understand the effect of explicit grammar instruction on grammar knowledge and writing proficiency. Prior to beginning their regular university studies where they would study French, 12 native English speakers were given a five-month intensive course focusing on explicit grammar instruction. Both the treatment group and a control group were tested three times during the instructional period. The results showed that explicit grammar instruction led to some significant improvements in particular aspects of their grammar knowledge, but that it did not result in improvements in the grammatical accuracy of their writing. They concluded the following about developing grammatical accuracy in writing: (a) it involved a process that cannot be hurried, (b) development varies by individual, and (c) it “requires continuous exposure to both positive and negative evidence in both receptive and productive tasks” (p. 321).

Despite these studies, however, there continues to be confusion about the place of grammar in our L2 instruction. While some studies point to possible benefits of explicit grammar instruction, others fail to see its influence in productive tasks such as writing. As Musumeci (1997) has indicated, it seems that conflicting results from research studies often leave teachers confused about what should be done in the classroom. This is reflected in comments by Ellis (2006) in the twenty-fifth anniversary edition of the

TESOL Quarterly, in which he reminds the theorists, researchers and practitioners that the field has yet to determine *whether* grammar should be taught explicitly; and if it should be taught, he suggests that we still need to identify *what* should be taught, *when* it should be taught and *how* it should be taught.

Error Correction

In addition to background about writing and grammar instruction, another important part of this literature review relates to error correction in L2 writing. While the need to help students write with greater grammatical accuracy has been a topic of notable interest among ESL teachers and researchers, it has not been without controversy. More than a decade ago, Truscott (1996) launched a popular debate with his claim that grammar correction should be eliminated from L2 writing classes. The basis for his assertion arose from a growing number of studies that have been unsuccessful in providing meaningful evidence that error correction improves the accuracy of student writing (for examples see Polio, Fleck & Leder 1998; Robb, Ross & Shortreed, 1986; Semke, 1984; Sheppard, 1992).

Truscott went on to make three insightful observations to help substantiate his position. First, he argued that the common approaches to grammar correction ignore research about L2 learning that suggests that the process by which learners acquire various grammatical structures is slow and complex. Second, he pointed out that many teachers are unable or unwilling to provide adequate feedback to students and that even when feedback is given, students are often unwilling or unable to utilize it effectively. Third, he suggested that grammar correction is inefficient because it wastes valuable time and resources that could be used for more productive learning activities.

Subsequently, Truscott's assertions initiated a flurry of debate over the appropriateness of grammar correction in L2 writing (Ellis, 1998; Ferris, 1999, 2002, 2004; Ferris & Hedgcock, 1998; Truscott, 1999). Ferris (1999) went on to suggest that Truscott may have been a bit hasty in his conclusions and that error correction has helped some students in limited contexts. Subsequently, some have questioned the validity of some of Truscott's conclusions (Chandler, 2003) and others have advocated caution in interpreting Truscott's claims based on subsequent studies (Bitchener, Young, & Cameron, 2005). Ultimately, Ferris and Truscott agreed that further research was needed to help us better understand some of the potential effects of error correction on L2 writing. They suggested that studies should examine whether particular approaches to corrective feedback lead to greater accuracy and whether such approaches will result in greater performance with certain grammatical forms than others (Ferris, 1999; Truscott, 1999).

To clarify some of this research, it may be helpful to define some of the terms associated with corrective feedback in the literature. Two important terms are *direct* and *indirect* feedback (Ferris & Hedgcock, 1998; Ferris & Roberts, 2001; Lalande, 1982; Robb, Ross, & Shortreed, 1986; Terry, 1989; Zamel, 1985). Though Ferris (2006) points out that such expressions have not always been used consistently among researchers, generally speaking, direct feedback is provided when a teacher gives the student a particular correction and indirect feedback is provided when the teacher simply marks the error but does not correct it. In providing indirect feedback, some teachers tend to *code* mistakes to indicate the precise location and type of error, while others provide *uncoded*

feedback that simply locates the error without disclosing the error type. Usually with uncoded feedback, it becomes the student's task to diagnose and correct the mistake.

However, despite the feedback that might be offered, not all ESL students may be able to use that feedback equally well. For example, students with lower proficiency levels may not have adequate linguistic awareness to correct mistakes, even if they are identified for them (Ferris, 2006, Ferris & Hedgcock, 1998). This could lend some support to the claims of Truscott (1996), who has argued that error feedback may be harmful. Nevertheless, Ferris (2004) has suggested that since students have demonstrated an overwhelming desire for feedback, to withhold feedback may be detrimental due to legitimate affective concerns that may undermine the learning process.

Though most learners want and expect feedback from their teachers, there is evidence to suggest that they tend to prefer direct over indirect feedback (Ferris & Roberts, 2001; Komura, 1999; Rennie, 2000; Roberts, 1999). However, there appears to be some evidence that suggests that indirect feedback may result in accuracy levels that are at least as effective, depending on what is being analyzed (Ferris & Helt, 2000; Frantzen, 1995; Lalande, 1982; Lee, 1997; Robb, Ross & Shortreed, 1986).

For example, Ferris (2002) observed that though direct feedback led to greater accuracy in text revisions, indirect feedback resulted in the production of fewer initial errors. Thus, some have suggested that students might be served best when the method of feedback is dictated by the error type and context (Chaney, 1999, Ferris, 2006; Hendrickson, 1980).

In addition, Ferris (1999, 2001) distinguished *treatable* errors from *untreatable* errors. Treatable errors are those that can be prevented through the application of

systematic grammar rules. These include verb tense and form, subject-verb agreement, article usage, plural and possessive noun endings, and sentence fragments. Untreatable errors are those that result from ignorance of idiosyncratic language rules that must be acquired over time. These would include many word choice and sentence structure errors.

In testing the value of these distinctions, Ferris, Chaney, Komura, Roberts and McKee (2000) report a number of mixed but useful results from error correction (as cited in Bitchener, Young and Cameron, 2005). For treatable errors, there was a dramatic improvement with verb tense and form along with a slight improvement with noun ending errors and worse performance with article errors. For untreatable errors, there were slight improvements from earlier lexical errors and worse performance with sentence structures. Also, in the analysis of text revisions, Ferris and Roberts (2001) found fewer verb and noun ending errors as well as greater accuracy in the use of articles.

Answering the call of Ferris (2004) for more research on the effect of corrective feedback, Bitchener, Young and Cameron (2005) examined whether the kind of feedback given learners affects their writing accuracy. They used 53 migrant learners, who were placed into one of three groups which met for 20, 10 or 4 hours per week respectively. The researchers hasten to note that despite the varying amounts of total class time, all three groups spent 4 hours per week on writing and grammar. The first group included 19 students, who received direct written feedback along with a five-minute conference with the researcher after completing each new composition. The second group included 17 students, who only received direct written feedback. The third group included 17 students, who were only given feedback on the quality of their content and organization, rather than feedback on the linguistic accuracy of their writing.

After a twelve week period, learners were asked to produce a novel piece of writing. Three kinds of errors were analyzed including the definite article, prepositions, and the simple past tense. These error types were chosen for analysis based on the fact that they represented the three most frequent error types in the initial composition. The researchers note that there were considerable inconsistencies in accuracy levels among the four pieces of writing used for the study. Though no overall effect was observed when the three error types were combined, the researchers reported that the combined effect of the written feedback and the conferences was significant for the definite article and the simple past tense. These and other recent findings suggest that certain kinds of error correction in particular contexts may be useful. Yet, it seems that there may be much more that is not well understood about the effects of various approaches to error correction on L2 writing.

Seeking to expand our knowledge of how error feedback may affect L2 writing, Ferris (2006) used 3 experienced teachers to study the writing of 92 ESL students, most of whom were pursuing undergraduate degrees. While 20% of the group was made up of international students, 80% were long-term residents of the United States. Though males and females were represented fairly evenly, nearly two thirds of the students were from Asian countries. The specific questions of the research dealt with short-term and long-term improvements, whether the feedback offered by teachers was complete, whether the various strategies teachers use to give feedback made a difference on L2 writing and whether error treatment affected different types of errors differently.

Ferris pointed out that students addressed over 90% of the errors identified by their teachers and that 80% of those revisions that were based on teacher feedback were

corrected appropriately by the students. She also reported that, according to independent researchers, the instructor feedback was “overwhelmingly accurate” (89.4%) and dealt with 83% of the errors (p. 83). Ferris concluded that these results “do not support the claims of previous researchers that teachers give incomplete and inaccurate error feedback and that students ignore teacher feedback or cannot utilize it effectively in revision” (p. 83).

In examining the actual error feedback provided by the teachers, Ferris found that the feedback included direct feedback, where the teacher gave the students the corrections, and indirect feedback where the errors were identified without the corrections. The indirect feedback included the 15 common error correction codes identified for use in the study, some additional, less common codes and some corrections deemed as unnecessary by the independent researchers. The identified errors include the following: word choice, verb tense, verb form, word form, articles, singular-plural, pronouns, run-on, fragments, punctuation, spelling, sentence structure, informal, idiom, subject-verb agreement and a miscellaneous category. Interestingly, treatable errors received indirect feedback in approximately 59% of the cases while untreatable errors received direct corrections in approximately 65% of the time. Ferris hypothesized that perhaps teachers instinctively give different types of feedback based on the type of error the student makes and what the teacher believed would be the most helpful to the student.

Although this study produced useful insights about the effects of feedback on L2 writing, there are also some obvious limitations. First, since only three teachers were used in this study, it is clear that the findings should not necessarily be generalized to other teacher populations. Second, since this study used multiple drafts of a particular

composition, it appears to examine the effect of feedback on student error correction, rather than on student writing itself. Although it is useful to study how students respond to corrective feedback through multiple drafts of a composition, an additional question that seems at least as important is whether these efforts help students to produce fewer errors in a new piece of writing, as was attempted in the study conducted by Bitchener et al. (2005). It seems that, ultimately, one of our primary goals should be to strive to provide our university-bound students with strategic skills sets that are portable and that L2 writers can use effectively to edit their own work without the assistance of the ESL teacher.

Though many studies have examined the effect of error correction on L2 writing, most have had a number of weaknesses that have made it difficult to interpret the results with a high level of confidence. This is particularly true when attempting to draw conclusions from the collective findings of these studies. For example, as Ferris (2004) indicates, many of these studies lacked a control group of learners who did not receive corrective feedback. Another potential weakness has been that many of these studies did not examine a new piece of writing.

Ferris (2004) pointed out that other challenges in comparing one study with another include problems with the sizes of the treatment and control groups, the length of the treatment, the types of writing examined, the kinds of feedback provided, who provided the feedback, and the methods for identifying errors and measuring improvement. Because of these weaknesses and inconsistencies, Ferris goes on to describe the state of research on error correction in writing with the following:

. . . despite the published debate and several decades of research activity in this area, we are virtually at Square One, as the existing research base is incomplete and inconsistent, and it would certainly be premature to formulate any conclusions about this topic. (p. 49)

Ferris (2004) also observed that most researchers studying L2 writing error feedback during the last few of decades have been “operating in a vacuum” (p. 55). She lamented the lack of a concerted and systematic approach to investigating the relevant questions and calls for greater care in the design and reporting of future studies to ensure that they are replicable. She outlined her recommendations in the following:

Specifically, what is needed, going forward, are studies that carefully (a) report on learner and contextual characteristics; (b) define operationally which errors are being examined (and what is meant by “error” to begin with); (c) provide consistent treatments or feedback schemes; and (d) explain how such errors (and revisions or edits) were counted and analyzed systematically. (p. 57)

In addition to outlining her recommendations on how to proceed with future research, she also identified specific questions that she believes should guide these future research efforts. Though drawing attention to some very preliminary evidence relating to some of these questions, she pointed out that current efforts to answer such questions have been entirely inadequate. Her proposed research agenda includes the following:

1. Is there a difference in student progress in accuracy if students are allowed or required to revise their papers after receiving feedback?
2. Does supplemental grammar instruction (especially if it is tied to the concerns or error categories addressed in teacher feedback) affect student progress?

3. Does charting of written errors help students to engage cognitively in error analysis and facilitate long-term improvement?
 4. Are certain types of errors (lexical, morphological, syntactic) more amenable to treatment than others?
 5. Does the relative explicitness of teacher feedback (direct, indirect, location, labeling, etc.) have an impact on student uptake and long-term progress?
- (pp. 57-58)

Measures of L2 Writing Production

Although the primary focus of this study deals with the effect of manageable and immediate feedback on L2 writing accuracy, we have also highlighted the need for L2 writers to develop a high level of rhetorical competence. Without adequate rhetorical skills, a high level of writing accuracy would not be sufficient to help the L2 writer to produce quality writing. In addition to linguistic accuracy and rhetorical competence, writing fluency and writing complexity are also commonly used by researchers to measure writing development (for examples see Bonzo, 2005; Ellis & Yuan, 2004, Larsen-Freeman, 2006; Ojima, 2006; Spiliotopoulos, 2003). Though these notions of rhetorical competence, writing fluency and writing complexity are only secondary to accuracy in this study, it was assumed that including them would help contextualize findings and expose possible unintended consequences of the treatment on L2 writing production.

For example, since the treatment required participants to write and rewrite every day, it seemed reasonable to think that the fluency and complexity of their writing might improve over time. However, since participants were quite aware of the emphasis on

linguistic accuracy, it seemed equally plausible that potential gains in fluency and complexity might be stifled or even reversed due to excessive monitoring or avoidance of structures with which students may not have been comfortable.

Therefore, it was assumed that answering these additional questions of rhetorical competence, fluency, and complexity would help contextualize findings about the effect of the treatment on linguistic accuracy. Subsequently, it was necessary to find or create appropriate measures of each of these indicators of writing development. One useful resource came from Wolfe-Quintero, Inagaki and Kim (1998), who reviewed 39 studies to analyze the validity and reliability of more than 100 objective measures of L2 writing accuracy, fluency and complexity as correlated with L2 writing proficiency. Additional help in this search for appropriate measures came from the work of others such as Ortega (2003), who reviewed 25 studies of writing complexity. The following includes a brief discussion of some of the most common measures for writing accuracy, fluency and complexity.

Writing accuracy. Wolfe-Quintero et al. (1998) defined accuracy simply as “the ability to be free from errors while using language to communicate” (p. 33). In search of the most appropriate measure of accuracy, they examined 42 measures based on a variety of frequencies, ratios and indices. Since the primary question in this study dealt with the effect of the treatment on linguistic accuracy, the two measures favored most by Wolfe-Quintero et al. were used with the hope that each would present a complementary picture of L2 writing performance. Each of these measures will be described below.

The first measure of accuracy they recommended was the error-free T-unit ratio (EFT/T), or the total number of error-free T-units per total number of T-units in a given

piece of writing. They point out that while the EFT/T generally has not been effective at showing short-term changes, it has been an important research tool and that a majority of the studies they examined demonstrated high and moderate correlations with measures of L2 writing proficiency. For convenience and uniformity in this study, this and many of the other measures were converted to a 100-point scale. Thus, this measure of overall accuracy was calculated as (EFT/T) multiplied by 100. Since this and a number of subsequent measures utilize the T-unit, a brief discussion of the T-unit may be useful.

The T-unit was originally developed by Hunt (1965) as a way of measuring writing maturity to overcome problems associated with using sentences as units of production. Hunt observed that less mature writers would often generate run-on sentences that were simply coordinated with *and*. Such practices distorted sentence boundaries and made it difficult to analyze and interpret data. For example, writers with inadequate punctuations skills seemed to be more advanced because their sentences appeared relatively large and complex. Hunt defined a T-unit as “one main clause plus the subordinate clauses attached to or embedded within it” (p. 49). For example, the two-word sentence *Bill went* contains one main or independent clause and would be considered one T-unit. On the other hand, consider an expanded version of this sentence: *Before coming home, Bill went to the library*. Though this sentence also contains a subordinate or dependent clause, it would still be counted as only one T-unit.

However, consider one additional expansion, albeit erroneously punctuated: *Before coming home, Bill went to the library and he checked out several books and he went to his apartment and he studied most of the night*. Though punctuated as one sentence by the writer, it actually contains four T-units as identified in the following

breakdown: (a) *Before coming home, Bill went to the library*, (b) *he checked out several books*, (c) *he went to his apartment*, and (d) *he studied most of the night*. Thus, analyzing T-units rather than sentences provided researchers with a more stable measure of writing development.

Notwithstanding this straightforward definition of the T-unit, Wolfe-Quintero et al. (1998) point out that various researchers have presented conflicting interpretations of the T-unit when dealing with various sentence structure errors. For example, Bardovi-Harlig and Bofman (1989) and Tapai (1993) counted sentence fragments as T-units if they had been punctuated as a sentence by the writer, while Hirano (1991), Ishikawa (1995) and Vann (1979) suggested that fragments should not be counted as a T-unit. Similarly, Hunt (1965) counted T-units across multiple sentence boundaries according to the punctuation provided, while Homberg (1984) and Ishikawa (1995) only counted T-units within sentence units as dictated by the punctuation of the L2 writer. Despite these conflicting definitions, it seems quite possible that these various approaches may be more or less appropriate depending on the specific purpose of the measurement.

For the purpose of measuring overall accuracy in this study, fragments were counted as a T-unit in the sense that a fragment represented an unsuccessfully attempted T-unit. The rationale for this approach is that it would provide greater discrimination of accuracy. For example, consider a native English speaker who produced 30 error-free T-units out of 30 total T-units. Now consider an L2 writer who produced 25 error-free T-units and 5 fragments out of 25 T-units as defined above. If the fragments are not counted, then the accuracy scores for both writers would be 100.00 [(30/30)100 and (25/25)100, respectively]. However, if the fragments are included, then the L2 writer's

score would be 25 EFTs out of 30 T-units $[(25/30)100]$ or 83.00, most likely a more appropriate reflection of writing accuracy.

Similarly, for the purposes of this study, run-on sentences were analyzed according to the number of T-units they contained. However, each T-unit needed to have an appropriate form of punctuation preceding and following it before it could be considered error free. For example, if a run-on sentence contained three T-units but lacked appropriate punctuation that would have correctly separated the T-units, then the run-on would be counted as three T-units with no error-free T-units. Of course, it should be remembered that the presence of any type of error would make a particular T-unit ineligible to be counted as an EFT. Where multiple T-units were strung together with coordinating conjunctions (i.e. and, or, but), the conjunctions were counted in the T-unit that followed it. Using the EFT/T in this way provided one consistent, objective measure of overall accuracy of student writing.

In addition to this general measure of L2 writing accuracy, another measure of writing accuracy was used in this study as recommended by Wolfe-Quintero et al. (1998). This consisted of the total number of errors per the total number of T-units (E/T). While the traditional approach to examining E/T has involved one overall measure of error production, two innovations were incorporated in this study. First, rather than using E/T to measure the overall inaccuracy of a piece of writing, this approach was used to examine varying performance levels among seven different types of errors within three error families as illustrated in Figure 2. It was hoped that such an approach would provide insight on the effect of the treatment on specific error types.

<p><u>I. Grammatical Error Family</u></p> <p>Sentence Structure Errors</p> <ol style="list-style-type: none"> 1. Run-on sentences 2. Incomplete sentences 3. Sentence-level punctuation <p>Determiner Errors</p> <ol style="list-style-type: none"> 1. Articles 2. Possessive nouns/Pronouns 3. Numbers 4. Indefinite pronouns 5. Demonstrative pronouns <p>Verb Errors</p> <ol style="list-style-type: none"> 1. Subject-verb 2. Verb tense 3. Other verb form problems <p>Numeric Shift Errors</p> <ol style="list-style-type: none"> 1. Count-non-count 2. Single-plural <p>Semantic Errors</p> <ol style="list-style-type: none"> 1. Unclear Meaning 2. Awkwardness 3. Word order 4. Insertion/omission 	<p><u>II. Lexical Error Family</u></p> <p>Vocabulary Errors</p> <ol style="list-style-type: none"> 1. Word Choice (spelled correctly but wrong word) 2. Word Form (spelled correctly but wrong form of an appropriate word) 3. Prepositions (spelled correctly but wrong preposition) <p><u>III. Mechanical Error Family</u></p> <p>Mechanical Errors</p> <ol style="list-style-type: none"> 1. Spelling (misspelled) 2. Capitalization 3. New paragraph 4. Non-sentence level punctuation
--	---

Figure 2. Error families and error types used to analyze writing accuracy

Second, rather than focusing on the ratio of errors, or the inaccuracy of the L2 writing for a particular error type, the other innovation in this study was the use of the formula, $(1 - E/T)100$, to express the accuracy of performance (or absence of errors) for each error type. For example, if a student produced 6 determiner errors within 30 total T-units, the accuracy score for determiners would be $[(1 - 6/30)100]$, or 80.00. Such an approach produces a score that is expressed positively, rather than negatively, and is more comparable to the overall accuracy score illustrated earlier. Examples of these error types listed in Figure 2 can be seen in Appendix A.

Generally speaking, widely accepted guidelines of Standard English were used for error identification, and each mistake was counted as one error with no attempt to weight its egregiousness. Rather than weighting errors, it was believed that combining them into their respective error groups would provide a similar kind of information since some error families appear to be more problematic than others. Consider the follow example of a flawed production: *She watch sunset every night*. Here one error would be counted because the subject and verb are not in agreement and another error would be counted because of the missing determiner that would need to precede the word *sunset*.

However, it should be pointed out that such errors were only counted when the mistake was obvious or when a missing component was obligatory. For example, the preposition *over* in the grammatically acceptable sentence, *He came over my house*, would not automatically be counted as an error unless clear evidence from the context demonstrated that what had been written was not the intended meaning. If the rater had strong evidence, for instance, to assume that the writer intended to mean, *He came to my house* or *He came over to my house*, one error was counted. Thus, errors resulted from an

inappropriate inclusion, an inappropriate omission, or an inappropriate form of a word or phrase that otherwise would have produced an accurate construction.

The semantic error group, the last in the grammatical error family, requires some additional explanation. Since some of these error types affect meaning to varying degrees, an attempt was made to account for some of that variability while maintaining procedures that could be executed reliably. One error was counted for every word that was inserted inappropriately or every time an obligatory word was missing. One error was counted for every word order error where one shift (whether of one word or a group of words) could correct the error. For example, consider the sentence *I have for three years lived in the US*. This simple word order error could be corrected by one shift that inserts *lived in the US* between the *have* and the *for* to produce *I have lived in the US for three years*. Thus, one error would be counted. In addition, the notion of awkwardness was defined as a type of production error that was obviously distracting or conspicuously nonnative-like, though the meaning of the construction was clear to the rater. Such productions were also counted as one error.

Perhaps the most complex errors in the semantic error group were those labeled *unclear meaning*. These were calculated as the minimum number of words that would need to be revised to clarify the meaning of the production. To qualify for having a clear meaning, a particular word would need to make sense with the word preceding and following it. For example, consider the following construction: *After working all day, the work come TV bed sleep early*. The breakdown in this construction begins with the word *work*. Though the word *work* is preceded acceptably by the word *the*, the word *come* that follows it does not make sense after the word *work*. Therefore, the error counting begins

with the word *work* and continues through the word *sleep* for a total of 5 *unclear meaning* errors. The word *early* is not counted as an error because its preceding word, *sleep*, can fit appropriately with the word *early*. Though it seemed that this approach might yield relatively more errors than other groups, it was believed that such an approach would discriminate better than counting one error for an entire string of words that lacked a clear meaning.

In addition to these ungrammatical constructions, each word choice mistake was also counted as one error. For example, the word *universe* in the following sentence would have been rated as one error: *After four years of diligent study, the young man graduated from the universe.* However, no errors were counted unless the inaccurate nature of the word choice was obvious. For example, consider the sentence: *After getting a flat tire along the highway, he realized that he knew nothing about installing tires.* Though the more common collocation is *changing tires* rather than *installing tires*, the word *installing* seems adequate in this context, notwithstanding legitimate differences between the meanings of *to change* and *to install*. Another important point about evaluating word choice errors was that only correctly spelled words were eligible for the word choice error category. Otherwise such words were considered spelling errors rather than word choice errors.

As indicated, misspellings and mistaken punctuation were also counted as errors. Each mistake with punctuation or capitalization was also counted as an error, but only when they were obligatory in the specific context. For example, as a proper noun in the noun phrase “Mr. Brown,” capital letters would be required for “Mr.” and “Brown.” However, in other contexts capitalization was seen as optional such as for the word

following a colon. Similarly, a comma was needed to separate items in a series, to end a dependent clause that preceded an independent clause in a sentence, or to set apart sentence connectors such as *therefore* or *however*.

In addition to including these two measures of accuracy, this study also utilized three other measures of writing development including fluency, complexity and rhetorical competence. While the main focus of this study is L2 writing accuracy, these additional measures were included with the intent of providing a way to determine whether the treatment may have had an adverse or unintended effect on other important measures of writing development.

Writing fluency. Wolfe-Quintero et al. (1998) defined fluency as “a measure of the sheer number of words or structural units a writer is able to include in their writing within a particular period of time” (p. 14). They differentiate between fluency frequencies and fluency ratios and suggest that the latter is generally more meaningful. Of the nine measures examined across the several studies, the total number of words produced in a set time appears to be the most appropriate measure of fluency, though the authors point out some possible questions about the validity of this measure due to mixed results from the studies they examined.

They indicated that while 10 studies showed a high correlation between the number of words and proficiency level, and that one study showed a moderate correlation, seven studies demonstrated no correlation. However, they hastened to mention that all but one of these seven studies analyzed the writing of learners that were approximately at the same proficiency level, which may explain why no difference was observed. In addition, the authors mentioned that some of the studies suggest that there

may be a “ceiling effect” at the advanced level where the number of words may plateau or even decrease (p. 17). Though the authors did not address interplay among fluency, complexity and accuracy, these observations do not seem surprising in the context of this study since it is conceivable that fluency may be affected by student focus on accuracy. With an awareness of these potentially confounding effects in mind, fluency was simply defined in this study as the total number of words written in 30 minutes.

Writing complexity. In addition to their efforts to find the most effective ways to measure accuracy and fluency in L2 writing, Wolfe-Quintero et al. (1998) also analyzed 33 measures of L2 writing complexity in the form of various frequencies, ratios and indices. They defined complexity as “grammatical variation and sophistication” in three possible units of production including clauses, T-units, and sentences (p. 69). They differentiated between two types of complexity measures. The first analyzes these three production units in relation to themselves. Such measures might include the number of clauses per T-unit or the number of T-units per sentence. The second analyzes the occurrence of specific structures within these production units. Such measures might include the number of passives per sentence or the number of dependent clauses per T-unit.

Despite mixed results from their analysis, Wolfe-Quintero et al. favored the T-unit complexity ratio, (the total number of clauses per T-unit) for measuring L2 writing complexity since generally it appeared to increase along with proficiency. Of the 18 studies that utilized the T-unit complexity ratio, one was highly correlated with proficiency, six were moderately correlated, four were weakly correlated and seven showed no correlation. They also highlighted two additional measures as potentially viable alternatives including the number of dependent clauses per total clauses and the

number of dependent clauses per T-unit. Though these latter measures also showed a general linear increase with proficiency, they had been used much less frequently (3 studies each) compared to the more popular T-unit complexity ratio.

Despite their recommendations, Wolfe-Quintero et al. (1998) admit various difficulties associated with attempting to measure L2 writing complexity. First, they explained that a number of researchers have generated conflicting definitions for the units of production when attempting to measure complexity. This was particularly true for clauses. For example, while Hunt (1965) limited the definition of a clause to independent clauses, along with all dependent clauses, including nominal clauses, adverbial clauses, and adjective or relative clauses, Bardovi-Harlig and Bofman (1989) expanded their definition of a “clause” to include gerunds, participles, and infinitive verb phrases.

At the same time, others such as Homburg (1984) emphasized the difference between independent and dependent clauses but excluded nominal clauses as being embedded but not dependent. Still others such as Tapia (1993) distinguished among three types of clauses including independent, dependent and embedded clauses, and concluded that the latter included all adjective and nominal clauses. Needless to say, these varied definitions, emphases and categorizations have made comparing research findings or designing subsequent studies somewhat problematic.

However, in a more recent review of 25 studies of writing complexity, Ortega (2003) examined six of the same measures of complexity including:

. . . mean length of sentence [MLS], mean length of T-unit [MLTU], ...mean length of clause [MLC]..., mean number of T-units per sentence [TU/S]..., mean

number of clauses per T-unit [C/TU], and mean number of dependant clauses per clause [DC/C]. (p. 498)

Although these measures were among those previously analyzed by Wolfe-Quintero et al. (1998), Ortega claims that there is insufficient evidence to suggest than any one of these six measures is more valid than the others.

Moreover, despite the prominent role that clauses have assumed in measuring writing complexity in many studies, Rimmer (2006) presents additional concerns about such production units that go beyond the difficulties of conflicting definitions. First, he points out that structural and semantic ambiguity among cases of coordination and subordination of clauses leads to reliability problems when attempting to measure complexity. Second, he claims that clauses may often be too crude of a measure to capture subtle differences in writing development. Though his ultimate point is to encourage researchers to use corpus linguistics to inform complexity measurements (a notion beyond the scope of this study), his discussion of the limitations of traditional measures of complexity is quite insightful.

Interestingly, in his early research with T-units, Hunt (1965) concluded that the MLTU (or the average number of words per T-unit) was the best indicator of L1 writing development because it accounted for the highest percentage gain from one age group to another. Moreover, in assessing the value of the MLTU, Wolfe-Quintero et al. (1998) add:

[A] comparison of the means across studies show that here is a range from 6.0 words per T-unit for the lowest level learners to 23.0 for the most advanced, with word per T-unit increasing in a linear relationship with proficiency, regardless of

how proficiency was measured or whether the results were significant. This repeated sampling reliability of the linear nature of the words per T-unit measure across studies suggests that it may be a very useful measure indeed. (p. 25)

Despite their awareness of the effectiveness of the MLTU, however, Wolfe-Quintero et al. (1998) did not include it in their analysis of complexity measures because they interpreted it as a measure of fluency rather than complexity. Though they admit that “most researchers have treated T-unit length as a measure of grammatical complexity,” and that “T-units include complexity as part of their definition,” they defended their position by explaining that the MLTU does not identify the cause of length increase, and may or may not reflect greater grammatical complexity (p. 25).

Notwithstanding Wolfe-Quintero et al.’s rejection of the MLTU as a measure of complexity, it seems clear that the measure is robust and well suited for this study for at least three reasons. First, though the MLTU may not be a certain measure of *grammatical* complexity, at a minimum, it appears to be a strong measure of *linguistic* complexity. As such it seems quite adequate for the purpose of this study, which is to determine whether the treatment increased linguistic accuracy without diminishing writing complexity.

Second, since the MLTU deals with words (the smallest and most numerous unit of production) rather than clauses, it may be a more sensitive to smaller differences in L2 writer performance. In fact, based on her review of 25 studies of writing complexity, Ortega (2003) indicated critical magnitudes that show significant differences between proficiency levels such that as few as two words for the MLTU could indicate that writers may belong to different proficiency samples. Third, rather than attempting to grapple with the conflicting definitions, ambiguity and subjectivity associated with the analysis of

clauses, the MLTU may actually be much easier to calculate and more reliable. Because of these reasons, the MLTU was utilized in this study as the measure of linguistic complexity.

Rhetorical competence. In addition to the three objective measures of linguistic accuracy, fluency and complexity, the rhetorical competence of the observed writers was also assessed. To determine the level of rhetorical writing competence, a rubric was adapted from the TOEFL iBT (Test of English as a Second Language Internet-based Test), developed by the Educational Testing Service (ETS). The following includes a brief discussion of why this rubric was used.

Though the writing component of the iBT has only been around since 2005, it is largely an improved version of ETS's Test of Written English, which had been used since the mid 1980s. While a special rubric could have been created exclusively for this study, the iBT rubric was chosen because it is the product of years of refinement and has been used extensively to assess the writing of ESL learners at approximately the same proficiency level as the students included in this study.

The primary purpose of the TOEFL is to assess the English proficiency of nonnative speakers to determine their readiness to begin university-level study in English. ETS consistently provides TOEFL scores to over 6,000 institutions in 110 nations worldwide. Though the former version of the TOEFL did not measure productive language skills, the newer iBT includes a writing component. Based on data from the first year of use (September of 2005 to December of 2006), ETS reports a reliability estimate of .78 for the iBT writing component and a standard error of measure of 2.65 on a scale of 0 to 30 (Educational Testing Service, 2007).

Despite ETS's carefully planned approach to writing assessment, a holistic rating scheme such as this may not be without some controversy. Though a holistic rubric would probably help facilitate greater reliability among raters than an analytic rubric, some may argue such an approach would be inappropriate since writing is often seen as multidimensional. For example, Hamp-Lyons and Kroll (1996) warned that scoring procedures would be more valid when developers attend to the "mix of strengths and weaknesses often found in ESL writings" (p. 233). While such commentary seems appropriate in instructional settings where specific feedback needs to be given on the development of discrete skills, this approach seems less suitable in a research context where the objective is simply to measure one aspect of writing such as rhetorical competence. Moreover, it should be noted that outside of instructional settings, consumers of writing in authentic communicative contexts almost always view writing holistically for its global content rather than for its analytical components.

Another potential concern with the iBT rubric was whether it would be sensitive enough to detect subtle differences between various levels of writing. Though the rubric itself only has six categories (0-5), it should be noted that in practice, half levels are awarded when the average score of two trained raters are different by one score. For example, .5, 1.5, 2.5, 3.5, and 4.5 can also be awarded for a total of 11 possible scores. Thus, the 11 possible scores suggested by this rubric seemed to be adequate to capture a great deal of potential variation among writers within a fairly narrow proficiency level. For these reasons, the adapted ETS rubric was believed to be an appropriate instrument for this study. The adapted version of this rubric can be found in Appendix B.

Summary

Though developing writing ability is a primary objective of university training, many ESL writers struggle to produce adequate writing. This is particularly true of the challenge many face with linguistic accuracy. Although the process writing model seems valuable for helping ESL writers develop rhetorical writing competence and for meeting a variety of experiential objectives, it alone seems quite inadequate for improving linguistic accuracy. One particular reason for this may be the excessive number of errors that teachers and students attempt to manage at once. Another reason may be associated with the fact that corrected feedback is often delayed and occurs too infrequently to benefit the students.

However, the literature associated with the effects of grammar instruction and error correction in L2 writing is not entirely clear about how best to improve the accuracy of what ESL writers produce. While some have suggested that error correction is ineffective, or even harmful (Truscott, 1996, 1999, 2007), there appears to be some evidence that certain kinds of error correction may be useful in some contexts. Nevertheless, many of the studies that have pursued questions about error correction have had challenging flaws. Consequently, Ferris (2004) has proposed a fairly focused research agenda and has invited careful researchers to contribute toward a greater understanding of the effects of error feedback in L2 writing.

This study, therefore, seeks to contribute to this line of inquiry by testing the effects of a particular method of L2 writing pedagogy that aims to complement the benefits of process writing with learning activities designed to improve the linguistic accuracy of ESL writers. Though the specific details of this treatment will be addressed

in greater detail in Chapter 3, the following is a brief summary as it relates to the six research questions proposed above by Ferris. This treatment includes the following features: (a) students are required to revise writing after receiving feedback, (b) explicit instruction is tied to the needs of the learners in a dynamic syllabus that responds to student performance, (c) students are cognitively engaged in error analysis and chart their errors and their progress, (d) all errors of linguistic accuracy in student writing are identified by the teacher, (e) feedback is indirect but includes the type of errors and the location of those errors. In addition to these emphases suggested by Ferris, two central features of this treatment are: (f) the intent that the volume of errors is much more manageable for both teacher and students (because feedback is based on short, 10-minute writings), and (g) the fact that these compositions facilitate a fairly constant flow of feedback to students (because feedback is given on a daily basis).

Since the teachers and ESL writers participating in this study were part of real classes imbedded in a broader curriculum with specific goals and objectives, it was not possible to completely isolate every variable associated with the students' learning experiences. Rather, this method for teaching L2 writing was viewed as one treatment, though there were a variety of individual components. As Ferris (2004) put it, the dilemma is an ethical one. Given the various needs of the learners, a specific curriculum has been developed with the hope that it would provide student with the best learning experience possible. Thus, it would seem "unethical to withhold it . . . simply for research purposes" (p. 51). However, though not perfectly aligned with Ferris' research agenda, it was assumed that the treatment under investigation might shed light on at least some of the research question she has posed.

Though the central question of this study dealt with the effect of the treatment on the linguistic accuracy of L2 writing, another important question dealt with whether such an emphasis would diminish other measures of L2 writing development such as rhetorical competence, writing fluency and writing complexity. Since each of these measures help contribute to good writing, each has important implications for pedagogical practice in the classroom. It seems that gains in linguistic accuracy would be the most meaningful if they did not occur at the expense of other important features of good writing.

In addition, one question that emerged in the literature and that seemed to have important implications for programmatic assessment was the appropriateness of the multiple-choice grammar tests that are intended to measure grammatical competence. While the literature seemed inconclusive, it appeared that there was a body of evidence to suggest that traditional objective grammar tests may not be effective predictors of ESL learner performance in productive tasks. Of course, the irony is that productive tasks seem to be the most meaningful contexts in which grammatical accuracy would be important for the ESL learner.

This chapter has presented a variety of relevant literature to help contextualize this study. This review has addressed writing instruction, process writing, grammar instruction, error correction, as well as an examination of common methods of measuring L2 writing accuracy, fluency, complexity and rhetorical competence.

Research Questions

With these considerations in mind, we are prepared to form our research questions. Though the following questions are stated generally here, they will be defined operationally in the next chapter.

1. To what extent will the treatment produce greater linguistic accuracy in new writing when compared to the traditional instructional method?
2. To what extent will the treatment produce equivalent levels of fluency, complexity, and rhetorical competence in new writing when compared to the traditional approach?
3. What is the relationship between explicit grammar knowledge and grammar use in a productive writing task?

CHAPTER 3: METHOD

The purpose of this chapter is to describe the research methodology used to answer this study's research questions. It provides a description of the participants, including the students, the teachers, and those who scored or rated various aspects of the student writing. This chapter also presents a brief rationale for the research design, including an explanation of the design for establishing evidence of reliability. In addition, it contains a description of the instruments and elicitation procedures used to gather data. Finally, it presents an operationalized version of the research questions.

Participants

This section provides useful information about the 47 students who participated in this study as either members of the control group or treatment group. It also includes the background of the various teachers who taught these students before they took the pretest and posttest. Finally, this section briefly describes those who provided scores or ratings of student essays.

The students. The writing students used in this study included 47 Level 5 ESL students who were studying at Brigham Young University's English Language Center (ELC) in Provo, Utah. Level 5 represents the highest proficiency level at the ELC and, using the guidelines established by the American Council of Foreign Language Teachers, the proficiency level for most of these students was estimated to range from advanced-low to advanced-mid. While the writing instruction given to the control group took place over a 15-week summer semester, between May and August of 2006, the instruction for the treatment group occurred one year later, during the same 15-week semester in 2007.

The control group was made up of 19 students with ages ranging from 18 to 33, with a mean of approximately 25 years and 9 months. On the other hand, the treatment group included 28 students, nearly one and a half times the size of the control group, with ages ranging from 18 to 45, with a mean of approximately 24 years and 9 months. Table 1 summarizes the composition of the control and treatment groups in terms of native language and gender. While males and females are reasonably represented in the treatment group, it should be noted that females outnumber the males in the control group nearly four to one.

Table 1

Experimental Groups by Native Language and Gender

Native Language	Experimental Groups					
	Control Group			Treatment Group		
	Male	Female	Total	Male	Female	Total
Spanish	2	4	6	10	9	19
Korean	0	3	3	4	2	6
Mandarin	1	2	3	0	0	0
Portuguese	1	2	3	0	0	0
Japanese	0	0	0	1	1	2
French	0	1	1	1	0	1
Mongolian	0	1	1	0	0	0
Romanian	0	1	1	0	0	0
Russian	0	1	1	0	0	0
Totals	4	15	19	16	12	28

Since such a disparity is somewhat unusual in the Level 5 classes, and since the control group included a disproportionate number of females compared to the treatment group, a repeated measures ANOVA was used to analyze the effect of gender on accuracy scores derived from student writing (these scores were based on error-free T-units over the total T-units described in the previous chapter). Table 2 presents the means

and standard deviations for this analysis and Table 3 includes the results of the analysis in an ANOVA summary table. While these data suggest that mean accuracy scores for both males and females appears to have improved ($p = .02$), there was no significant difference between mean accuracy scores of males and females ($p = .96$). In the absence of any additional evidence that gender might influence the results of this study, the assumption was made that the disproportional number of females in the control group was not likely to affect the outcomes of this study.

Table 2

Descriptive Statistics for Gender and Accuracy Scores

Group		Pretest	Posttest	Means
Males (n = 20)	Mean	15.53	20.44	17.99
	SD	16.42	17.56	16.99
Females (n = 27)	Mean	14.54	19.64	17.09
	SD	11.11	17.64	14.38
Total (N = 47)	Mean	14.94	19.97	17.46
	SD	13.35	17.42	15.39

Table 3

ANOVA Summary Table for Gender and Accuracy Scores

Source	SS	df	MS	F	p
Between Subjects		46			
Gender	18.15	1	18.15	.046	.83
Error	17889.03	45	397.53		
Within Subject		47			
Time	567.38	1	567.38	6.02	.02
Time x Gender	.22	1	.22	.002	.96
Error	4241.19	45	94.25		
Total	4808.79	93			

In addition to our discussion of gender, some commentary about the L1s in Table 1 may also be helpful. This breakdown student L1s is useful for examining the potential effect of language distance, or the notion that similarities or differences between various L1s and English may account for at least part of the relative difficulty or speed with which a learner may acquire an L2 such as English (Odlin, 1989). Corder (1981) pointed out that, based on language distance, native speakers of western European languages such as Spanish would likely experience less difficulty learning English while native speakers of Asian languages such as Chinese, Japanese or Korean would likely experience greater difficulty. While the percent of native speakers of western European languages in the control group was just under 53%, the percent in the treatment group was just over 71%. In addition, the native speakers of Chinese, Japanese and Korean made up just over 31% of the control group and 29% of the treatment group.

Additional insights from Ringbom (1987) suggest that if language distance influenced performance levels of the respective groups at all, the influence would likely be rather small. First, he noted that L1 influence is stronger for younger learners than for older learners. Second, he observed that L1 influence is greatest for those with lower proficiency levels and less significant for those at higher proficiency levels. Third, he concluded that L1 influence is greater in highly communicative tasks and less significant when more monitoring takes place. Unlike those learners who would most likely be affected by language distance issues, students in this study were advanced-level adult learners who were engaged in writing tasks which allowed for substantial monitoring. Therefore, it was assumed that the influence of language distance on student performance would be minimal if not negligible.

The teachers. Due to the practical constraints of dealing with intact classes, it was not possible to control for teacher differences among the learners in the treatment and control groups. It is assumed that some teacher effect was present since different teachers instructed various students throughout the periods being examined. However, the following attempt was made to clarify and contextualize some of these potential differences.

Four teachers taught students in the control group winter semester of 2006 prior to their pretest, and three different teachers taught the same students in the summer semester prior to their posttest. Although all of the teachers had taught for at least three years, their level of experience and the number of students they taught varied by teacher. This information is provided in Table 4. Experience levels for teachers were defined according to the following: a “novice” teacher had taught for five or fewer years, an “experienced” teacher had taught for six to ten years, and a “veteran” teacher had taught for eleven or more years.

Table 4

Control Group Students by Term, Teacher and Teacher’s Experience

Semester	Teacher	Experience Level	Number of Students
Winter 2006 (Level 4)	A	Veteran	11
	B	Experienced	3
	C	Novice	3
	D	Experienced	2
	Total		19
Summer 2006 (Level 5)	E	Experienced	8
	F	Experienced	6
	G	Experienced	5
	Total		19

Table 5 provides the number of students who were taught by the several teachers during the treatment period and shows the experience level of each. All of the teachers who taught the students in the control group prior to their posttest were well experienced in teaching the traditional process writing approach. On the other hand, while Teacher P had previously taught students using the treatment method, Teachers Q and E had never used this approach before. Only one of the treatment group teachers (Teacher E) also taught students in the control group prior to their posttest. There were no other overlaps between teachers of students in the control group and treatment group. In addition, it should be noted that three students in the treatment group had not previously attended Level 4 classes but were new to the ELC when they were placed into Level 5 prior to the pretest at the beginning of the semester in the summer of 2007. For these students, the unknown teacher information is marked with an asterisk (*).

Table 5

Treatment Group Students by Term, Teacher and Teacher's Experience

Semester	Teacher	Experience Level	Number of Students
Winter 2006 (Level 4)	H	Experienced	9
	I	Novice	7
	J	Veteran	3
	K	Veteran	3
	L	Experience	2
	M	Veteran	1
	*	*	3
	Total		28
Summer 2007 (Level 5)	P	Veteran	10
	Q	Veteran	10
	E	Experienced	8
	Total		28

The scorers and raters. In an effort to estimate the reliability of the measures investigated in this study, the principle researcher was assisted by two additional individuals who helped score or rate essays or essay components. Both held master's degrees and had taught writing for a number of years at the ELC and other institutions. Both were well acquainted with the kinds of challenges L2 writers face in attempting to produce writing that is both accurate and rhetorically well developed. Additional information scoring and rating procedures are outlined in the subsequent section entitled *Reliability Design*.

Research Design

A pretest, posttest nonequivalent control group design was used for this study as described by Shadish, Cook and Campbell (2002). This design is illustrated in Table 6. Using a mixed model, repeated measures Analysis of Variance (ANOVA), the mean performance of students in the control group was compared with the mean performance of students in the treatment group (between subjects), and the mean performance of students on pretest measures was compared with the mean performance of students on posttest measures (within subjects).

Table 6

Pretest, posttest nonequivalent control group design

Experimental Group	Pretest	Treatment	Posttest
Treatment (32 Students in 2007)	O ₁	X	O ₂
Control (19 Students in 2006)	O ₁	~X	O ₂

Note: O = Testing Occasion, X = Experimental Treatment, ~X = No Treatment

A brief comment about repeated measures may be useful here. Variation in between-subjects comparisons can originate from the treatment, the individuals or the error associated with the experiment (Tanguma, 1999). However, since repeated measures observe the same individuals on multiple occasions, the individual variance is not included in the analysis. This results in greater statistical power and reduces the likelihood of making a Type II error, or failing to reject the null hypothesis when the alternative hypothesis is actually true (Stevens, 1996). For the purposes of this study, the mixed model ANOVA provided evidence to allow us to answer our research questions regarding differences between performance levels of students in the control and treatment groups.

The pretest and posttest observations illustrated in Table 4 include the essay written before the treatment or control and the essay written after these instructional periods were completed. Though each student produced only two essays, each was subjected to several analyses. For example, 12 measures were analyzed in this study. These included (a) accuracy scores and seven additional types of accuracy, including: (b) sentence structure accuracy scores, (c) determiner accuracy scores, (d), verb accuracy scores, (e) numeric accuracy scores, (f) semantic accuracy scores, (g) lexical accuracy scores, and (h) mechanical accuracy scores. Additional scores were also analyzed including: (i) writing fluency scores, (j) writing complexity scores, (k) rhetorical competence ratings, and (l) grammar knowledge scores.

To compute the mixed model ANOVA needed for this study, the Statistical Package for the Social Sciences (SPSS) was used. With a significance level set for .05, the within subject factor was labeled *Time* and included two levels (pre and posttest

observations). The between-subjects factor was labeled *Group* and also had two levels, the control and treatment group.

Instruments

To answer the research questions relevant to this study, student writing had to be assessed for (a) linguistic accuracy, (b) fluency, (c) complexity and (d) rhetorical competence. To do this, a number of instruments and procedures were devised as described in the previous chapter. The linguistic accuracy category was further broken down into eight separate components. A summary of each of the dependent variables along with its method of measurement is included in Table 7. While measures of linguistic accuracy, fluency and complexity, were determined through careful analyses of student writing, two instruments were used to elicit data for the rhetorical competence scores and the grammar knowledge scores. These include the rhetorical competence rubric and grammar knowledge test, each of which will be discussed briefly.

Table 7

Dependent Variables and Their Methods of Measurement

Dependent Variables	Method of Measurement
1. Overall Accuracy	(Error-free T-units/total T-units)
2. Sentence structure accuracy	$[1 - (\text{number of errors}/\text{total T-units})] \times 100$
3. Determiner accuracy	$[1 - (\text{number of errors}/\text{total T-units})] \times 100$
4. Verb accuracy	$[1 - (\text{number of errors}/\text{total T-units})] \times 100$
5. Numeric accuracy	$[1 - (\text{number of errors}/\text{total T-units})] \times 100$
6. Semantic accuracy	$[1 - (\text{number of errors}/\text{total T-units})] \times 100$
7. Lexical accuracy	$[1 - (\text{number of errors}/\text{total T-units})] \times 100$
8. Mechanical accuracy	$[1 - (\text{number of errors}/\text{total T-units})] \times 100$
9. Fluency	(number of words written in 30 minutes)
10. Complexity	(Mean length of T-units/total T-units)
11. Rhetorical competence	(ratings based on Adapted iBT Rubric)
12. Grammar knowledge	(scores on grammar test)

The rhetorical competence rubric. To determine the level of rhetorical competence for student writing, an altered version of ETS's iBT rubric was chosen as was discussed previously. Though the rubric was adapted slightly for use in this study, nearly 80% of the original rubric content remained intact. Essentially, the only adaptations that were made to the rubric were the deletions of references to linguistic accuracy since these were to be assessed through the measures of writing accuracy discussed previously. The remaining content of the rubric included references to the same kinds of rhetorical features taught at the ELC and many other institutions that teach process writing. These included emphases such as effectively addressing the topic or task, organization and development, appropriate examples, details or support, and unity and coherence. The rhetorical competence rubric can be found in Appendix B.

The grammar knowledge test. The other instrument used in this study was the multiple choice grammar test administered to the students in both the control and treatment groups at the completion of their semester of Level 5 classes. This exam consisted of 75 multiple-choice items that tested student ability to identify correctly and incorrectly formed grammatical structures. Each item provided students with four alternatives, labeled "a," "b," "c" or "d." After reading each item in their exam, students selected a response by filling in the corresponding circle on their answer sheet. Tests were then scored by computer.

Of the 75 items included on the test, 61 items presented the students with an incomplete sentence. The student's task was to choose the alternative that completed the sentence grammatically. These items used the format illustrated below:

We will _____ to New York for the New Year's celebration.

- a. been traveling
- b. be traveling
- c. traveling
- d. be travel

Eight of the items required the students to identify the mistake in a sentence by choosing the appropriate letter as illustrated in the following sample:

I am going to do my homework after dinner last Wednesday, but I fell asleep.

A

B

C

D

The remaining six items were all part of the same response set associated with a short paragraph. Like the first item type, students were required to choose among alternatives to form grammatical sentences, though in this case students needed to operate at the paragraph level rather than simply the sentence level.

Reliability Design

For the findings of this study to be meaningful, it was necessary to provide appropriate estimates of reliability for the included measures. Of the 12 dependent variables examined in this study, 11 were based on scores derived from those methods outlined above in Table 4. However, the rhetorical competence variable was based on ratings rather than scores. The different approaches used for estimating the reliability of the scoring and the reliability of the ratings are described below.

Scoring. The scoring simply involved counting the specific number of the various occurrences being examined. Though computer analysis provided the grammar knowledge score and the number of words in each essay needed to calculate writing fluency, human scoring was used for the remaining scored variables. Since nine of the measures were based on the number of T-units, it was essential that the T-units be counted accurately. Therefore, a criterion of absolute agreement for the number of T-

units for each essay was established between Scorer 1 (S1), the principle researcher, and Scorer 2 (S2), a credentialed ESL teacher trained to use the scoring criteria outlined previously. Though it was not possible for S1 to be a completely blind scorer due to his involvement in data management, S2 was totally blind to student, group and essay. It was determined that if any discrepancies emerged from a particular essay, S1 and S2 would reexamine the essay together and decide the number of T-units jointly.

After the number of T-units for each essay was established, the first eight measures listed in Table 4 still required additional scoring. While S1 scored all 94 essays on each of the eight measures, S2 scored just over half of the essays as outlined in Table 8. In an attempt to have the essays that would be scored by S2 reflect the variability of the larger population, essays were chosen through simple random sampling from one of six stratified groups. The strata were determined by essay (pretest and posttest) and proficiency level for each group. Proficiency level was based on teacher ratings that were submitted along with student grades at the conclusion of the semester. Since each student had four teachers, the four proficiency ratings were averaged and this overall proficiency rating was used to place students into one of three proficiency levels: *low*, *middle* and *high*.

Three control group essays and five treatment group essays were chosen from each stratum in an attempt to reflect the approximate ratio of students in the control group compared with students in the treatment group. A Pearson correlation coefficient was then produced for each of the eight scored variables based on counts provided by S1 and S2. The results of these correlations are reported in the following chapter.

Table 8

Stratification for the Second Scorer's Random Sampling

Testing Occasion	Proficiency Level	Control Group Students	Treatment Group Students
Pretest	High	3	5
	Middle	3	5
	Low	3	5
Posttest	High	3	5
	Middle	3	5
	Low	3	5
	Total	18	30

Rating. Unlike the scoring of essays used to establish accuracy, a different approach was taken to estimate reliability of ratings used to establish rhetorical competence. Though a number of methods might have been used, two valuable approaches are utilized in this study: (a) the Many-facets Rasch Model (MFRM) and (b) the intraclass correlation. Each is discussed briefly below.

Perhaps the most informative approach is the MFRM. Building on the seminal work of Rasch (1960), Linacre (1994) developed the MFRM in an effort to deal with inequitable cases of rater severity and leniency. Unlike a traditional interrater correlation coefficient, the MFRM can provide a wealth of additional information as it seeks to account for differences in the ability of examinees, the difficulty of the various items, and both interrater and intrarater inconsistencies. In their introductory text, Bond and Fox (2007) explain that Rasch modeling also enables researchers to measure the interactions among these facets. Thus ratings can be adjusted for fluctuations in item difficulty or rater severity.

Though many traditional rating scales use integer values, such values rarely reflect equidistant intervals. Therefore, the MFRM utilizes the logit value, a true interval measure based on a probabilistic log-linear scale that allows researchers to map examinee ability, item difficulty, rater severity and expected ratings all on the same scale. With this in mind, the MFRM constructs a model in an attempt to account for the data. A rater who is too severe, too lenient or too inconsistent in his rating will not fit the model. In addition, Linacre (1994) points out that either too much or too little error variance also undermines validity. For example, a very low variance for a particular rater would indicate his tendency to cluster his ratings at the center of a scale, resulting in ratings that would be less discriminating and less useful.

One important feature of the MFRM is its use of fit statistics. Linacre (1994) explains that when the model fails to account for enough observed variation, “misfits” are identified from a “mean-square fit statistic, based on the ratio of observed error variance to modeled error variance” (p. 10). While the expected value to the fit statistic is 1.0, they have a possible range from 0 to infinity. Wright and Linacre (1994) suggest that for high-stakes testing, mean square values should be within a range of .8 and 1.2. A mean square statistic of .8, for example, would indicate that the rater demonstrated 20% less variance than was predicted by the model, and a mean square of 1.2 would indicate 20% more variance than predicted. For clinical observations, such as the current study, Wright and Linacre suggest that a range of .5 to 1.7 may be acceptable.

Others such as Pollitt and Hutchinson (1987), Lynch and McNamara (1998), Park (2004), and Kim (2006) have recommended a similar but more precise test. For example, Kim explains that “a value lower than the mean minus twice the standard deviation would

indicate too little variation, or overfit, while a value greater than the mean plus twice the standard deviation would indicate too much unpredictability, or misfit” (p. 22). Since the model cannot correct for misfitting data, such data should be carefully analyzed by the researcher to identify how these problems might be corrected.

Another feature of the MFRM is the reliability separation index. Unlike the conventional Pearson correlation coefficient, for example, where higher values would indicate greater strength in the linear relationship between two raters, Myford and Wolf (2004) explain that the reliability separation index shows the amount of undesirable variance due to statistically significant differences in severity or leniency among raters. While a higher index would be appropriate for examinees, a lower value would be more desirable among raters and would indicate greater rater stability.

Though the most informative rating design would be fully crossed, where each rater would provide ratings on each essay for each student, such a design was not feasible in the context of this study. Moreover, it was noted that the rating design requirements for the MFRM and the intraclass correlation are not the same. Therefore, the following includes a brief description of how the rating was designed to meet both sets of requirements. In describing rating design requirements for the MFRM, Schumacker (1999) stated that traditional rating designs that nest ratings within task are inappropriate for comparing facets because they lack the requirement of connectivity or linking. However, he pointed out that “a mixed design can be used to achieve a common vertical ruler when the frame of reference permits commensurate measures to be linked” (p. 323).

He went on to explain that a useful method “to achieve connectivity for the creation of a common vertical ruler is to have at least one judge crossed with all elements

of the facet” (p. 325). In other words, if one rater were to rate all students on all tasks, then a fully crossed design would not be necessary if at least one additional rater completed ratings for each student on each task. Therefore, the most efficient (and minimally acceptable) use of raters would utilize one rater (R1) who would rate all students on both the pretest and posttest essays. A second rater (R2) would rate all of the pretest essays, and a third rater (R3) would rate all of the posttest essays.

Though such a design would meet the requirements for the MFRM, it would be inadequate for calculating an intraclass correlation coefficient. The latter would require a fully crossed design, including ratings from each rater for each student on each essay. Therefore, to provide an estimate of the intraclass correlation, R2 and R3 needed to rate additional essays. Appendix C illustrates how this was done. The principal researcher served as R1 and rated both pretest and posttest essays of all 47 students (94 essays) using the rhetorical competence rubric described in Chapter 2. The second rater, R2, rated each of the 47 pretest essays and was fully crossed with R1 and R3 on 46 essays, for a total of 70 individual essays. Similarly, R3 rated each of the 47 posttest essays and was fully crossed with R1 and R2 for the same 46 essays mentioned above, for a total of 70 essays. During this process, R2 and R3 were not informed of the rating design and were completely blind to student, group and essay.

With this rating design in mind, some additional discussion of the intraclass correlation is in order. McGraw and Wong (1996) point out that the ubiquitous Pearson r is an *interclass* correlation because it compares bivariate data sets that are not assumed to share the same metric or variance. However, they explain that an *intraclass* correlation

coefficient (ICC) can also be calculated for two or more sets of variables when the metric and variance are shared.

However, to calculate an ICC accurately, McGraw and Wong (1996) explain that researchers must choose the appropriate model for the specific context because different models utilize different calculations. Of the three possible models, the two-way mixed model was chosen because the raters, each of whom rated each essay from a random sample of students, were viewed as random effects as were the students and essays. In addition to the information provided by the MFRM, the ICC provided an average measure of consistency among all three raters in the form of a correlation coefficient. The results of the MFRM and the ICC are presented in the next chapter.

Instructional Methods

The 19 students included in the control group participated in Level 5 classes during the summer semester of 2006. Coursework included four 65-minute class periods per day, Monday through Thursday. These classes focused on reading, writing, listening, speaking and grammar. During the course of the semester, students in the control group received traditional process writing instruction and produced a total of four major papers. Each paper included multiple drafts where feedback focused on rhetorical conventions as well as linguistic accuracy.

In addition to working through several drafts of the major papers, classroom activities ranged from formal instruction on rhetorical conventions that might include organization, paragraph development, transitions, providing adequate examples and support, to in-class writing exercises and activities designed to help the students master particular skills. In addition, students would occasionally engage in peer-editing exercises

where they would evaluate the writing of a classmate, or they might spend time in the computer lab where they would practice writing a timed essay or participate in research exercises. Many of these efforts directly contributed to the writing portfolios, which were used to help determine each student's writing proficiency grade. Portfolios included two of their major papers, along with their drafts and a thirty-minute essay written in the computer lab at the end of the semester.

In a number of ways the learning experiences of those in the control group and treatment group were similar. For example, like the control group, the treatment group met for one semester and took four classes with 65-minute class periods, Monday through Thursday. Students in the treatment group also received formal instruction and participated in a variety of writing exercises and activities. Like the control group, students in the treatment group periodically went to the computer lab to practice writing thirty-minute essays and prepared a writing portfolio for the end of the semester.

Despite these similarities between the control group and the treatment group, there were also some important differences that were essentially the focus of this study. Perhaps the greatest single difference that set the treatment group apart from the control group was the daily writing and analysis of a ten-minute paragraph. This time limit was set intentionally with the general assumption that ten minutes was long enough to provide a representative sample of student writing over time while still being short enough that feedback could still be relatively immediate and manageable for both the teacher and the students.

These paragraphs were usually written during the first ten minutes of the class period and were written virtually every day the class met. Topics were diverse, ranging

from opinions, analysis on social issues, science, history, popular culture and so on. While students in the three classes that made up the treatment group wrote to the same topics most of the time, once each week topics varied from one class to another because they were chosen based on particular content students in different sections were reading and talking about in their other classes.

Each day the teachers read the paragraphs, marked each of the errors using a system of indirect coding and then returned the paragraphs to the students at the beginning of the following class period. Figure 3 illustrates the indirect coding symbols used by the Applied Grammar teachers. Though additional errors and suggestions were identified periodically, this list represents the most common error types that were emphasized in the classes. Appendix A illustrates how these codes were used in context.




D	= Determiner	S/PL	= Singular/Plural
SV	= Subject Verb Agreement	C/NC	= Count/Noncount
VF	= Verb Form	?	= Meaning is not clear
ro	= Run-on Sentence	AWK	= Awkward Wording
inc	= Incomplete sentence		= Word Order
VT	= Verb Tense	C	= Capitalization
PP	= Preposition	P	= Punctuation
SPG	= Spelling		= Omit
WF	= Word Form		= Something is missing
WC	= Word Choice	¶	= New Paragraph

Figure 3. Indirect coding symbols used to mark L2 student writing

After the teachers returned the marked paragraphs, the students corrected the marked errors and resubmitted a typed copy of their paragraph. Since the primary focus

of this course was linguistic accuracy, emphasis was placed on editing (correcting linguistic errors), rather than on revision (changing or enhancing the content of the paragraph). Students were usually given eight days from the time their hand-written draft was returned to submit an error-free version of their paragraph. For example, if a particular paragraph topic were assigned on Tuesday, the teacher marked the paragraphs and returned them to the students on Wednesday. Students would then use the teacher's feedback to edit their paragraph with the goal of resubmitting the paragraph without any linguistic errors before the Wednesday of the following week.

If errors were perpetuated in subsequent drafts of the paragraph, students would continue to rewrite the paragraph with additional feedback from the teacher as many times as was needed until all of the mistakes had been corrected or until the deadline had arrived. Invariably, such an approach resulted in students working on various drafts of different paragraphs at one time. The intent of having students produce error-free versions of their paragraphs was to provide the student with an opportunity to become more acquainted with the linguistic rules that were applied inaccurately and to provide the students with an accurate sample of writing that could be referenced in the future. The intent of imposing a deadline was to help motivate the student and to keep feedback manageable for both the teacher and the students. The goal of the teacher throughout this process was always to return edited drafts the next class period after they had been submitted.

In addition to editing and keeping track of all of these paragraphs, throughout this process the students kept a running total of the type and frequency of their errors on a tally sheet. The purpose of this sheet was to help the students to become well acquainted

with their most frequent error types. It was hoped that with this heightened awareness, students would become more familiar with how to overcome their greatest linguistic weaknesses so they could produce more accurate writing. This information also helped to shape the ongoing classroom instruction. A sample of this Error Tally Sheet is illustrated in Appendix D.

Along with the Error Tally Sheet, students maintained other records to track their progress. These included an Edit Log, which was used to track how many times students edited their writing before all of the errors had been corrected for each paragraph (see a sample in Appendix E) and an Error List, which was used to record every sentence or clause that contained some type of error (see a sample in Appendix F). Although the Error Lists would become quite lengthy over time for most students, these error samples provided students with insight into their progress in the class as well as provided them with a personalized reference document that could be used to help them review key principles needed to continue to improve the accuracy of their writing.

Though the treatment group received formal instruction like the control group, instruction for students in the treatment group was only loosely organized around a list of grammatical structures in the syllabus and was driven primarily by the specific needs of the students at any given time. Daily classroom instruction and activities often included analysis of student writing from the paragraphs written the day before. In other words, rather than following a predetermined syllabus, the syllabus for this course was dynamic in that it responded to students needs as demonstrated in the Error Tally Sheets and Error Lists. In this way, classroom instruction was flexible enough to focus on a particular

grammar point longer than might have been anticipated or to return to a grammar point after it had been taught previously.

Like the control group, students in the treatment group wrote four or five 30-minute essays during the course of the semester. There were three main purposes for these timed essays. First, these essays helped the students to apply what they were learning in the broader context of a larger, more complex piece of writing. Unlike the daily paragraphs, the expectations for these essays included a much greater level of rhetorical complexity such as an introduction with a clearly articulated thesis, a body with well-suited topic sentences and support, an appropriate conclusion, effective transitions between paragraphs and so on. Second, like the daily paragraphs, these essays were a rich source of error feedback for the students. Third, these essays provided the students with some experience with the 30-minute essay format that was used to elicit data at the conclusion of the course. The intent was to provide students with enough experience writing for 30 minutes so they could approach the final writing task with appropriate expectations and an accurate sense of the timing required to complete that task successfully.

Elicitation Procedures

As mentioned previously, the students in the control group took the pretest at the end of the winter semester in 2006 and students in the treatment group took the pretest at the end of winter semester 2007. The pretest task was simply to write for 30 minutes in response to the following prompt:

Do you agree or disagree with the following statement? Only people who earn a lot of money are successful. Use specific reasons and examples to support your answer.

These same students took the posttest at the end of the summer semester 2006 and at the end of the summer semester 2007 respectively. The posttest task was to write for 30 minutes in response to the following prompt:

In your opinion, what is the most important characteristic (for example, honesty, intelligence, a sense of humor) that a person can have to be successful in life? Use specific reasons and examples from your experience to explain your answer. When you write your answer, you are not limited to the examples listed in the question.

In both instances the elicitations occurred in a computer lab where students typed their responses during the regular final exam period in a secure testing environment. In-house computer software had been developed for delivery of the writing test under time conditions. Once students entered their identification numbers, the prompt appeared at the top of the screen along with a space to type the essay. Although the software allowed the students three common word processing options, including *cut*, *copy* and *paste*, no other word processing tools were available. While students worked, the remaining time for the task was displayed in the lower left of the screen. Once the time ran out, the software prevented the students from being able to continue to type and transitioned to additional portions of their exam that focused on other skills such as listening and speaking.

All of the students in the control group used the in-house software to take the pretest and the posttest, each of which was administered to the entire group at the same time. Though the instructional period was the same for the control group and treatment group, the time between the pretest and posttest was slightly shorter for half of the students in the treatment group. While 14 students in the treatment group took the pretest at the end of the 2007 winter semester, the other 14 took the pretest 18 days later at the beginning of the 2007 summer semester. The reason for this delay was because these students were either new to the ELC at the beginning of the 2007 summer semester or they had not yet been placed into Level 5 by the end of the 2007 winter semester. Though 26 of the 28 students in the treatment group took the posttest at the same time using the in-house system, two students who could not take the test at the planned time took the posttest a day before the group administration. These students were carefully proctored as they used AppleWorks software, a basic word processing tool with the same features that were available in the in-house system.

Essays were then saved and catalogued according to grouping and test administration times. Though the prompts remained with the essays so raters could evaluate the extent to which each writer completed the task successfully, no names or group information was included on material provided to those who scored or rated essays. From this point, essays were handled and analyzed using an identification code which could be traced back by the principal researcher to the corresponding group and test administration.

A brief comment about the identification codes may be useful. Codes were made up of four characters, a letter followed by three numbers. The letter indicated whether the

essay was from the control or treatment group. Identification codes beginning with vowels (A, E, I, O, U) represented the control group, and codes beginning with consonants represented the treatment group. These letters were randomly assigned to essays according to their respective grouping. The first of the three numbers indicated whether the essay was from the pretest or the posttest. Numbers 1-4 represented pretest essays, and numbers 6-9 represented posttest essays. Similarly, numbers were randomly assigned to essays depending on the test occasion. The final two numbers indicated the specific L2 writer. Figure 4 illustrates this coding scheme with two examples.

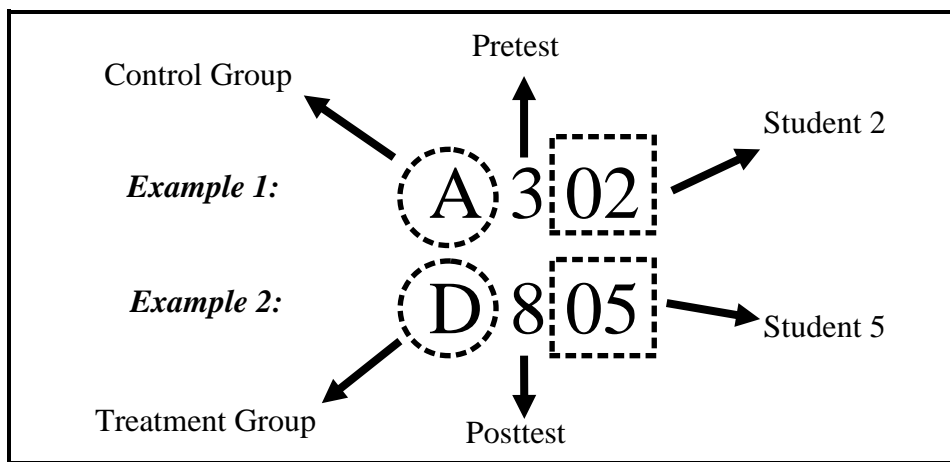


Figure 4. Illustration of components of identification coding

Research Questions Operationalized

With this additional background, we are now ready to restate the research questions operationally:

1. To what extent will the treatment produce greater linguistic accuracy in new writing when compared to the traditional instructional method?

Operationally: Will mean accuracy scores from posttest essays be significantly greater for the treatment group?

2. To what extent will the treatment produce equivalent levels of rhetorical competence, fluency and complexity on a new piece of writing when compared to the traditional approach?

Operationally:

- a. *Rhetorical competence:* will mean rhetorical competence scores from posttest 30-minute essays be significantly lower for the treatment group?
 - b. *Fluency:* will the total number of words written from pretest and posttest 30-minute essays be significantly fewer for the treatment group?
 - c. *Complexity:* will the average number of words per T-unit written from pretest and posttest 30-minute essays be significantly fewer for the treatment group?
3. What is the relationship between explicit grammar knowledge and grammar use in a productive writing task?

Operationally: What proportion of the variance in grammar use on the 30-minute essay can be explained by grammar knowledge as demonstrated by the Level 5 grammar test?

It should be noted that the research questions were divided into Phase I and Phase II questions and that the first three research questions represented the Phase I questions. The *a priori* decision was to simultaneously run the five separate tests included in

Questions 1-3, using the Bonferroni correction to safeguard against the chance possibility that a particular result of the treatment might appear significant when actually it was not. With an original significance level of .05, divided by the five tests, the resulting significance level was .01. However, it was also decided that if the results for Question 1 were significant (i.e. the treatment produced significantly greater accuracy scores), then analysis would continue on to Research Question 4, consisting of the Phase II questions. Research Question 4 is operationally defined below:

4. Which, if any, of the following accuracy scores from pretest and posttest essays will be significantly greater for the treatment group? These include (a) sentence structure accuracy scores, (b) determiner accuracy scores, (c) verb form accuracy scores, (d), numeric accuracy scores, (e) semantic accuracy scores, (f) lexical accuracy scores, and (g) mechanical accuracy scores.

Notwithstanding these seven additional tests, it was decided *a priori* that rather than continuing to fragment the significance level into increasingly smaller values, the prior significance level of .01 would be retained as a rough pseudo Bonferroni correction (as described by Huck, 2008). This is because efforts to avoid both error types were deemed as having equal value in this study. While the need for protection against type I errors grows with each additional test, the risk of increasing type II errors also grows proportionally with increasingly more stringent significance values. Therefore, rather than function as a rigid cutoff point, it was intended that this significance level would function as a rough approximation and that if tests were found that would have been

significant prior to the Bonferroni correction, they should be carefully analyzed for evidence of practical significance.

CHAPTER 4: RESULTS

The purpose of this chapter is threefold. First, the chapter provides the results of the various methods used to estimate the reliability of the measures analyzed in this study. This includes reporting the Pearson correlation coefficients for various counts of error types by the two scorers as well as the intraclass correlation coefficient and the results of the Rasch Modeling for the three raters as was described in Chapter 3. Second, the chapter presents the several repeated measures ANOVA results needed to answer the research questions. Finally, the chapter provides a brief rationale for two additional repeated measures tests, which, *a posteriori* to the data analysis, were developed to help provide additional insight for answering the research questions.

Reliability Estimates

Before presenting results from the statistical tests chosen to help answer our research questions, we need to examine the reliability of the measures used in this study. The procedures designed to provide evidence for reliability were followed as outlined in Chapter 3. Two scorers (S1 and S2) independently counted the total number of T-units for each essay. Where discrepancies occurred, specific essays were reviewed, and the scorers decided the total number of T-units jointly. S1 then scored all 94 essays on eight categories of accuracy, and the resulting accuracy scores were derived from the proportion of accurate T-units for a given accuracy type over the total number of T-units as outlined previously in Table 3.

Scoring reliability. Though it was not possible for all of the essays to be double scored due to practical constraints, S2 scored a stratified random sample of 48 essays as illustrated previously in Table 5. Pearson correlation coefficients by accuracy type were

generated for each score set. These are listed in Table 9 and show a range of coefficients from .81 to .98. Though the relative strength of these correlation coefficients varied from one accuracy type to another, it was assumed that they provided sufficient evidence of reliability to justify the use of the scores for the subsequent repeated measures tests.

Table 9

Pearson Correlation Coefficients by Accuracy Type between S1 and S2

	Types of Accuracy	<i>r</i>
1.	Mechanical accuracy scores:	.98
2.	Overall accuracy scores:	.97
3.	Determiner accuracy scores:	.94
4.	Semantic accuracy scores:	.92
5.	Verb accuracy scores:	.90
6.	Sentence Structure accuracy scores:	.86
7.	Numeric accuracy scores:	.83
8.	Lexical accuracy scores:	.81

Rating reliability. In addition to examining the reliability of the accuracy scores, we also need evidence of the reliability of ratings used to determine rhetorical competence. As described in Chapter 3, two methods were used: (a) an intraclass correlation coefficient (ICC), and (b) the Many-facets Rasch Model (MFRM). Three raters (R1, R2 and R3) used the Rhetorical Competence Rubric included in Appendix B. Though R1 rated all 94 essays, R2 and R3 rated 71 essays each—48 of which were triple rated and 23 of which were double rated as illustrated in the rating design included in Appendix C.

Since intraclass correlations require a fully crossed design, the ICC could only be calculated for the 48 essays that were triple rated for rhetorical competence. SPSS was used for this calculation and generated an average measures ICC for the three raters of .87 ($df1 = 47, df2 = 94, p < .001$). This statistic is calculated as the ratio of the covariance from the ratings compared with the total variance. While this correlation coefficient provided some positive evidence of the reliability of the rhetorical competence ratings, the MFRM was also used to provide additional complementary information beyond what the ICC could provide on its own.

FACETS output. Before focusing on the reliability of the rhetorical competence ratings, however, it may be helpful to discuss some of the relevant information that can be generated through Rasch Modeling. In addition to providing an analysis of the reliability of facet data and the potential to strengthen reliability by utilizing adjustments recommended by the model, the MFRM also allows researchers to identify whether data generally functions as expected. Therefore, this section will present a number of figures and tables that will help clarify whether data functioned as predicted as well as the reliability of that data. First, we will examine FACET output data related to a student essays. Second, we will analyze rater performance, and third, we will examine the function of the rhetorical competence rubric itself.

First, we will examine Figure 5, which displays the vertical logit scale on the far left, followed by student essays, raters and the categories of the rhetorical competence rubric all plotted on the same scale. For student essays, the higher toward the top of the vertical scale, the greater the rhetorical competence demonstrated by the essay. At this point it is important to note that in order to facilitate sorting and data analysis for Rasch

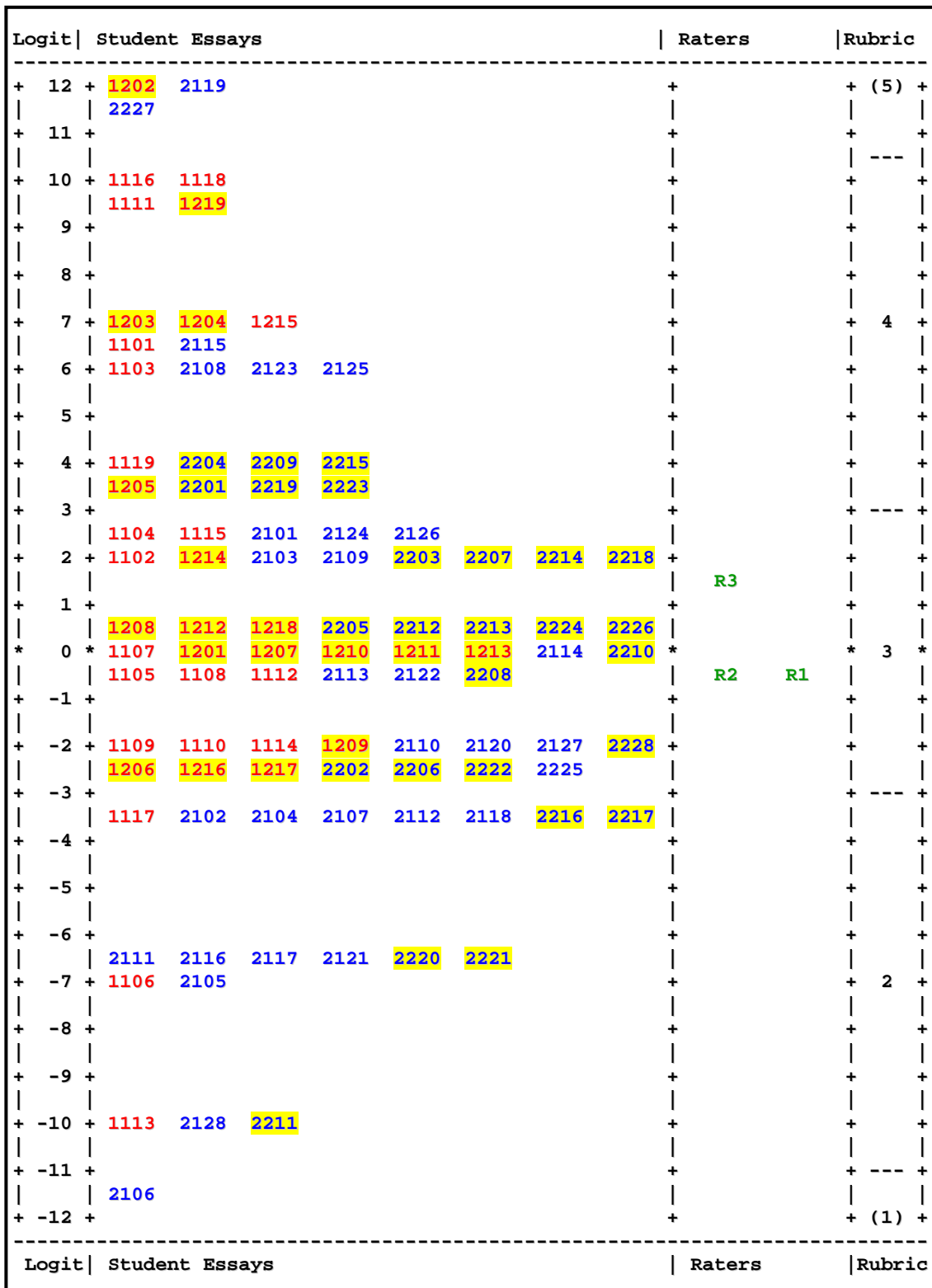


Figure 5. Vertical plot of student essays, raters, and rubric levels in logits.

Modeling, student essays were provided with new identification codes as they were input into FACETS. The first number of the four-digit codes that appear in Figure 5 represents which experimental group each essay came from. For example, essays beginning in “1” indicate the control group and those beginning in “2” indicate the treatment group. The second number represents the essay, where “1” stands for the pretest, and “2” stands for the posttest. The last two digits signify the individual student number.

In an effort to accentuate these coding differences in the figure, the control group codes appear in red, and treatment group codes appear in blue. In addition, posttest essay codes for both groups have been highlighted with a yellow background. The third and fourth columns in the figure place the three raters (R1, R2 and R3) and the rubric levels on the same logit scale as the essays. Although an inspection of Figure 5 shows no obvious patterns in terms of experimental groups or test occasions, ratings appear to approach a normal distribution about the mean.

In addition, Table 10 provides a summary of FACETS output for student essays which includes the means and standard deviations for ratings, ability, standard error and infit statistics. The table also reports a reliability separation index of .86 and the separation of 2.43, which suggest that individual essays are fairly reliably separated from each other in terms of the levels of rhetorical competence demonstrated by each. J. Linacre provided a general benchmark of at least .80 for a reliability separation index and at least 2.0 for separation (J. Linacre, unpublished training material, 2008). The reliability separation index shows how reliably different student scores are from each other. Thus, the higher the value of the index is, the greater the discrimination. On the other hand, given the theoretical notion of a *true distribution* for a set of data, the separation indicates

how many separate measures can be reasonably differentiated based on the number of error distributions or error strata that appropriately fit within the true distribution.

Table 10

Summary of FACETS Output for Student Essays

Students	Observed Mean	Ability Measure	Standard Error	Infit MS
Mean	3.1	0.38	1.79	0.63
Stand Dev.	0.8	4.96	.57	0.84

Separation = 2.43, Reliability separation index = .86, Chi-square = 849.8, $df = 93$, $p < .00$

With this information related to student essays in mind, we will now examine rater performance. As demonstrated in Figure 5, though all three raters are generally clustered around the mean, R3's ratings appear more severe than the ratings of R1 and R2. The information displayed in Table 11 provides a more precise analysis of this observation. The table column labeled *observed mean* presents the average rating awarded by each rater, and the *rater severity* column shows the relative severity of respective raters measured in logits. For example, R3 is identified as the most severe (1.36) and R1 as the most lenient (- 0.72), for a complete range of 2.08 logits. Since ideally there would be no differences among raters, this range of more than 2 logits tends to undermine the reliability of these ratings.

This problem is further illustrated by a reliability separation index of .91 and a separation of 3.14. While a higher index would be appropriate and desirable for examinees, in the case of raters, this represents undesirable variance in severity or leniency. Therefore, these statistics suggests that the raters in this study were fairly

Table 11

Summary of FACETS Output for Raters 1, 2 and 3

Raters	Observed Mean	Rater Severity	Standard Error	Infit MS
Rater 1	3.1	– .72	.26	0.51
Rater 2	3.1	– .64	.30	1.16
Rater 3	2.8	1.36	.31	0.84
Mean	3.0	0.00	.29	0.84
Stand Dev.	0.2	0.96	.02	0.27

Separation = 3.14, Reliability separation index = 0.91, Chi-square = 30.8, $df = 2$, $p = < .01$

reliably inconsistent. Although this reliability separation index is high and certainly is not ideal, values as high or higher are not uncommon for studies using multiple raters (for examples, see Bachman, Lynch & Mason, 1995; Haladyna & Hess, 1994; McCollum, 2006; Park, 2004). Fortunately, raters were consistent enough that the MFRM was able to produce a “fair average” for each essay rating that adjusts for differences in severity from one rater to the next. These adjusted ratings were used for subsequent analyses because they are more reliable and provide an estimate that is much more precise than would be obtained simply by averaging the three ratings.

Additional information in Table 11 that should be highlighted is the mean square infit statistic. According to Wright and Linacre (1994), these results would not be appropriate for high-stakes testing because R1’s infit statistic of 0.51 falls outside the desired range of 0.8 to 1.2. However, the 0.51 does fall within the 0.50 to 1.7 range they have established for clinical observation. Moreover, 0.51 also falls well within the acceptable range set by other researchers such as Pollitt and Hutchinson (1987), McNamara (1989), Park (2004) and Kim (2006). They have suggested that the infit statistic should not be less than or greater than the mean square mean plus or minus twice

the standard deviation. Data from Table 7 show that the mean square infit statistics would need to be within the range of .30 to 1.38 [$.84 \pm 2(.27)$]. Since the infit statistic for each of the three raters falls within this range, the model appears to account for enough observed variation to allow us to conclude that our adjusted “fair average” ratings would be sufficiently reliable to be used in our repeated measures test.

Additional sources of useful information from the FACETS output are summarized in Table 12 and Figure 6, and help us understand how well the Rhetorical Competence Rubric functioned. Table 12 displays the rubric categories in the left column, followed by the step calibration values, the counts for each category selection, and the accompany breakdown of category use in percentages. The step calibration values correspond to the logit scale and mark the intersections between two probability curves where the probability of a rater awarding one rating is equal to the probability of the same rater award the adjacent rating. For example, the intersection of Categories 1 and 2 is marked by -10.83 and the intersection of Categories 4 and 5 is marked by 10.59. Since step calibration values represent the intersections of two probability curves, there will be one fewer intersection than rubric categories. However, perhaps the most important characteristic of the step calibration values is that they are properly ordered as is demonstrated in Table 12.

Much of this same information is graphically depicted in Figure 6, which illustrates the probability curves for the rhetorical competence ratings. Rhetorical competence is plotted along the horizontal axis with those essays demonstrating the least competence on the left and those demonstrating the most competence on the right. Probability, ranging from 0 to 1, is plotted long the vertical axis. The figure displays one

Table 12

Summary of FACETS Output for the Rhetorical Competence Rubric

Rubric Categories	Step Calibrations	Counts Used	Percentage Used
1	--	4	2 %
2	-10.83	59	26 %
3	-2.87	108	47 %
4	3.11	54	23 %
5	10.59	6	3 %

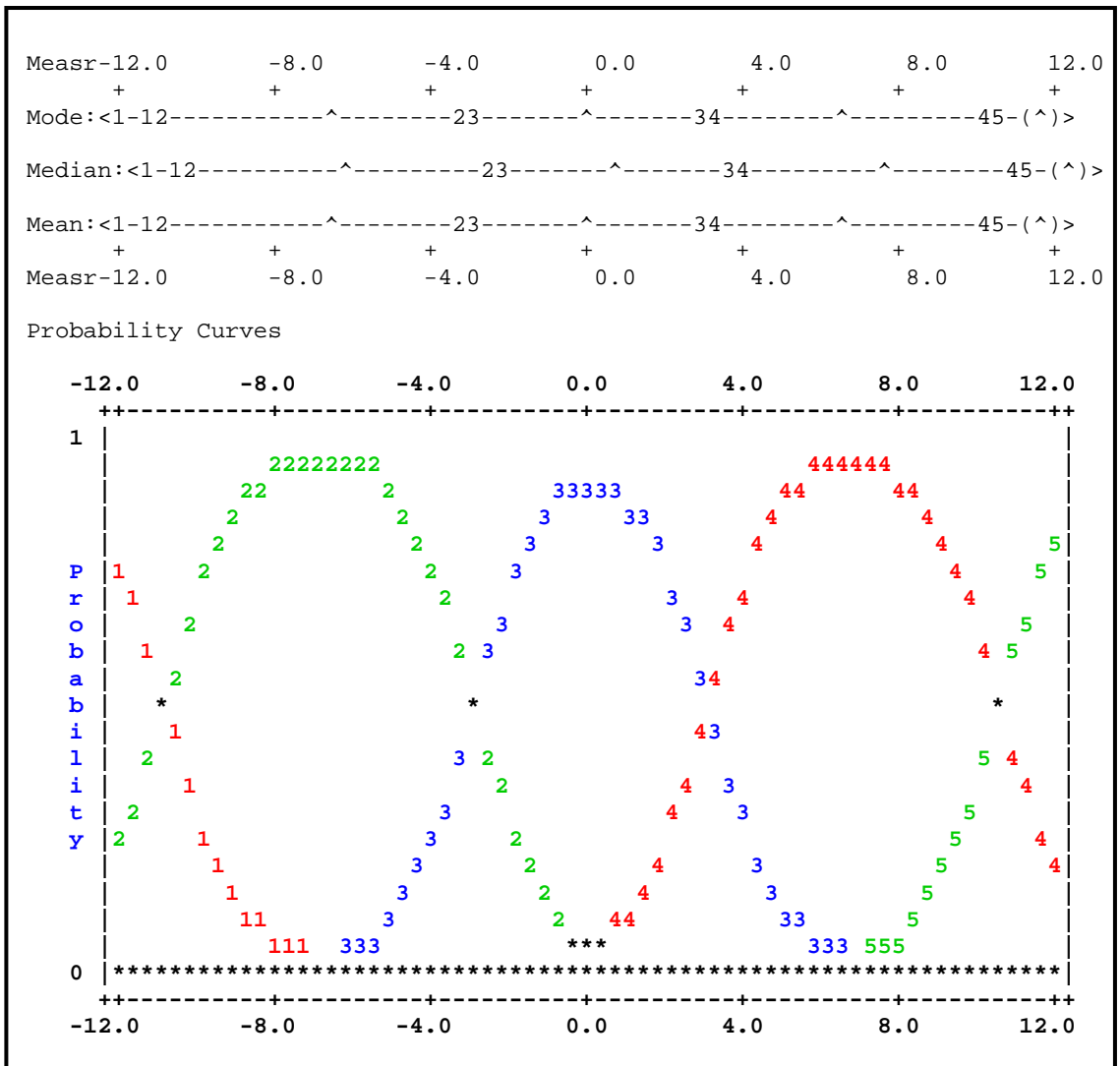


Figure 6. Probability curves for rhetorical competence ratings

curve for each level of the rubric used by the raters, and the five numbers constituting the actual curves represent the five levels included in the rubric.

Ideally each level of the rubric would be represented by a distinct peak, and each peak would be evenly spaced horizontally. Such conditions would show that a particular category awarded by raters would be the most probable for a given portion of the rhetorical competence distribution. If different category curves ended up being superimposed, were stacked vertically, or were not evenly spaced, such problems would provide evidence that the rubric categories are not functioning as expected and would suggest that the rubric may need to be revised. Problems such as these would undermine our ability to see clear probabilities for essays to be assigned a specific level in a given portion of the rhetorical competence distribution. However, inspection of the probability curves in Figure 6 show data that appears nearly idealistic, suggesting that the rubric functioned as expected.

Effect Size

In addition to examining the reliability of our measures, we also need to discuss how this study addresses the issue of effect size. Over the past few decades, researchers and practitioners have seen growing criticism of the limitations of research methods that simply rely on significance testing. Many have advocated methods that emphasize identifying the effect size of independent variables in order to place tests of significance into a more meaningful context. For example, some have noted that the results of some research may be statistically significant while practical significance is negligible.

Conversely, the results of some research may not be significant though there may be a

great deal of practical significance (see Cortina & Hossein, 2000, Grissom & Kim, 2005; Kline, 2004).

Moreover, the fifth edition of the Publication Manual of the American Psychological Association (2001) recommends that researchers report estimates of effect size, even when results are not significant. Similarly, many professional research journals now require their authors to provide some indication of the magnitude of the effects reported in their articles. Despite this emphasis, a number of researchers have struggled to understand which measures of effect size might be the most suitable in various research contexts. This is an appropriate question because, as Grissom and Kim (2005) stress, “no effect size or estimator is without one or more limitations” (p. 124).

Tabachnick and Fidell (1996) explain this notion of effect size or strength of association as

. . . the proportion of variance in the DV [dependent variable] associated with levels of an IV [independent variable] . . . Statistical significance testing assesses the *reliability* of the association between the IV and the DV. Strength of association measures *how much* association there is. (p. 53)

Although there are numerous measures of effect size that might be considered, and research and debate about the appropriateness of different methods in various contexts is ongoing, this study utilizes the partial eta squared statistic (η_p^2) along with the eta squared (η^2) statistic and the simple main effects for those interactions that are significant. Though no method of effect analysis will be ideal for every context, the rationale for these three approaches is that they seem the best suited for the specific

context of this study. The intent is to provide the reader with adequate information to draw appropriate conclusions about the various phenomena under investigation.

First, it should be noted that η^2 and η_p^2 are not the same. Tabachnick and Fidell (1996) express η^2 as

$$\eta^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}}$$

However, they point out that, as the proportion of the total variance attributed to a particular effect, the η^2 is flawed in that the strength of association depends on how many independent variables are included in the design and how significant those variables are. Thus, the reliability of the η^2 statistic as an estimate of effect size seems somewhat context dependent.

They go on to explain that η_p^2 is an attempt to correct for this defect and is expressed as

$$\eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error}}}$$

Though Bakeman and Robinson (2005) refer to the η_p^2 as “more useful” (p. 239), it is important for researchers to understand that what the η_p^2 calculates is quite different from the η^2 and that the η_p^2 also has its own limitations that need to be understood. For example, Bakeman and Robinson point out that the η_p^2 would not be recommended for “comparing effects of a particular variable across studies that use different designs” (p. 239). In addition, Tabachnick and Fidell (1996) clarify that the η_p^2 should not be used to

draw inferences about a larger population, and Pedhazur (1997) claims that η_p^2 “is an overestimate of the actual effect size” (p. 509).

Despite these limitations, however, Bakeman and Robinson (2005) explain that unlike the η^2 , the η_p^2 is rather successful at isolating the effect of a specific variable. For this reason they recommend its use, particularly “in the context of repeated-measures designs” (p. 239). Fortunately, Cohen (1988) has provided useful guidelines for interpreting the η^2 and η_p^2 statistics. Cohen proposed that .01 represented a small effect, that .06 represented a moderate effect, and that .14 represented a large effect (also see Huck, 2008). Nevertheless, since the η_p^2 will often produce a larger value than the η^2 , many researchers have warned of the need for great care in clarifying which statistic is used (see Bakeman & Robinson, 2005, Pierce, Block & Anguis, 2004), noting that some researchers and journals have reported η_p^2 statistics that were mistakenly referred to as η^2 . Therefore, since the η^2 and η_p^2 both have strengths and weaknesses in the context of this study, the η^2 will be reported along with the η_p^2 when results are significant.

In addition to estimating the effect size, Shaughnessy, Zechmeister and Zechmeister (2003) recommended that a test of simple main effects can be used when an interaction in a mixed model, repeated measures ANOVA is statistically significant. They point out that, “A simple main effect is the effect of an independent variable at only one level of a second independent variable” and that calculating simple main effects is helpful for identifying “the source of an interaction” (p. 441). Therefore, simple main effects were also calculated for those ANOVA tests that included a significant interaction.

ANOVA Test Results

Having addressed how this study will deal with issues of effect size, we are now ready to examine the results of the repeated measures ANOVA tests designed to help answer our research questions. It should be kept in mind that though Question 1 was deemed as the most important, data analyzed to answer Question 1 were tested simultaneously with data for the three parts of Questions 2 as well as data for Question 3. As explained in the previous chapter, a pseudo Bonferroni correction was used for these five tests, resulting in an adjusted significance level of .01.

The first question stated “To what extent will the treatment produce greater linguistic accuracy in new writing when compared to the traditional instructional method?” This was operationally defined as: “Will accuracy scores from pretest and posttest essays be significantly greater for the treatment group?” As described previously, these accuracy scores were derived from the total number of error-free T-units over the total number of T-units in each essay. Table 13 provides the means and standard deviations for accuracy scores for the control and treatment groups. The ANOVA summary in Table 14 demonstrates an interaction effect showing that significantly higher accuracy scores were produced by those who received the treatment than those who had been instructed with the traditional approach. Figure 7 plots this interaction effect and Table 15 summarizes the simple main effects of the interaction.

Though Table 14 shows a significant main effect ($p = .04$) for the “time” factor, this must be qualified by the significant interaction effect ($p = .001$) illustrated in Figure 7 and Table 14. Together these show that while pretest group differences were not

Table 13

Descriptive Statistics for Accuracy Scores

Group		Pretest	Posttest	Means
Control (n = 19)	Mean	16.30	13.78	15.04
	SD	10.70	11.81	11.26
Treatment (n = 28)	Mean	14.02	24.16	19.09
	SD	15.00	19.46	17.23
Total (N= 47)	Mean	14.94	19.97	17.46
	SD	13.35	17.42	15.39

Table 14

Mixed ANOVA Summary Table for Accuracy Scores

Source	SS	df	MS	F	p	η_p^2
Between Subjects		46				
Group	371.05	1	371.05	0.95	.33	.02
Error	17536.12	45	389.69			
Within Subject		47				
Time	329.01	1	329.01	4.44	.04	.09
Time x Group	908.19	1	908.19	12.26	.001	.21
Error	3333.22	45	74.07			
Total	22477.59	93				

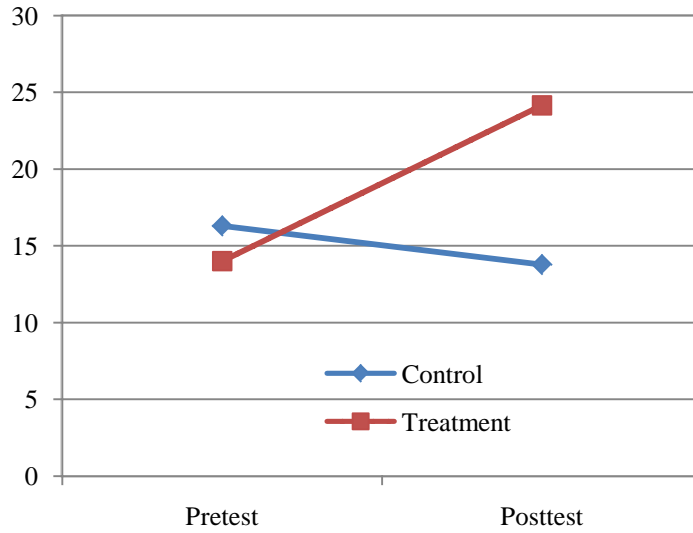


Figure 7. Pretest and posttest means for accuracy scores

Table 15

Simple Main Effects for Pretest and Posttest Accuracy Scores

Source	SS	df	MS	F	<i>p</i>
Between Groups at Pretest	2.66	1	2.66	2.30	0.14
Between Groups at Posttest	55.87	1	55.87	48.30	0.00000001
Error	53.21	46	1.16		

significant ($p = .14$), posttest differences between experimental groups were significant ($p < .000$), suggesting that the treatment had a positive effect on writing accuracy.

In addition, we should consider the effect size of this interaction. While the η^2 of .04 (derived from the $SS_{\text{effect}}/SS_{\text{total}}$ included in Table 14) suggests a small to moderate effect size, the η_p^2 of .21 suggests a large effect size. Though we should keep in mind that these eta statistics actually measure different things, together, they seem to provide enough evidence to suggest that the treatment had a practical effect on writing accuracy as measured by EFTs in the pretest and posttest essays.

The three parts of Question 2 are articulated in the following: “To what extent will the treatment produce equivalent levels of rhetorical competence, fluency and complexity on a new piece of writing when compared to the traditional approach? Operationally, these included (a) *Rhetorical competence*: “Will rhetorical competence scores from posttest 30-minute essays be significantly lower for the treatment group?” (b) *Fluency*: “Will the total number of words written from posttest 30-minute essays be significantly fewer for the treatment group?” (c) *Complexity*: “Will the average number of words per T-unit written from posttest 30-minute essays be significantly fewer for the treatment group?”

As explained previously, rhetorical competence scores were derived from ratings based on the rubric in Appendix B. Table 16 presents the descriptive statistics for rhetorical competence ratings for writers in the control and treatment groups. The ANOVA summary in Table 17 shows that differences in the rhetorical competence

Table 16

Descriptive Statistics for Rhetorical Competence Ratings

Group		Pretest	Posttest	Means
Control (n = 19)	Mean	3.18	3.30	3.24
	SD	0.76	0.65	0.71
Treatment (n = 28)	Mean	2.82	3.00	2.91
	SD	0.81	0.63	0.72
Total (N= 47)	Mean	2.97	3.12	3.05
	SD	0.81	0.65	0.73

Table 17

Mixed ANOVA Summary Table for Rhetorical Competence Ratings

<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η_p^2
Between Subjects		46				
Group	2.44	1	2.44	3.47	.07	.07
Error	31.60	45	0.70			
Within Subject		47				
Time	0.51	1	0.51	1.51	.23	.03
Time x Group	0.03	1	0.03	0.09	.77	.002
Error	15.22	45	0.34			
Total	49.81	93				

ratings generated by the two groups were not significantly different and that effect sizes were nearly negligible.

Similarly, Table 18 provides the descriptive statistics for fluency scores generated by learners in the control and treatment groups. Table 19 suggests that while writing fluency was not significantly different from one group to the next ($p = .19$), both groups appear to have significantly improved their writing fluency during the instructional period ($p = .01$). Moreover, the significance of this result is underscored by an effect size estimate that is *small* ($\eta^2 = .03$) or *moderate to large* ($\eta_p^2 = .13$), regardless of the instructional method.

Table 18
Descriptive Statistics for Writing Fluency Scores

Group		Pretest	Posttest	Means
Control (n = 19)	Mean	359.53	409.11	384.32
	SD	73.03	95.51	84.27
Treatment (n = 28)	Mean	357.36	372.75	365.06
	SD	89.08	117.19	103.14
Total (N = 47)	Mean	358.23	387.45	372.84
	SD	82.14	109.34	95.74

Table 19

Mixed ANOVA Summary Table for Writing Fluency Scores

<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η_p^2
Between Subjects		46				
Group	8399.56	1	8399.56	.56	.46	.01
Error	679499.55	45	15099.99			
Within Subject		47				
Time	23890.96	1	23890.96	6.49	.01	.13
Time x Group	6614.28	1	6614.28	1.80	.19	.04
Error	165748.66	45	3683.30			
Total	884153.01	93				

Although the interaction effect of time by group ($p = .19$) was not statistically significant at the .05 level in Table 19, the effect size, though small ($\eta_p^2 = .04$) warranted additional exploration. This interaction is plotted in Figure 8, and Table 20 presents the simple main effects for the pretest and posttest fluency scores. Table 20 suggests that while mean fluency scores from the two experimental groups were not significantly different at the pretest occasion ($p = .69$), the control group demonstrated significantly higher fluency scores at the posttest occasion ($p < .000$).

While an additional test of simple main effects for each experimental group between pretest and posttest occasions showed significant increases in fluency scores for both the control group ($p < .000$) and the treatment group ($p = .03$), these data suggest that the students in the control group increased their fluency significantly more than the students in the treatment group. Thus, the treatment appears to have favored the control group with a small but practical advantage over the treatment group in terms of the development of L2 writing fluency.

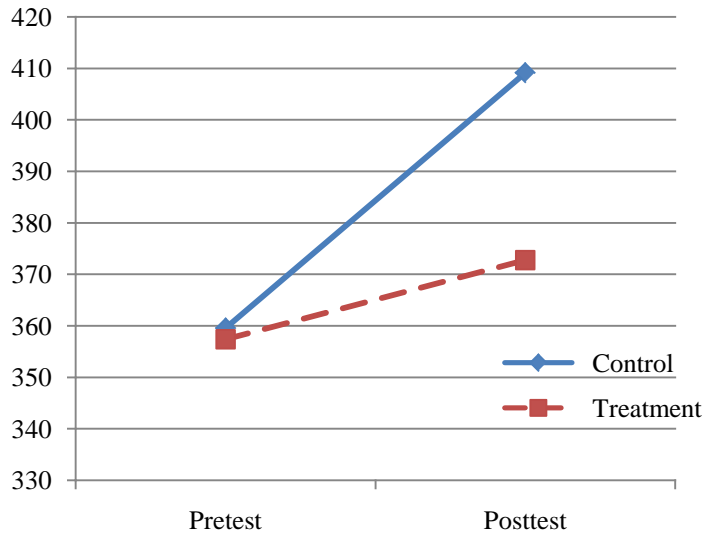


Figure 8. Pretest and posttest means for fluency scores

Table 20

Simple Main Effects for Pretest and Posttest Fluency Scores

Source	SS	df	MS	F	<i>p</i>
Between Groups at Pretest	2.45	1	2.45	0.16	0.69
Between Groups at Posttest	685.25	1	685.25	46.16	0.00000002
Error	682.80	46	14.84		

The third part of Question 2 addressed the issue of writing complexity.

Complexity was defined as mean length of T-units divided by the total number of T-units for a given essay. Table 21 presents descriptive statistics on complexity scores for the control and treatment groups, and Table 22 provides the ANOVA summary. Though the interaction effect of “time by group” was not significant ($p = .079$) at a .05 alpha, it is interesting to note its estimated effect size of *small* ($\eta^2 = .02$) to *moderate* ($\eta_p^2 = .067$). The pretest and posttest means for complexity scores are plotted in Figure 9.

Table 21

Descriptive Statistics for Writing Complexity Scores

Group		Pretest	Posttest	Means
Control (n = 19)	Mean	12.56	14.13	13.35
	SD	2.68	3.90	3.29
Treatment (n = 28)	Mean	13.69	13.55	13.62
	SD	2.51	2.50	2.51
Total (N = 47)	Mean	13.23	13.78	13.51
	SD	2.61	3.12	2.87

Table 22

Mixed ANOVA Summary Table for Writing Complexity Scores

Source	SS	df	MS	F	p	η_p^2
Between Subjects		46				
Group	1.74	1	1.738	.152	.698	.003
Error	513.37	45	11.408			
Within Subject		47				
Time	11.62	1	11.619	2.261	.140	.048
Time x Group	16.62	1	16.617	3.234	.079	.067
Error	231.22	45	5.138			
Total	774.56	93				

Though the effect of the experiment on writing complexity may seem small, Figure 9 and Tables 23 offer additional information that may provide further insight into the possible effects of the instructional method. While Figure 9 plots the interaction effect, Table 23 clarifies that between group differences were significant on both occasions such that students in the treatment wrote with significantly greater complexity on the pretest and students in the control group wrote with significantly greater complexity on the posttest.

While an additional test of simple main effects for each experimental group between pretest and posttest occasions showed a significant increase in complexity scores for the control group ($p < .000$), differences in the complexity scores for the treatment group were not significant ($p = .55$). This suggests that while the complexity of student writing in the control group increased over the course of the experimental period, the complexity of student writing in the treatment group seems to have been unaffected by the treatment.

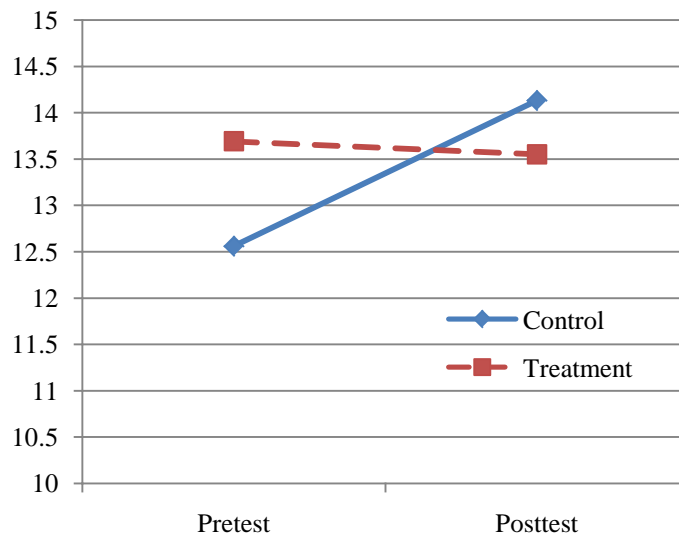


Figure 9. Pretest and posttest means for complexity scores

Table 23

Simple Main Effects for Pretest and Posttest Complexity Scores

Source	SS	df	MS	F	p
Between Groups at Pretest	0.66	1	0.66	62.63	.0000000004
Between Groups at Posttest	0.18	1	0.18	16.63	.0002
Error	0.49	46	0.01		

Although it is unclear why L2 writing from students in the control group may have produced greater fluency and greater complexity when compared to the writing from the treatment group, there are at least two possibilities. First, this could be the result of some inherent group differences that were not controlled in the design of this study. Second, it is equally possible that these effects could result from the treatment itself. For example, it is conceivable that as students strive to write more accurately, the ongoing development of fluency and complexity of their writing may be inhibited.

However, two important points should be kept in mind regarding these findings. First, the effect of the treatment on accuracy scores, which favored the treatment group, was large while the effects of the treatment on fluency and complexity, which favored the control group, were much smaller. Second, these findings are not suggesting that the students in the treatment group decreased in their writing fluency or complexity, only that they did not increase in their fluency or complexity at the same rate as the students in the control group. With these insights in mind, we are not ready to move on to the next question.

Question 3 was the last of the Phase I research questions. It stated “What is the relationship between explicit grammar knowledge and grammar use in a productive

writing task?” This was operationally defined as: “What proportion of the variance in the accuracy of grammar use on the 30-minute essay can be explained by grammar knowledge as demonstrated by the Level 5 grammar test?” To answer this question a simple bivariate regression analysis was conducted with the grammar knowledge scores used as an explanatory variable for the accuracy scores. A summary of this regression analysis can be seen in Table 24.

Table 24

Summary of Bivariate Regression Analysis

Source	Unstandardized Coefficients		Standardized Coefficients	<i>t</i>	<i>p</i>
	B	Std. Error	Beta		
(Constant)	-27.087	14.494		-1.869	.069
Grammar Knowledge	.627	.188	.462	3.338	.002

The results of this regression analysis suggest that the grammar knowledge scores were a significant ($p = .002$) predictor of linguistic accuracy on the writing task and that grammar knowledge accounted for approximately 20% ($r^2 = .214$) of the variance in the linguistic accuracy scores as demonstrated by the EFT/T ratio in the 30-minute essays. Figure 10 plots these data along with the corresponding regression line (plotted as the black line). However, an inspection of the figure reveals a possible outlier (marked with a red circle) that may slightly distort these results.

Though there was no clear evidence to suggest that the validity of this student’s performance should be questioned other than the student’s isolated location on the plot, a careful review of this individual’s scores in the classroom and on other measures from this study showed a pattern of below average performance and occasional performance

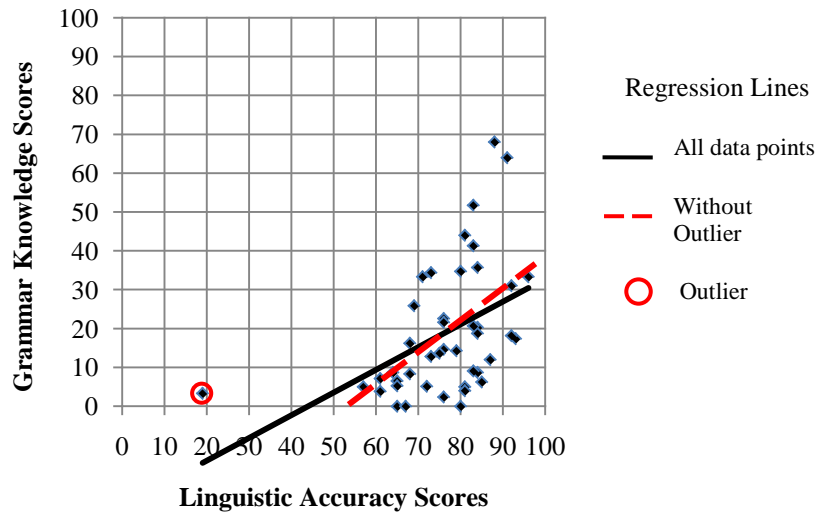


Figure 10. L2 writer performance plotted by grammar knowledge and linguistic accuracy

below the first quartile of those examined. Notwithstanding this observation, the student's location on the plot seemed to differ considerably from what might be expected. Therefore, an additional line was calculated without this possible outlier (plotted as the dotted red line). This second regression line suggests that grammar knowledge accounted for nearly 25% ($r^2 = .244$) of the variance in the linguistic accuracy scores.

While these data from the grammar knowledge scores and the linguistic accuracy scores have an obvious relationship, they also suggest that nearly 75-80% of the observed variance is unrelated to grammar knowledge as demonstrated by the grammar knowledge test. Though this may seem like a great deal of unexplained variance, Bakeman and Robinson (2005) reminded us that the r^2 and η^2 statistics measure essentially the same thing—the proportion of the total variance attributed to a particular effect. With this in mind, these data show a much stronger relationship than was expected and suggest that in this study grammar knowledge had a fairly positive effect on writing accuracy. While grammar knowledge by itself may be insufficient to produce highly accurate writing,

these findings suggest that a solid knowledge of grammar is likely to be an important asset for those who desire to write accurately.

In addition to the three Phase I questions we have examined, there were also a number of Phase II questions, examined in this study. As explained previously, since the ANOVA tests associated with Research Question 1 showed that students who received the treatment generated significantly higher accuracy scores than those who did not, Phase II data to answer Research Question 4 were also analyzed in an attempt to provide additional insight about the effects of the treatment on writing accuracy.

Though this involved an additional seven tests, the *a priori* decision was to retain the .01 significance level used previously as a broadly interpreted pseudo-Bonferroni correction. The rationale for this decision was a thoughtful attempt to balance efforts to safeguard against both type I and type II errors. Moreover, it was decided that rather than function as a rigid cutoff point, this significance level would work as a general value to guide our analysis. For example, it was decided that if tests were found that would have been significant prior to the Bonferroni correction, they would also be analyzed for evidence of practical significance. It was also determined that regardless of significance levels, test results would be carefully examined whenever warranted by effect size.

Research Question 4 was operationally defined as: “Which, if any, of the following accuracy scores from posttest essays will be significantly greater for the treatment group? These include (a) sentence structure accuracy scores, (b) determiner accuracy scores, (c) verb accuracy scores, (d), numeric accuracy scores, (e) semantic accuracy scores, (f) lexical accuracy scores, and (g) mechanical accuracy scores.” Each of these will be examined, beginning with the first Phase II sub-question regarding

sentence structure accuracy scores. Table 25 provides the descriptive statistics for the sentence structure accuracy scores for the control and treatment groups. In addition, Table 26 shows that differences in these sentence structure accuracy scores were not significant and that effect sizes were quite small.

Table 25

Descriptive Statistics for Sentence Structure Accuracy Scores

Group		Pretest	Posttest	Means
Control (n = 19)	Mean	96.07	96.47	96.27
	SD	4.58	3.52	4.05
Treatment (n = 28)	Mean	95.80	97.86	96.83
	SD	5.17	3.05	4.11
Total (N= 47)	Mean	95.91	97.30	96.61
	SD	4.89	3.28	4.09

Table 26

Mixed ANOVA Summary Table for Sentence Structure Accuracy Scores

<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η_p^2
Between Subjects		46				
Group	7.08	1	7.08	.351	.56	.008
Error	906.61	45	20.15			
Within Subject		47				
Time	34.26	1	34.26	2.31	.14	.049
Time x Group	15.51	1	15.51	2.31	.31	.023
Error	668.07	45	14.85			
Total	1631.53	93				

The second of the Phase II sub-questions examined the determiner accuracy scores. Table 27 displays the descriptive statistics, and Table 28 presents the ANOVA summary table. Though the p -value of .017 is not smaller than the roughly established significance level of .01, it is low enough to be of interest. Moreover, the effect size could be considered *small* ($\eta^2 = .04$) or near the border between *moderate* to *large* ($\eta_p^2 = .12$). Thus, these statistics provides some evidence that the treatment may have resulted in a meaningful improvement in the accurate use of determiners. In addition, Figure 11 plots the pretest and posttest means for each group, and Table 29 displays the simple main effects for pretest and posttest determiner accuracy scores. This additional information not only depicts the nature of the interaction effect, but it also shows significant group differences in pretest ($p < .000$) and posttest ($p < .000$) scores.

Table 27

Descriptive Statistics for Determiner Accuracy Scores

Group		Pretest	Posttest	Means
Control (n = 19)	Mean	86.62	79.44	83.03
	SD	15.05	17.14	16.10
Treatment (n = 28)	Mean	79.66	84.81	82.24
	SD	16.05	15.28	15.67
Total (N= 47)	Mean	82.48	82.13	82.31
	SD	15.87	16.01	15.94

Table 28

Mixed ANOVA Summary Table for Determiner Accuracy Scores

<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η_p^2
Between Subjects		46				
Group	14.04	1	14.04	.04	.85	.001
Error	16294.85	45	362.11			
Within Subject		47				
Time	23.38	1	23.38	.166	.69	.004
Time x Group	860.49	1	860.49	6.11	.017	.120
Error	6337.95	45	140.84			
Total	23530.70	93				

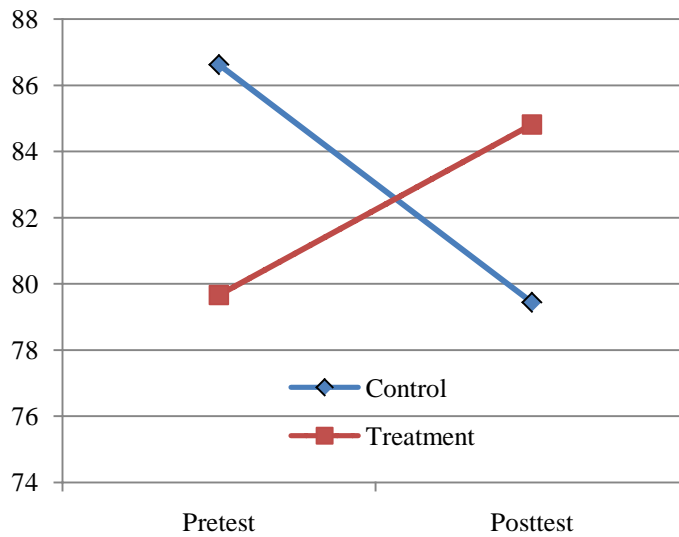


Figure 11. Pretest and posttest means for determiner accuracy scores

Table 29

Simple Main Effects for Pretest and Posttest Determiner Accuracy Scores

Source	SS	df	MS	F	<i>p</i>
Between Groups at Pretest	25.09	1	25.09	108.14	.0000000000001
Between Groups at Posttest	14.42	1	14.42	62.14	.00000000004
Error	10.67	46	0.23		

The third Phase II sub-question dealt with verb accuracy. Table 30 presents the descriptive statistics and Table 31 displays the ANOVA summary table. Although these results show that mean performance of the control and treatment groups on verb accuracy was not significantly different ($p = .08$) at a .05 alpha, we should note that its effect size is estimated as *small* ($\eta^2 = .02$) to *moderate* ($\eta_p^2 = .07$). Though minor, this effect size estimate warranted additional examination. To further understand this effect, pretest and posttest means for verb accuracy scores have been plotted in Figure 12, and simple main effects analyses for experimental grouping by testing occasion are provided in Table 32. This table suggests that while mean performance for verb accuracy scores was not significantly different for the control and treatment groups at the pretest occasion ($p = .24$), there was a significant difference between experimental groups at the posttest occasion ($p < .000$).

The fourth Phase II sub-question addressed numeric accuracy. This included accurate use of count and non-count nouns as well as the accurate production of singular and plural constructions. Table 33 presents the means and standard deviations for the control and treatment groups, and Table 34 displays the ANOVA summary table. These

Table 30

Descriptive Statistics for Verb Accuracy Scores

Group		Pretest	Posttest	Means
Control (n = 19)	Mean	74.34	69.32	71.83
	SD	14.80	25.40	20.10
Treatment (n = 28)	Mean	72.61	79.37	75.99
	SD	18.58	19.56	19.07
Total (N= 47)	Mean	73.31	75.31	74.31
	SD	17.00	22.40	19.70

Table 31

Mixed ANOVA Summary Table for Verb Accuracy Scores

<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η_p^2
Between Subjects		46				
Group	390.76	1	390.76	.72	.40	.07
Error	24597.93	45	546.62			
Within Subject		47				
Time	17.04	1	17.04	.07	.79	.002
Time x Group	784.59	1	784.59	3.33	.08	.07
Error	10602.75	45	235.62			
Total	36393.07	93				

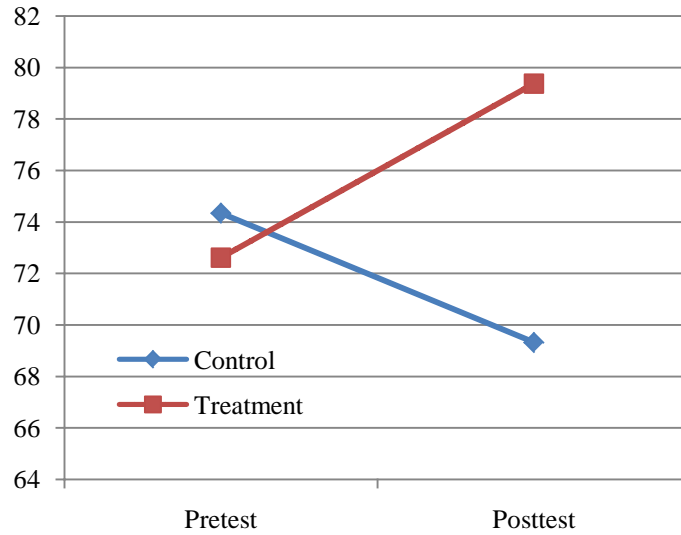


Figure 12. Pretest and posttest means for verb accuracy scores

Table 32

Simple Main Effects for Pretest and Posttest Verb Accuracy Scores

Source	SS	df	MS	F	<i>p</i>
Between Groups at Pretest	1.55	1	1.55	1.40	.24
Between Groups at Posttest	52.36	1	52.36	47.40	.00000001
Error	50.81	46	1.10		

results show no significant difference between groups and that the effects size estimates were negligible.

The fourth Phase II sub-question addressed numeric accuracy. This included accurate use of count and non-count nouns as well as the accurate production of singular and plural constructions. Table 33 presents the means and standard deviations for the control and treatment groups, and Table 34 displays the ANOVA summary table. These results show no significant difference between groups and that the effects size estimates were negligible.

Table 33

Descriptive Statistics for Numeric Accuracy Scores

Group		Pretest	Posttest	Means
Control (n = 19)	Mean	95.31	96.25	95.78
	SD	5.08	4.68	4.88
Treatment (n = 28)	Mean	93.63	93.31	93.47
	SD	6.90	8.23	7.57
Total (N= 47)	Mean	94.31	94.50	94.41
	SD	6.22	7.11	6.67

Table 34

Mixed ANOVA Summary Table for Numeric Accuracy Scores

Source	SS	df	MS	F	p	η_p^2
Between Subjects		46				
Group	120.85	1	120.85	2.08	.16	.04
Error	2611.96	45	58.04			
Within Subject		47				
Time	2.15	1	2.15	.07	.79	.002
Time x Group	8.96	1	8.96	.30	.59	.007
Error	1362.87	45	30.29			
Total	1373.98	93				

The next Phase II sub-question dealt with semantic accuracy. Table 35 displays the means and standard deviations associated with the semantic accuracy scores for the control and treatment groups, and Table 36 presents the ANOVA summary. While the main effect of “time” in Table 36 suggests a significant improvement of semantic accuracy scores, this main effect must be viewed in the context of the significant interaction effect for the within subjects “time by group” factor. Pretest and posttest means for semantic accuracy scores are plotted in Figure 13, and the simple main effects for the pretest and posttest semantic accuracy scores are displayed in Table 37.

Table 35

Descriptive Statistics for Semantic Accuracy Scores

Group		Pretest	Posttest	Means
Control (n = 19)	Mean	68.83	70.75	69.79
	SD	20.14	13.92	17.03
Treatment (n = 28)	Mean	64.85	81.11	72.98
	SD	25.26	13.81	19.54
Total (N= 47)	Mean	66.46	76.93	71.70
	SD	23.18	14.64	18.91

Table 36

Mixed ANOVA Summary Table for Semantic Accuracy Scores

Source	SS	df	MS	F	p	η_p^2
Between Subjects		46				
Group	230.45	1	230.45	.41	.52	.009
Error	25084.08	45	557.42			
Within Subject		47				
Time	1870.61	1	1870.61	10.40	.002	.19
Time x Group	1161.46	1	1161.46	6.46	.015	.13
Error	8091.83	45	179.82			
Total	36438.43	93				

Though the significance of the interaction effect ($p = .015$) was not smaller than the .01 alpha, the effect size could be estimated as near the border of *small* to *moderate* ($\eta^2 = .03$) or *moderate* to *large* ($\eta_p^2 = .13$). In addition, Table 37 shows a significant difference between the control and treatment group at both the pretest and the posttest occasions. However, an additional simple main effects analysis for each experimental group clarifies that while the increased semantic accuracy for the treatment group was significant ($p < .000$), improvement for the control group was not ($p = .42$). This seems

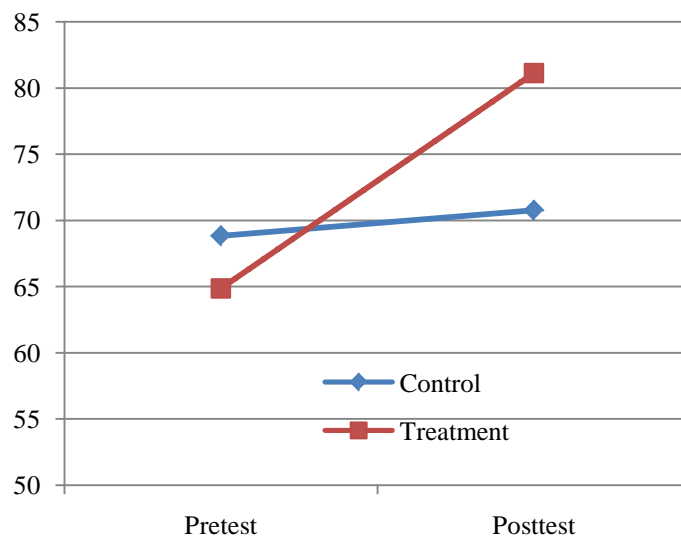


Figure 13. Pretest and posttest means for semantic accuracy scores

Table 37

Simple Main Effects for Pretest and Posttest Semantic Accuracy Scores

Source	SS	df	MS	F	p
Between Groups at Pretest	8.21	1	8.21	7.96	.007
Between Groups at Posttest	55.66	1	55.66	53.96	.00000003
Error	47.46	46	1.03		

to provide additional evidence that the treatment may have helped students in the treatment group to write with greater semantic accuracy.

The final Phase II sub-questions dealt with lexical accuracy and mechanical accuracy, both of which are of interest. Table 38 displays the descriptive statistics associated with the lexical accuracy scores for the control and treatment groups, and Table 39 presents the ANOVA summary. Though the significance for the interaction effect of the treatment on lexical accuracy ($p = .014$) was not smaller than .01, the estimated effect size was near the border between *moderate* and *large* for both the η^2 (.12) and the η_p^2 (.13). To help illustrate this interaction, Figure 14 provides a plot of pretest and posttest means for lexical accuracy scores, and Table 40 displays the relevant simple main effects. This additional information not only helps describe the nature of this interaction effect, but it also seems to underscore potential differences between the two groups prior to the treatment.

Table 38

Descriptive Statistics for Lexical Accuracy Scores

Group		Pretest	Posttest	Means
Control (n = 19)	Mean	81.23	76.84	79.04
	SD	10.00	11.54	10.77
Treatment (n = 28)	Mean	71.68	79.31	75.50
	SD	16.79	15.17	15.98
Total (N = 47)	Mean	75.54	78.31	76.93
	SD	15.07	13.73	14.40

Table 39

Mixed ANOVA Summary Table for Lexical Accuracy Scores

<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η_p^2
Between Subjects		46				
Group	283.57	1	283.57	1.03	.315	.02
Error	12339.82	45	274.22			
Within Subject		47				
Time	59.00	1	59.00	.47	.498	.01
Time x Group	817.18	1	817.18	6.48	.014	.13
Error	5679.46	45	126.21			
Total	6555.64	93				

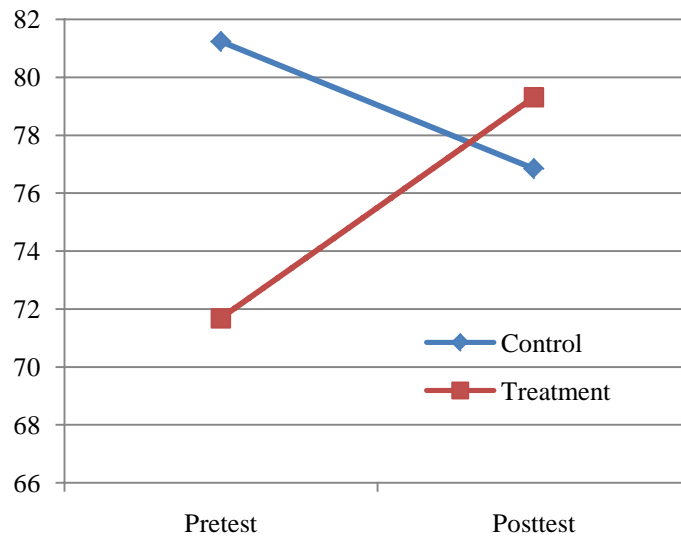


Figure 14. Pretest and posttest means for lexical accuracy scores

Table 40

Simple Main Effects for Pretest and Posttest Lexical Accuracy Scores

Source	SS	df	MS	F	p
Between Groups at Pretest	47.39	1	47.39	49.29	0.000000008
Between Groups at Posttest	3.16	1	3.16	3.29	0.08
Error	44.23	46	0.96		

Similarly, Tables 41 and 42 provide the descriptive statistics and the ANOVA summary for the mechanical accuracy scores. Mechanical errors were by far the most pervasive for both groups, and this is reflected in the fact that some of the means included in Table 41 are negative values. Nevertheless, Table 42 shows a significant interaction effect ($p = <.000$) and effect size estimates that range from between *small* and *moderate* ($\eta^2 = .04$) to *large* ($\eta_p^2 = .24$). Figure 15 and Table 43 provide additional information about this interaction, suggesting that the treatment group improved their mechanical accuracy while the control group did not.

Table 41

Descriptive Statistics for Mechanical Accuracy Scores

Group		Pretest	Posttest	Means
Control (n = 19)	Mean	15.00	-2.99	6.01
	SD	39.97	56.51	48.24
Treatment (n = 28)	Mean	-14.46	13.99	-0.24
	SD	68.99	60.58	64.79
Total (N = 47)	Mean	-2.55	7.13	2.29
	SD	60.01	58.94	59.48

Table 42

Mixed ANOVA Summary Table for Mechanical Accuracy Scores

<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η_p^2
Between Subjects		46				
Group	880.99	1	880.99	.15	.71	.003
Error	273882.10	45	6086.27			
Within Subject		47				
Time	619.89	1	619.89	.72	.40	.02
Time x Group	12204.39	1	12204.39	14.26	.0005	.24
Error	38509.43	45	855.77			
Total	326096.8	93				

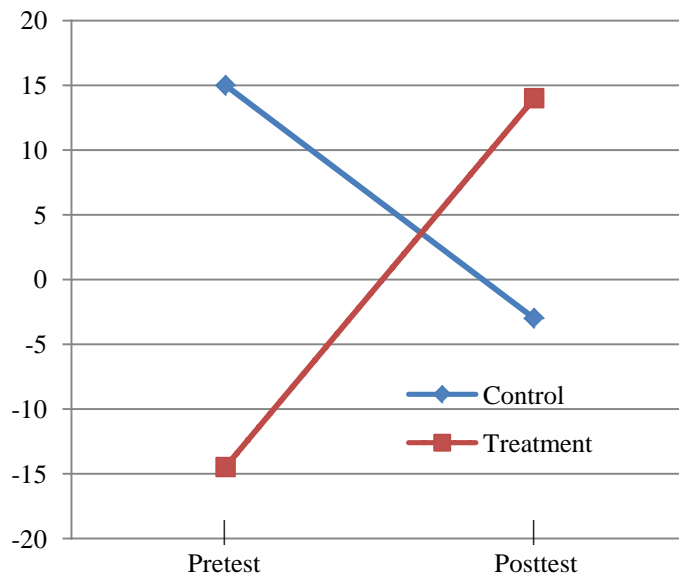


Figure 15. Pretest and posttest means for mechanical accuracy scores

Table 43

Simple Main Effects for Pretest and Posttest Mechanical Accuracy Scores

Source	SS	df	MS	F	p
Between Groups at Pretest	449.85	1	449.85	68.89	.0000000001
Between Groups at Posttest	149.47	1	149.47	22.89	.00002
Error	300.38	46	6.53		

Although this completes all of the statistical procedures originally planned to help answer the research questions in this study, two additional post hoc analyses were devised and implemented in an effort to better understand the effect of the treatment on writing accuracy. The results of the first test will be referred to as the Accuracy Index and the results of the second test will be referred to as the Grammatical Accuracy Index.

Linguistic Accuracy Index

At this point, it may be helpful to provide a brief rationale for these additional procedures. Although the statistical tests used up to this point have been beneficial, these additional procedures were developed in an attempt to overcome limitations that were not evident in the original planning stages of this research. First, let us consider the Linguistic Accuracy Index (LAI). Despite the relatively high correlations between the various sets of accuracy scores displayed in Table 6, it was noted that some correlations were much stronger than others. Moreover, it appeared that while scorers were almost always united in identifying a particular error, there was an occasional difference in how they classified the same error. It was assumed that these instances of scorer error represented the loss of valuable information, some of which could be recovered to form an overall linguistic accuracy index by (a) totaling all the errors for each essay, and (b) using this number in

the same formula that produced the various accuracy scores examined earlier: $[1 - (\text{total errors}/\text{total T-units})]100$.

Since it was assumed that this new procedure might provide additional insight about overall performance levels between the two groups, error totals were generated for each essay. Based on scoring data, these totals produced a Pearson correlation coefficient of .98. However, in an effort to avoid negative numbers for the convenience of the reader, the following formula was used, which divides the total errors by seven:

$$LAI = \left\{ 1 - \left[\frac{(\text{Total Errors} / 7)}{\text{Total T-units}} \right] \right\} 100$$

Then the same repeated measures procedure was used as had been utilized to produce the previous accuracy scores. The descriptive statistics for the LAI are included in Table 44, and the ANOVA summary is included in Table 45.

Table 44

Descriptive Statistics for the Linguistic Accuracy Index

Group		Pretest	Posttest	Means
Control (n = 19)	Mean	73.91	69.44	71.68
	SD	9.92	11.63	10.78
Treatment (n = 7)	Mean	62.27	75.68	68.98
	SD	15.37	13.59	14.48
Total (N = 26)	Mean	69.36	73.16	71.26
	SD	13.84	13.07	13.46

Table 45

Mixed ANOVA Summary Table for the Linguistic Accuracy Index

<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η_p^2
Between Subjects		46				
Group	11.25	1	11.25	.04	.85	.001
Error	13832.73	45	307.39			
Within Subject		47				
Time	138.03	1	138.03	3.58	.065	.07
Time x Group	1091.26	1	1091.26	28.30	.000003	.39
Error	1735.27	45	38.56			
Total	16808.54	93				

Table 45 shows a significant ($p < .000$) interaction effect and effect size estimates that range from *moderate* ($\eta^2 = .07$) to *large* ($\eta_p^2 = .39$). Figure 16 and Table 46 provide additional information that not only shows the nature of this interaction but that indicate significant group differences prior to the treatment. Nevertheless, they also show that on average writers in the treatment group experienced marked improvement in overall accuracy of their writing while those in the control group did not. Although the original accuracy score reported earlier was derived from a different type of calculation, the LAI appears to be more discriminating and relevant. In addition to producing a dramatically smaller p -value compared to the original accuracy scores (from $p = .001$ to $p < .0000$), more importantly, the effect size of the LAI is nearly twice as large as the original accuracy score, suggesting that measurement methods matter a great deal.

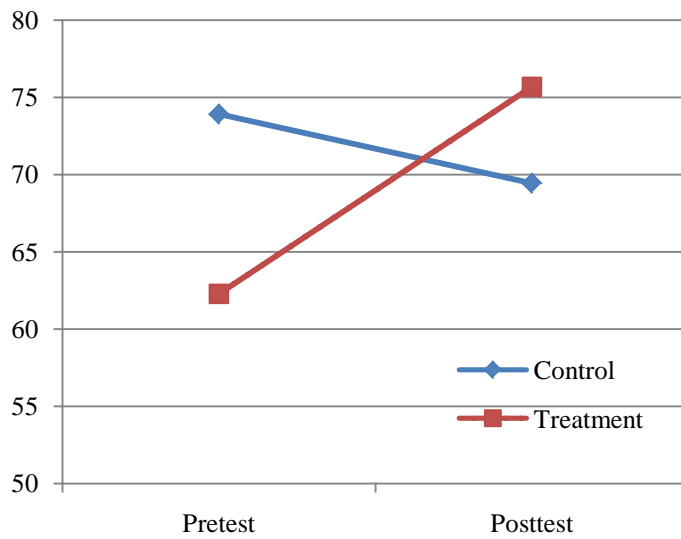


Figure 16. Pretest and posttest means for the linguistic accuracy index

Table 46

Simple Main Effects for Pretest and Posttest Linguistic Accuracy Index

Source	SS	df	MS	F	<i>p</i>
Group at Pretest	70.97	1	70.97	64.29	.0000000003
Group at Posttest	20.19	1	20.19	18.29	.00009
Error	50.78	46	1.10		

Grammatical Accuracy Index

In addition to the LAI, a brief rationale for devising and calculating the Grammatical Accuracy Index (GAI) might also be helpful. This study examined three broad areas of writing accuracy: grammatical, lexical and mechanical. While the evidence of the effect of the treatment on improved mechanical accuracy seemed quite compelling, evidence for improved lexical and grammatical accuracy, though clearly present, was not equally robust or tended to produce mixed results. Therefore, following the same logic that produced the LAI, an effort was made to minimize as much scorer error as possible to provide a more accurate indicator of the general effect of the treatment on grammatical accuracy. To do this, the formula was altered to subtract out the mechanical and lexical errors for each essay, leaving only the grammatical errors examined in this study:

$$\text{GAI} = \left\{ 1 - \left[\frac{\text{Total Errors} - (\text{Mechanical Errors} + \text{Lexical Errors})}{\text{Total T-units}} \right] \right\} 100$$

With the mechanical and lexical errors removed, the two sets of values provided by the scorers produced a correlation coefficient of .92. The descriptive statistics for the GAI for the control and treatment groups are displayed in Table 47, and the ANOVA summary is presented in Table 48. Though the within subjects “time” factor appears significant, this result must be viewed in light of the significant interaction for the “time by group” factor ($p < .000$).

Table 47

Descriptive Statistics for the Grammatical Accuracy Index

Group		Pretest	Posttest	Means
Control (n = 19)	Mean	21.17	12.23	16.70
	SD	44.88	38.94	41.91
Treatment (n = 7)	Mean	6.64	36.46	21.55
	SD	51.96	38.93	45.45
Total (N= 26)	Mean	12.51	26.66	19.59
	SD	49.24	42.86	46.05

Table 48

Mixed ANOVA Summary Table for the Grammatical Accuracy Index

<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η_p^2
Between Subjects		46				
Group	530.74	1	530.74	.149	.70	.003
Error	160492.39	45	3566.50			
Within Subject		47				
Time	2467.38	1	2467.38	4.19	.05	.085
Time x Group	8501.12	1	8501.12	14.43	.0004	.243
Error	26504.66	45	588.99			
Total	198496.29	93				

The nature of this interaction effect is further clarified by Figure 17, which plots the pretest and posttest means for the GAI, and Table 49, which displays the simple main effects for the pretest and posttest GAI. These show that the significance of the “time” factor can be attributed to improvements in the grammatical accuracy of the writing of those in the treatment group and that those in the control group did not improve their grammatical accuracy. In addition, effect size estimates range from *small* ($\eta^2 = .04$) to *large* ($\eta_p^2 = .24$), suggesting that there was a practical, positive effect of the treatment on improved grammatical accuracy.

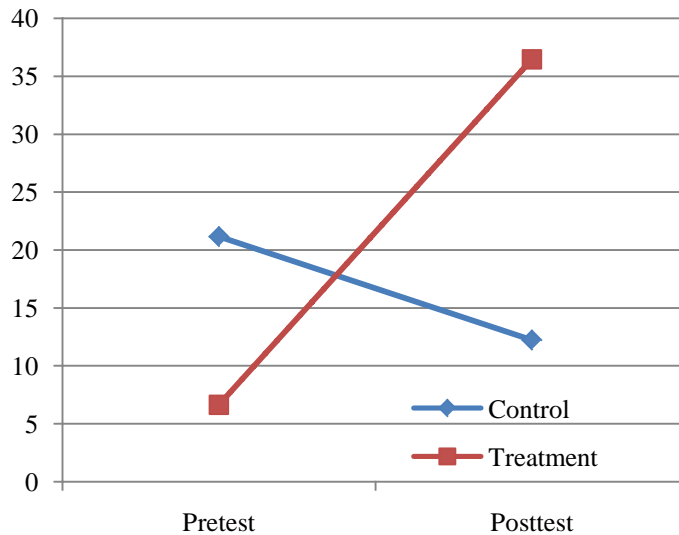


Figure 17. Pretest and posttest means for the grammatical accuracy index

Table 49

Simple Main Effects for the Pretest and Posttest Grammatical Accuracy Index

Source	SS	df	MS	F	p
Group at Pretest	109.45	1	109.45	25.84	0.000007
Group at Posttest	304.26	1	304.26	71.84	0.00000000006
Error	194.81	46	4.24		

CHAPTER 5: DISCUSSION AND CONCLUSION

The purpose of this chapter is to summarize and synthesize the results of this study, particularly in terms of the research questions and the practical implications of the study's findings. In addition to a reflective discussion of these findings, this chapter also addresses a number of limitations of the study, presents some pedagogical implications, and provides suggestions for further research.

Discussion

Although writing ability is one of the most important outcomes of higher education, many L2 writers continue to struggle to produce writing that is linguistically accurate. While some researchers such as Truscott (1996, 1999, 2007) have claimed that error correction is ineffective or that it may be harmful to learners, others have suggested that corrective feedback may provide some benefit to students in certain contexts (Bitchener & Cameron, 2005; Ferris, 2004, 2006). However, such researchers have struggled to find conclusive evidence of the value of corrective feedback. Therefore, the aim of this study was to contribute to this line of research by examining one innovative approach to L2 writing pedagogy and its effects on various aspects of L2 writing accuracy. The underlying assumptions were that accuracy might improve if feedback were more manageable, timely, meaningful, and constant.

Nevertheless, it was assumed that if the treatment produced improved writing accuracy, such improvements would be the most meaningful if they did not come at the expense of other important measures of writing development such as rhetorical competence, writing fluency or writing complexity. Moreover, it was assumed that certain aspects of writing accuracy might be more difficult to master than others, or that

the treatment might affect certain aspects of writing accuracy differently. Therefore, an attempt was made to identify how individual aspects of writing accuracy were affected by the treatment. Equally important, however, was the attempt to also provide general indicators of the effect of the treatment on the overall accuracy of L2 writing.

After a careful examination of the reliability of the scores and ratings analyzed in this study, 14 statistical tests were utilized to help answer the research questions. Twelve of these procedures were planned *a priori* and two were *a posteriori* tests devised and conducted in an effort to clarify and contextualize the results of the *a priori* tests. Since this study includes many different statistical tests, it may be helpful to provide a synopsis of these findings. Table 50 summarizes the test results used to answer the primary research question. The table includes the relevant dependent variables, the related *p*-values, the *eta* statistics used for estimating effect size, and a simple evaluation of the effect of the treatment on writing performance, indicating whether the evidence of an effect was *negligible*, *small*, *moderate* or *large*.

Table 50

A Summary of Findings Used to Answer the Primary Research Question

Dependent Variables	<i>p</i> -value	η^2	η_p^2	Effect Estimate
General Accuracy Scores	.001	.04	.21	Small to Large
Complexity Scores	.08	.02	.07	Small to Moderate
Fluency Scores	.19	< .00	.04	Negligible to Small
Rhetorical Competence Scores	.77	< .00	.002	Negligible

Table 50 shows that while the treatment seems to have significantly improved general writing accuracy, it does not appear to have improved the rhetorical competence

of the L2 writers. Though the effect size of the treatment on accuracy seems relatively large, the small to moderate effects of the treatment on complexity and fluency also need to be acknowledged. Although neither complexity nor fluency were statistically significant factors at the .05 level, these data provide enough evidence to suggest that the treatment may have had a small stifling effect on the development of writing complexity and fluency for students in the treatment group. It is possible that as some writers focus more on accuracy, they may be slightly less willing or able to produce writing that is as fluent and complex as writing produced without the same regard for accuracy.

In addition to analyzing the Accuracy Scores, it was decided *a priori* that if the treatment group demonstrated significantly higher Accuracy Scores, then additional Phase II tests would be conducted to determine which dimensions of writing accuracy might be affected the most by the treatment. Table 51 summarizes the results of these procedures. Perhaps the most salient result reflected in this table is that some dimensions of accuracy seemed to be affected more by the treatment than others. At this point, it is not possible to determine why the treatment affected production of the various dimensions of accuracy differently. It may have something to do with different levels of awareness required for accurate production in the various dimensions examined. One noteworthy observation, however, is that the two dimensions of accuracy that appear to have been affected the most include two of the three error families originally presented in Figure 2. These include the mechanical error family and the lexical error family.

Though these non-grammatical dimensions of accuracy clearly have an impact on the quality and intelligibility of one's writing, some such as Truscott (2007) have

Table 51

A Summary of Findings Used to Answer the Phase II Research Questions

Dependent Variables	<i>p</i> -value	η^2	η_p^2	Effect Estimate
Mechanical Accuracy Scores	.0005	.04	.24	Small to Large
Lexical Accuracy Scores	.014	.12	.13	Moderate
Semantic Accuracy Scores	.015	.03	.13	Small to Large
Determiner Accuracy Scores	.017	.04	.12	Small to Large
Verb Accuracy Scores	.08	.02	.07	Small to Moderate
Sentence Structure Accuracy	.31	< .00	.02	Negligible to Small
Numeric Accuracy Scores	.59	< .00	.007	Negligible

emphasized the distinction between non-grammatical and grammatical errors. While he claims that the non-grammatical errors, such as spelling, are much simpler and often can be treated in isolation with observable improvement, he maintains that grammatical errors are much different because they arise from a much more complex system. Though admitting that there is still “a need for focused research” (p. 258), he cited a number of studies (see Chandler, 2003; Frantzen, 1995; Frazio, 2001; Kempner, 1991; Lanade, 1982; Polio, Fleck & Leder, 1998; Sheppard, 1992) to support his contention that “correction may have value for some non-grammatical errors but not for errors in grammar” (p. 258). He underscored this point by concluding that “research has found correction to be a clear and dramatic failure” (p. 271).

Although Table 51 shows that the greatest evidence of the treatment effect is observed for non-grammatical error types, as suggested by Truscott’s observations, it should be noted that helping students improve the accuracy of the non-grammatical aspects of their writing may be just as important as the grammaticality of their writing. Though further research may be needed in this area, many mechanical errors such as

punctuation and spelling may be as likely to undermine effective communication as those that would be considered errors of grammar.

Nevertheless, it is interesting to note that in addition to the apparent effect of the treatment on non-grammatical aspects of accuracy, there were three grammatical dimensions of accuracy in this study where improvement seems noteworthy. These include semantic accuracy, the accurate use of determiners, and verb accuracy. While improvements in verb accuracy for the treatment group were not significant at the .05 level and had a small to moderate effect size, this positive result should be acknowledged as suggesting at least some practical significance. On the other hand, both semantic accuracy and the accurate use of determiners demonstrated effect sizes on the border of moderate to strong.

In the case of determiners, these findings are in harmony with the results of other recent studies. For example, Ferris and Roberts (2001) found greater accuracy in the use of articles following error correction. Bitchener, Young and Cameron (2005) noted improved accuracy with the definite article for those who received error correction along with teacher conferences. Similarly, Sheen (2007) examined the performance of two different types of treatment groups and noted that those who received error correction performed better than a control group on the accurate production of articles.

Interestingly, it seems that the mechanisms that underlie the production of these various types of writing accuracy are quite different. For example, though daunting for many L2 writers, determiners, are used according to a finite set of grammar rules; this is particularly true of article use. This could also be said of verb use. On the other hand, semantic accuracy, as defined and measured in this study, encompassed the application of

a much more complex body of knowledge that cannot be reduced to a simple set of rules (similar to the untreatable errors posited by Ferris, 1999, 2001). In addition to applying knowledge of appropriate word order and the obligatory contexts for certain types of words, this notion of semantic accuracy includes the appropriate use of collocations that help a writer avoid language that is awkward, unclear or simply unintelligible.

This may suggest that in addition to raising awareness of finite rules of grammar production, the methodology used in this study may have benefitted L2 writers in aspects of their writing accuracy that appear instinctive in L1 writers but that seem much too complex to reduce to a simple set of rules. Though certainly not definitive, these observations provide additional evidence of the benefit of corrective feedback for grammar errors. Moreover, these findings suggest that it may be better to examine the effects of corrective feedback on individual error types rather than using an all-inclusive grouping of “grammar errors.” This is because greater understanding of trends in L2 writing accuracy for specific grammar error types is likely to benefit and inform pedagogical practice.

The argument for corrective feedback for non-grammatical errors as well as grammatical errors becomes even stronger when we examine Table 52. Not only did the L2 writers in the treatment group benefit a great deal from general corrective feedback as demonstrated by the LAI, but they also specifically improved the grammatical accuracy of their writing as seen by the GAI. Though the most exacting statistician might interpret the positive effect of the treatment on semantic, determiner, or verb accuracy with some hesitancy, this general indicator, focusing exclusively on grammatical accuracy, is much more difficult to discount.

Table 52

A Summary of Findings for a posteriori Test

Dependent Variables	<i>p</i> -value	η^2	η_p^2	Effect Estimate
Accuracy Index	.000003	.07	.39	Moderate to Large
Grammatical Accuracy Index	.0004	.04	.24	Small to Large

Although a review of the findings of this study suggest a fairly clear benefit of the treatment on L2 writing accuracy, perhaps it would be useful to examine in more practical terms how the treatment affected accuracy in light of its effects on fluency and complexity. One important assumption in this study was that gains in linguistic accuracy would be the most meaningful if they did not occur at the expense of other important features of well-developed writing such as fluency, complexity or rhetorical competence. Although we have seen that rhetorical competence was largely unaffected by the treatment, it seems that some additional discussion of the treatment's effect on fluency and complexity is in order.

While the data suggest that the positive effect of the treatment on accuracy was much greater than its negative effect on fluency and complexity, one may wonder whether the observed increase in accuracy is worth the small stifling effect the treatment seems to have had. One way to attempt to answer this question is to convert mean scores on various measures into practical units that can be discussed in more concrete terms. For example, consider fluency. Since a test of simple main effects revealed no significant differences between the control group and the treatment group on the pretest ($p = .69$), then the posttest scores can serve as a practical estimate of the effect of the treatment on fluency. An examination of posttest means suggest that on average the treatment group

wrote approximately 36 fewer words (about one and a half to two sentences) when compared to the control group out of an average of about 388 words written during the 30-minute time limit. While both groups significantly increased their fluency over the treatment period, these data suggest that on average students in the control group produced 9% more writing than the treatment group in the allotted time.

Similarly, we should also examine the treatment's effect on complexity. Unlike fluency, however, pretest means for the two groups were statistically different, making it much more difficult to interpret the posttest results, especially since this test also included an interaction effect (see Figure 9). Although the following comments may provide some additional insight into possible effects of the treatment, ultimately they must not be interpreted independently of the interaction effect. With this in mind, it is interesting to note that the posttest means show that the control group demonstrated approximately 4% more complexity than the treatment group despite the fact that the treatment group outperformed the control group in the pretest. This equates to a mean length of T-unit that favors the control group by about one half of a word. In addition, while a test of simple main effects shows that pretest and posttest means for the treatment group were not significantly different ($p = .55$), the control group demonstrated about 11% greater complexity from pretest to posttest. This is the rough equivalent of the control group increasing their mean length of T-unit by one and a half words.

Even with these more concrete descriptions, it still may be difficult to determine whether such effects on writing fluency or complexity might be within an acceptable range. Although every increment of writing development should be viewed as important, one might well ask questions such as (a) How fluent or how complex should the writing

of these students be? (b) If accuracy may come at the expense of some fluency or complexity, how much improvement in accuracy should be expected relative to the amount of fluency or complexity that might be sacrificed? Though the answer to such questions may be difficult to decide and may vary from one context to another, perhaps the best way to address such questions is by determining how much relative improvement in accuracy was observed.

In order to quantify improved accuracy, let us return to the accuracy scores used to answer the first research question. As was the case with fluency scores, pretest accuracy scores between the control and treatment groups were not significantly different ($p = .14$), making it somewhat easier to interpret the results. However, posttest means were significantly different and suggest that the writing of the students in the treatment group was approximately 43% more accurate than the writing of the students in the control group. In other words, when compared with the writing of the students in the control group, on average the writing of the students in the treatment group included about 43% more error-free T-units per total number of T-units generated.

In addition, it might also be useful to examine the effects of the treatment on the Accuracy Index and the Grammatical Accuracy Index. However, great caution should be used since both of these tests involved interaction effects where the control group outperformed the treatment group on the pretest measures and then produced significantly lower scores on the posttest measures. With this caution in mind, pretest and posttest means suggest that on average the writing of students in the treatment group was about 18% more accurate according to the Accuracy Index (which included all error types) and about 82% more accurate according to the Grammatical Accuracy Index

(which was limited to grammatical errors). Although the accuracy scores and the scores from the Accuracy Index and the Grammatical Accuracy Index measure different dimensions of accuracy, all three suggest that the treatment had a fairly positive impact on the accuracy of the student writing.

With these results in mind, we can now use more concrete terms to describe the possible trade off between increased accuracy on the one hand and somewhat stifled fluency and complexity on the other hand. The effects of the treatment included approximately 43% greater accuracy when compared to the control group. Also, in terms of grammatical accuracy, the treatment group improved about 82% from pretest to posttest administrations. In terms of fluency, it also included about one and a half to two fewer sentences, and in terms of complexity, it included a mean length of T-unit that was shorter by up to one and a half words.

While these findings seem promising, results such as these should not be generalized to other groups without additional studies that examine larger numbers of L2 writers and that randomly assign students into experimental groups. However, if these findings represent a fairly accurate description of what might be observed from other populations, then L2 writing teachers and administrators would need to weigh the possible benefits and tradeoffs of such an approach to L2 writing pedagogy for their specific teaching and learning context. Nevertheless, it seems safe to assume that most L2 writing teachers who value linguistic accuracy would welcome the levels of improved accuracy observed in this study despite the small stifling effects they may have on fluency and complexity.

Limitations

Despite these compelling results, there are a number of limitations in this study that should be considered. Ferris (2004) has described the plight of researchers who have been criticized in their attempts to examine the effects of error correction in L2 writing. Like previous research, this study will not be exempt from potential criticism. Since this study took place with the participation of students in actual ESL classes, which were part of a more comprehensive intensive English program, a number of practical constraints were encountered.

One notable limitation of this study is that subjects were not randomly selected from a broader population of ESL students, nor were they randomly assigned into groups. Though class assignments were completely arbitrary, no systematic process of random assignment was followed. Because experimental groups were based on intact classes rather than random assignment, it is possible that the groups may have been different in significant ways. Despite rigorous placement testing to ensure similar proficiency levels, similar L1 backgrounds, and similar classroom experiences, the control group outperformed the treatment groups on many pretest measures. For example, of the eleven analyses that examined pretest and posttest measures, the control group significantly outperformed the treatment group on five. On the other hand, while the treatment group significantly outperformed the control group on only one measure, an additional five pretest measures were not significantly different.

A related point is that this study produced many interaction effects that favored the treatment group. While the posttest scores of students in the treatment group were consistently as high or higher than pretest scores for each of the 11 measures which

compared groups on pretest and posttest measures, students in the control group produced lower scores on some of the posttest measures than they generated on pretest measures. One plausible explanation for these results is that the posttest was inherently more difficult than the pretest and that the instructional method helped the students in the treatment group to write more accurately relative to the accuracy of the writing of the students in the control group.

However, another potential explanation is that group differences may have been more pronounced than anticipated and that these group differences may have affected performance differently. Also, rather than occurring simultaneously, the instructional periods of the treatment group and control group were sequential; the treatment occurred during the summer semester 2007 and the students in the control group were enrolled the previous year in 2006. Though great care was taken to ensure an optimal testing environment, it is conceivable that some unknown factor could have influenced one group and not the other since they were not tested on the same occasion.

Another obvious limitation of this study is that the number of L2 writers whose essays were analyzed was rather small due to an unexpectedly high attrition rate, resulting in only 19 students in the control group and 28 students in the treatment group at the time of the posttest. Though the number of students was much higher at the time of the pretest at the end of Level 4, some did not qualify to move on to Level 5 and others were matriculated into a university elsewhere or were transferred into another intensive English program. Similarly, of those students who began the treatment in Level 5, a number left the program before completing the treatment or taking the posttest. Future

researchers may benefit from anticipating the possibility of high attrition rates in similar kinds of studies.

Moreover, under these conditions it was not possible to fully control for teacher effect. Though one of the teachers for the control group also taught a class in the treatment group, the remaining teachers were different individuals. However, it should be noted that on average, the teachers who taught control group classes were much more experienced with the traditional method for teaching process writing than were the teachers who taught the treatment group. Nevertheless, it is possible that some of the observed effect of the treatment could be attributed to teacher differences.

Many of the reasons for these limitations arise from a change in teaching methodology for the Level 5 students beginning winter semester 2007. Since it was assumed that feedback that was more manageable and immediate was pedagogically superior to the traditional approach, all of the students at Level 5 were taught with this method starting winter semester 2007. Not to do so would, of course, raise ethical questions about the appropriateness of withholding what appeared to be the most effective teaching methodology. Thus, it seemed that the only way to draw meaningful comparisons between methods would be to compare the performance of students in the treatment group with the performance of students who were enrolled immediately prior to the curricular change who served as the control group.

In addition to these logistical challenges, another potential limitation is in how the notions of accuracy were defined in this study. While some measures, such as the error-free T-unit (EFT), are well established and provide important information about at least one aspect of writing accuracy, possible arguments could be made against the use of

EFTs and some of the other measures of accuracy used in this study. First, since the analysis of EFTs results in a dichotomous assessment of each T-unit, the approach is limited in that it does not account for the varying levels of accuracy in the T-units that do not qualify as *error free*. Thus, potentially useful information about degrees of accuracy may be lost, resulting in a less precise measurement. Notwithstanding this possible limitation, however, the EFT seems to be a practical and effective way to quantify the amount of an essay that is truly accurate.

Second, many of the other aspects of accuracy examined in this study actually measured the absence of a particular error type rather than a measure of the accurate production of a particular linguistic feature. Such measures included sentence structure accuracy, determiner accuracy, verb accuracy and so on. The potential problem with the method used in this study is that it does not distinguish between the accurate production of these linguistic features and the absence of these features. Though an alternative approach might be to try to limit analysis to the accuracy of those linguistic features which are actually attempted, this is problematic for at least two reasons.

First, a particular linguistic feature may not be used at all or may be used so infrequently that statistical analysis is not possible. Second, since L2 writing may be laden with errors that may obscure the writer's intent, it may not always be possible to identify legitimate attempts at particular linguistic features. In other words, while an accurate production would be easy to identify, inaccurate attempts at the productions of a particular linguistic feature might be so obscure that the reader is unable to identify the type of error based on the intended meaning of the writer. Thus, limiting analysis to those linguistic features that appear to have been actually attempted by the writer may be

quite impractical or may produce data that distort our view of the phenomenon being examined.

Another potential argument that might be made against the method used in this study is the possible perception that the method rewards avoidance techniques more than the actual development of writing accuracy. However, an important assumption that is central to this study is that if errors can be reduced substantially without a loss of other important qualities of writing (i.e. rhetorical competence, fluency, complexity), then the L2 writer has improved his ability to write well regardless of whether he has utilized some kind of avoidance strategy or not. Extensive observation of L2 as well as L1 writing suggests that even L1 writers use a variety of avoidance strategies in the writing process. Therefore, despite the potential limitations in how the various aspects of accuracy were defined, it was believed that examining EFTs and the absence of particular error types in the student writing were rational approaches that provided practical and useful indications of L2 writing accuracy.

Pedagogical Implications

Despite the possible limitations of this study, these results suggest a number of practical pedagogical implications. This study has shown that a systematic approach to corrective feedback can have a positive effect on the accuracy of L2 writing for both non-grammatical and grammatical errors. Moreover, these findings underscore the assertion that a model for L2 writing pedagogy that simply adopts methods from L1 writing theory and instruction may be inadequate for maximizing L2 writing accuracy. While the skills developed through process writing and the activities that strengthen rhetorical competence, fluency and complexity are important pursuits in L2 writing, they need not

be pursued at the expense of linguistic accuracy. It seems that the pursuit of linguistic accuracy can and should occupy an appropriate place in the L2 writing curriculum.

However, if the findings of this study appropriately reflect the potential benefits of error correction on improved accuracy, one important question that emerges is why similar results have not been observed more frequently in previous studies that have utilized similar types of corrective feedback. The answer to this question may be found in the unique nature of the instructional method itself. Though the treatment was multifaceted, and it would be difficult to isolate which aspects of the treatment had the greatest influence on increased accuracy, there are at least four overarching and interrelated characteristics of the feedback that were used with the intent of increasing linguistic awareness and improving writing accuracy.

The core characteristic of the feedback in this instructional method was that it was manageable. Though manageability was vital to the method in its own right, keeping feedback manageable also made it possible to ensure that feedback was meaningful, timely and constant. The relationship among these characteristics is illustrated in Figure 18, which depicts the central role of keeping feedback manageable and how the other three characteristics are complementary and flow from this center of manageable feedback. Each of these characteristics of the feedback will be summarized in the following pages.

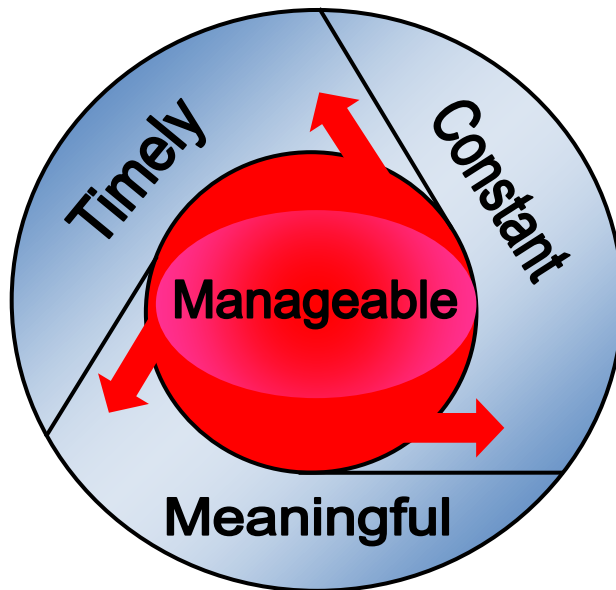


Figure 18. Characteristics of feedback designed to improve L2 writing accuracy

1. **Manageable.** Great attention was given to ensure that the corrective feedback for student writing was manageable for both teachers and students. The premise was that if the feedback load for the teacher was too great, then the quality or quantity of the feedback would likely suffer. Similarly, if the feedback given to the student was too voluminous, then the student would be more likely to be overwhelmed and less likely to be able to process, learn from and accurately apply the relevant concepts in subsequent writing. For the most part, manageability was maintained simply by limiting the new student writing to ten minutes per day. Longer compositions might have resulted in an unmanageable amount of work for both the teachers and the learners and undermined the learning process. Keeping feedback manageable

throughout the teaching and learning process also made it possible to ensure that feedback was meaningful, timely and constant.

2. **Meaningful.** For our purposes, this notion of meaningful feedback includes four related ideas. First, the feedback was meaningful to the students in that they understood the role and purpose of the feedback in the larger context of course objectives. They also knew how to interpret the codes provided by the teacher and they knew what they were expected to do with the feedback. They utilized various resources to keep track of their errors including the Error Tally Sheet (Appendix D), the Edit Log (Appendix E), and the Error List (Appendix F). They also used their feedback to rewrite compositions accurately. Second, a great deal of instruction and learning was centered on actual samples of writing generated by the students themselves. Thus, learning activities were meaningful in that students often learned from their own writing and the writing of their peers. This helped to make the learning experiences authentic and relevant to individual learner needs. Third, effort was made to help the students adequately process, learn from and apply the feedback in subsequent writing. Fourth, feedback was meaningful to the extent that it helped produced greater L2 writing accuracy. In this sense, only feedback that produced the desired results could be considered truly meaningful.

Timely. Corrective feedback was timely in that students consistently received feedback the next day following their writing experiences. Students were also expected to process this feedback in a timely manner using those resources

listed above. This notion allowed for many more cycles of student production and teacher feedback than would have been possible if these exchanges took longer. This kept students focused on their production and helped raise greater awareness as they continued to process new feedback that was based on work that was still fresh in their minds.

4. *Constant*. Closely related to the characteristics of manageable and timely, feedback was constant rather than sporadic over an extended period of time. Students wrote virtually every day, and they received feedback on their daily writing throughout the semester. It may be useful to point out anecdotally that according to the teachers, a fair number of the L2 writers included in this study had not made noticeable progress in their accuracy until the treatment was nearly half over. This constant cycle of receiving, processing and applying feedback over time may have helped the students reach a critical momentum in the feedback cycle that may have increased their awareness and accuracy beyond what might have been possible had they written and received feedback only once or twice per week.

While much more research needs to be done in order to understand exactly what should be implemented in the classroom and precisely how to implement it in diverse teaching and learning contexts, these four principles might serve as general guidelines for L2 writing classes where improved linguistic accuracy is a priority.

Suggestions for Further Research

In addition to the potential pedagogical benefits from this study, these findings also suggest a number of ideas for further research. For example, some of the statistical

tests revealed significant effects of the treatment on particular dimensions of accuracy while other tests revealed no effect. These unaffected dimensions of accuracy include sentence structure accuracy and numeric accuracy. The questions remain: “What is it about these aspects of grammar that might make them more difficult?” or “What is it about the other dimensions of accuracy that allowed learners to make significant improvements?” Moreover, since this study grouped error types into error groups and families, the specific effect of the treatment on particular error types within the error families or groups is not known. Additional research could clarify this by analyzing individual error types separately rather than examining them in groups or families. Greater understanding of trends in L2 writing accuracy for specific linguistic errors would be very useful for guiding pedagogy.

Moreover, the fact that some tests were significant while others were not raises the question of whether or not the most discriminating measurements were used in this study. For example, one legitimate question is whether the clause should have been used rather than the T-unit to measure accuracy. As was stated previously, there were a number of reasons the T-unit was chosen over the clause. First, the T-unit has arguably the best track record for measuring accuracy and has been recommended by many such as Hunt (1965) and Wolfe-Quintero et al. (1998). Second, the T-unit was a major component of many of the other measures included in this study. Therefore, using the T-unit rather than the clause as the basic unit of measurement simplified the study and made the work more efficient. Finally, researchers such as Rimmer (2006) have pointed out that using clauses can be problematic in terms of how best to define the clause and how

to deal with structural or semantic ambiguities that make it difficult for raters to identify clause reliably.

Despite these limitations, however, since a piece of writing will inevitably produce more clauses than T-units, the clause has the potential to be a more discriminating measurement if researchers carefully define what is meant by a “clause” and if they provide effective training for raters on how to deal systematically with potential ambiguities. Thus, one important focus of future research should include identifying the most discriminating way to test accuracy so subtle gains in accuracy are not overlooked.

Also, while this study focused on the effects of one instructional method with a number of different components, it is unclear whether certain elements of the method had a greater effect on improved accuracy or whether some elements were not as helpful. Additional research might help clarify this by isolating the various components of the instructional method in controlled experiments to identify those elements that have the greatest effect on improved accuracy. A related question deals with the appropriateness of this particular method for various proficiency levels. For example, how might the role of proficiency affect the improvement of accuracy at different levels? Could this methodology be equally useful with students who demonstrate lower proficiency levels such as intermediate-low or intermediate-high?

In addition to questions related to the instructional methodology, another compelling question deals with the effect of individual learner differences. For example, what might be the effect of various learner differences on accuracy such as motivation or the various ways learners intend to use English in the future? Though this study

demonstrated significant improvements in certain dimensions of accuracy for the collective group, informal observations revealed that some students made much more progress than others during the treatment period. Better understanding of individual learners could help refine methods and might better inform pedagogical practices. A related emphasis that could be given to similar studies in the future might include affective ways the treatment may have influenced the L2 writers. Thus, in addition to examining learner differences such as motivation, researchers could gather qualitative and quantitative data about student perceptions of the efficacy of the treatment, including which aspects of the treatment were the most challenging and which aspects of the treatment seemed to be the most useful.

Another important question deals with the fact that the data in this study were gathered over the course of only one semester. Therefore, one important question deals with how the results might have differed had the study continued over two or three semesters? For example, would student performance over a longer period continue to improve, plateau, or regress? Also, would we see improvement in the dimensions of grammatical accuracy that were not significant in this study such as sentence structure accuracy and numeric accuracy? In addition, would a longitudinal study result in different effects for rhetorical competence, fluency or complexity? These and many other questions could be pursued to increase our understanding about how we can help our students improve the accuracy of their L2 writing over time.

Conclusion

The purpose of this study was to determine the effect of one approach to writing pedagogy on L2 writing accuracy. A control group was taught with traditional process

writing while a treatment group was taught with an innovative approach that aimed to improve writing accuracy by raising learner awareness through error correction. This was achieved through a systematic method where students wrote for 10 minutes each day, received corrective feedback on their writing, tracked their progress, and worked toward implementing what they learned in new compositions.

Repeated measures tests using mixed model ANOVA revealed significant improvements in overall accuracy for the treatment group. The treatment also appeared to improve mechanical accuracy, lexical accuracy and some categories of grammatical accuracy. This study provides evidence that (a) grammatical accuracy as well as non-grammatical accuracy can be improved through corrective feedback, and (b) the specific methodology used for teaching L2 writing may be an important factor if linguistic accuracy is a primary objective in teaching and learning. Moreover, L2 writers may benefit the most when feedback designed to improve linguistic accuracy is manageable, meaningful, timely, and constant.

Though additional research is needed to further clarify how best to use formal teaching and learning opportunities to improve L2 writing accuracy, this study should give hope to teachers, administrators and students alike. While the path toward accurate L2 writing may be steep and strewn with challenges, substantial progress is possible. Explicit instruction coupled with ongoing practice and effective corrective feedback is likely to hasten many L2 learners along this important path in their language development.

References

- Alderson, J., Clapham, C. & Steel, D. (1997). Metalinguistic knowledge, language aptitude and language proficiency. *Language Teaching Research*, 1, 93–121.
- Arapoff, N. (1967). Writing: A thinking process. *TESOL Quarterly*, 1, 33–39.
- Bachman, L.F., Lynch, B.K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, 12, 238-257.
- Bakeman, R., & Robinson, B. F. (2005). *Understanding statistics in the behavioral sciences*. Mahwah, NJ: Erlbaum.
- Bardovi-Harlig, K. & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition*, 11, 17-34.
- Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. (2nd ed.) Mahwah, N.J.: Erlbaum.
- Berlin, J. (1984). *Writing instruction in nineteenth-century American colleges*. Carbondale: Southern Illinois University Press.
- Bialystok, E. (1982). On the relationship between knowing and using linguistic forms. *Applied Linguistics* 3, 181–206.
- Bitchener, J., Young, S. & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing*, 14, 191-205.

- Bonzo, J. D. (2005). *Who Is in Control? Topic Modulation in Spontaneous L2 Writing: Interest, Confidence, Fluency, and Complexity*. Unpublished doctoral dissertation, University of Texas, Austin.
- Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing, 12*, 267–296.
- Chaney, S. J. (1999). *The effect of error types on error correction and revision*. Unpublished Master's thesis, California State University, Sacramento.
- Coe, R. M. (1987). An apology for form: Or, who took form out of the process? *College English, 49*, 13-28.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Corbett, E. P. (1971). *Classical rhetoric for the modern student*. New York: Oxford University.
- Corder, S. P. (1981). *Error analysis and interlanguage*. London: Oxford University Press.
- Cortina, J.M., & Nouri, H. (2000). *Effect size for ANOVA designs*. Thousand Oaks, CA: Sage.
- Dvorak, T. (1986). Writing in the foreign language. In B. Wing (Ed.), *Listening, reading, writing: Analysis and application* (pp. 145-167). Middlebury, VT: Northeast Conference Reports.

Educational Testing Service. (2007). TOEFL iBT score reliability and generalizability.

Retrieved October 31, 2007, from the Word Wide Web:

http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_iBT_Score_Reliability_Generalizability.pdf

Ellis, R. (1998). Teaching and research: Options in grammar teaching. *TESOL Quarterly*, 32, 39–60.

Ellis, R. (2006). Current issues in the teaching of grammar: An SLA perspective. *TESOL Quarterly*, 40, 83-107.

Ellis, R. Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition*, 26(1), 59-84.

Erazmus, E. (1960). Second language composition teaching at the intermediate level. *Language Learning*, 10, 25-31.

Ferris, D. R. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). *Journal of Second Language Writing*, 8, 1-11.

Ferris, D. R. (2001). Teaching writing for academic purposes. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 298–314). Cambridge: Cambridge University Press.

Ferris, D. R. (2002). *Treatment of error in second language student writing*. Ann Arbor: University of Michigan Press.

Ferris, D. R. (2004). The "grammar correction" debate in L2 writing: Where are we, and where do we go from here? (and what do we do in the meantime ... ?). *Journal of Second Language Writing*, 13, 49-62.

- Ferris, D.R. (2006). Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues*. (pp. 81-104). Cambridge: Cambridge University Press.
- Ferris, D. R., Chaney, S. J., Komura, K., Roberts, B. J., & McKee, S. (2000). Perspectives, problems, and practices in treating written error. In Colloquium presented at International TESOL Convention, Vancouver, B.C., March 14–18, 2000.
- Ferris, D. R., & Hedgcock, J. S. (1998). *Teaching ESL composition: Purpose, process, and practice*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Ferris, D.R., & Helt, M. (2000). Was Truscott right? New evidence on the effects of error correction in L2 writing classes. In *Proceedings of the American Association of Applied Linguistics Conference*, Vancouver, B.C., AAAL.
- Ferris, D. R., & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing*, 10, 161–184.
- Flower, L. & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, 32, 365-387.
- Frantzen, D. (1995). The effects of grammar supplementation on written accuracy in an intermediate Spanish content course. *Modern Language Journal*, 79, 329-344.
- Green, P. & Hetch, K. (1992). Implicit and explicit grammar: an empirical study. *Applied Linguistics*, 13, 168-184.
- Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Lawrence Erlbaum.

- Haladyna, T. & Hess, R. K. (1994). The detection and correction of bias in student ratings of instruction. *Research in Higher Education, 35*, 669-87.
- Hall, T. C. (1991). A comparison of full and marginal code error correction methods in ESL writing. Unpublished master's thesis, Brigham Young University, Provo, UT.
- Hamp-Lyons, L. & Kroll, B. (1996). Issues in ESL writing assessment: An overview. in T. Silva & P. K. Matsuda (2001). *Landmark essays on ESL writing* (pp. 225-240). Mahwah, NJ: Hermagoras Press.
- Han, Y., & Ellis, R. (1998). Implicit knowledge, explicit knowledge and general language proficiency. *Language Teaching Research, 2*, 1-23.
- Hendrickson, J. M. (1980). The treatment of error in written work. *The Modern Language Journal, 64*, 216-221.
- Hinkel, E. (2004). *Teaching academic ESL writing: practical techniques in vocabulary and grammar*. Mahwah, NJ: Lawrence Erlbaum.
- Hirano, K. (1991). The effect of audience on the efficacy of objective measures of EFL proficiency in Japanese university students. *Annual Review of English Language Education in Japan, 2*, 21-30.
- Hoffman, J. V. (1998). When bad things happen to good ideas in literacy education: Professional dilemmas, personal decisions, political traps. *The Reading Teacher, 52*, 102-112.
- Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly, 18*, 87-106.

- Horowitz, D. (1986). Process, not product: Less than meets the eye. *TESOL Quarterly*, 20, 141-144.
- Howatt, A. P. (1984). *A history of English language teaching*. Oxford: Oxford University Press.
- Hulstin, J. H. & Hulstin, W. (1984). Grammatical errors as a function of processing constraints and explicit knowledge. *Language Learning*, 34, 23-43.
- Huck, S. W. (2008). *Reading statistics and research* (5th ed.). Boston: Pearson Education, Inc.
- Hunt, K. W. (1965). *Grammatical structures written at three grade levels*. Urbana , IL: The National Council of Teachers of English.
- Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing*, 4, 51-70.
- James, C. (1998). *Errors in language learning and use: Exploring error analysis*. London: Longman.
- Johns, A. M. (1995). Genre and pedagogical purposes. *Journal of Second Language Writing*, 4, 181-190.
- Kaplan, R. B. (1966). Cultural thought patterns in intercultural education. *Language Learning* 16, 1–20.
- Kent, T. (Ed.). (1999). *Post-process theory: Beyond the writing-process paradigm*. Carbondale, IL: Southern Illinois University Press.
- Kern, R., & Schultz, J. M. (1992). The effects of composition instruction on intermediate level French students' writing performance: Some preliminary findings. *The Modern Language Journal*, 76, 1-13.

- Kim, H. (2006). Providing validity evidence for a speaking test using FACETS. *Teachers College, Columbia University Working Papers in TESOL and Applied Linguistics*, 6, 1-37.
- Kline, R. B. (2004). *Beyond significance testing: reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.
- Komura, K. (1999). *Student response to error correction in ESL classrooms*. Unpublished Master's thesis, California State University, Sacramento.
- Krashen, S. D., Jones, C., Zelinski, S., & Usprich, C. (1978). How important is instruction? *English Language Teaching Journal*, 32, 257-261.
- Lalande II, J. (1982). Reducing composition errors: An experiment. *The Modern Language Journal*, 66, 140-149.
- Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, 27, 590-619
- Lee, I. (1997). ESL learners' performance in error correction in writing: Some implications for college-level teaching. *System*, 25, 465-477.
- Linacre, J.M. (1994). *Many-faceted Rasch measurement*. Chicago: MESA.
- Lynch, B.K. & McNamara, T.F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, 15, 158-180.
- Macaro, E. & Masterman, L. (2006). Does intensive explicit grammar instruction make all the difference? *Language Teaching Research*, 10, 297-327

- Macrory, G. & Stone V. (2000). Pupil progress in the acquisition of the perfect tense in French: the relationship between knowledge and use. *Language Teaching Research*, 4, 55-82.
- Mason, C. (1971). The relevance of intensive training in English as a foreign language for university students. *Language Learning*, 21, 197-204.
- Matsuda, P. K. (2001). Reexamining audiolingualism: On the genesis of reading and writing in L2 studies. In D. Belcher & A. Hirvela (Eds.), *Linking literacies: Perspectives on L2 reading-writing connections* (pp. 84-105). Ann Arbor: University of Michigan Press.
- Matsuda, P. K. (2003). Second language writing in the twentieth century: A situated historical perspective. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 15-34). New York: Cambridge University Press.
- McCollum, R. M. (2006). Validating the rating process of an English as a second language writing portfolio exam. Unpublished master's thesis, Brigham Young University, Provo.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30-46.
- McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.
- Murray, D. M. (1978). Internal revision: A process of discovery. In R. Cooper & L. Odell (Eds.), *Research on composing: Points of departure* (pp. 85-103). Urbana, IL: National Council for Teacher Education.

- Musumeci, D. (1997). *Breaking the tradition: An exploration of the historical relationship between theory and practice in second language teaching*. New York: McGraw Hill.
- Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5, 189-227.
- Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning*. New York: Cambridge University Press.
- Ojima, M. (2006). Concept mapping as pre-task planning: A case study of three Japanese ESL writers. *System*, 34, 566-585.
- Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24, 492-518.
- Park, T. (2004). An investigation of an ESL placement test of writing using multi-faceted Rasch measurement. *Teachers College, Columbia University Working Papers in TESOL and Applied Linguistics*, 4, 1-21.
- Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). Orlando, FL: Harcourt Brace.
- Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and Psychological Measurement*, 64, 916-924.
- Pincas, A. (1962). Structural linguistics and systematic composition teaching to students of English as a second language. *Language Learning*, 12(3), 185-194.

- Pincas, A. (1982). *Teaching English writing*. London: Macmillan.
- Polio, C., Fleck, C., & Leder, N. (1998). "If only I had more time": ESL learners' changes in linguistic accuracy on essay revisions. *Journal of Second Language Writing*, 7, 43-68.
- Pollitt, A. & Hutchinson, C. (1987). Calibrated graded assessment: Rasch partial credit analysis of performance in writing. *Language Testing*, 4, 72-92.
- Publication Manual of the American Psychological Association* (5th ed.). (2001). Washington, DC: American Psychological Association.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- Raimes, A. (1985). What unskilled writers do as they write: A classroom study of composing. *TESOL Quarterly*, 19, 229–258.
- Raimes, A. (1986). Teaching ESL writing: Fitting what we do to what we know. *The Writing Instructor*, 5, 153–166.
- Raimes, A. (1991). Out of the woods: Emerging traditions in the teaching of writing. *TESOL Quarterly*, 25, 407–430.
- Rennie, C. (2000). Error feedback in ESL writing classes: What do students really want? Unpublished master's thesis, California State University, Sacramento.
- Richards, J.C. & Rodgers, T.S. (2001). *Approaches and Methods in Language Teaching* (2nd ed.). Cambridge: Cambridge University Press.
- Rimmer, W. (2006). Measuring grammatical complexity: the Gordian knot. *Language Testing*, 23, 497-519.

- Ringbom, H. (1987). *The role of the first language in foreign language learning*.
Clevedon, England: Multilingual Matters Ltd.
- Robb, T., Ross, S., & Shortreed, I. (1986). Salience of feedback on error and its effect on EFL writing quality. *TESOL Quarterly*, 20, 83–91.
- Roberts, B.J. (1999). Can error logs raise more than consciousness? The effects of error logs and grammar feedback on ESL students' final drafts. Unpublished master's thesis, California State University, Sacramento.
- Roca de Larios, J., Murphy, L., Marin, J. (2002). A critical examination of L2 writing process research. In G. Rijlaarsdam (Serie Ed.), *Studies in Writing, Vol. 11* & S. Ransdell & M.-L. Barbier (Volume Eds.), *New Directions in Research on L2 Writing* (pp. 11-47). Dordrecht: Kluwer Academic Publishers.
- Scherer, G. A. & Wertheimer, M. (1964). *A psycholinguistic experiment in foreign language teaching*. New York: McGraw-Hill.
- Schumacker, R.E. (1999). Many-faceted Rasch analysis with crossed, nested, and mixed designs . *Journal of Outcome Measurement*, 3(4), 323-338.
- Scott, V. (1989). An empirical study of explicit and implicit teaching strategies in French. *The Modern Language Journal*, 73, 14-22.
- Scott, V. (1996). *Rethinking foreign language writing*. Boston: Heinle.
- Seliger, H. W. (1979). On the nature and function of language rules in language teaching. *TESOL Quarterly*, 13, 359-369.
- Semke, H. (1984). The effect of the red pen. *Foreign Language Annals*, 17, 195-202.

- Shadish, W., Cook, T. and Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inferences*. New York: Houghton Mifflin Company.
- Shaughnessy, J. J., Zechmeister, E. B., & Zechmeister, J. S. (2003). *Research methods in psychology* (6th Ed.). Boston, MA: McGraw-Hill.
- Sheen, Y. (2007). The effect of focused written corrective feedback and language aptitude on ESL learners' acquisition of articles. *TESOL Quarterly*, 41, 255-283
- Sheppard, K. (1992). Two feedback types: Do they make a difference? *RELC Journal*, 23, 103-110.
- Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly*, 27, 657–675.
- Sommers, N. (1982). Responding to student writing. *College Composition and Communication*, 33, 160–169.
- Sorace, A. (1985). Metalinguistic knowledge and language use in acquisition-poor environments. *Applied Linguistics*, 6, 239-54.
- Spiliotopoulos, V. (2003). *ESL Academic writing and electronic bulletin boards: The viability of technological supplements for writing improvement and socio-cultural development*. Unpublished doctoral dissertation, University of British Columbia, Vancouver.
- Stevens, J. (1996). *Applied multivariate statistics of the social science* (3rd ed.). Mahwah, NJ: Erlbaum.
- Susser, B. (1994). Process approaches in ESL/EFL writing instruction. *Journal of Second Language Writing*, 3, 31-47.

- Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins.
- Tanguma, J. (1999). Analyzing repeated measures designs using univariate and multivariate methods: A primer. In B. Thompson (Ed.), *Advances in social science methodology* (vol. 5, pp. 233-250). Stamford, CT: JAI Press, Inc.
- Tapai, E. (1993). *Cognitive demand as a factor in interlanguage syntax: A study in topics and texts*. Unpublished Dissertation, Indiana University, Bloomington.
- Taylor, B. P. (1981). Content and written form: A two-way street. *TESOL Quarterly*, 15, 5–13.
- Terry, R. (1989). Teaching and evaluating writing as a communicative skill. *Foreign Language Annals*, 22, 43-54.
- Tobin, L. (1994). Introduction: How the writing process was born—and other conversion narratives. In L. Tobin & T. Newkirk (Eds.), *Taking stock: The writing process movement in the '90s* (pp. 1–14). Portsmouth, NH: Boynton/Cook Heinemann.
- Tobin, L. (2001). Process pedagogy. In G. Tate, A. Rupiper, & K. Schick (Eds.), *A guide to composition pedagogies* (pp. 1–18). New York: Oxford University Press.
- Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning*, 46, 327-369.
- Truscott, J. (1999). The case for “the case for grammar correction in L2 writing classes”:
A response to Ferris. *Journal of Second Language Writing*, 8, 111-122.
- Truscott, J. (2007). The effect of error correction on learners’ ability to write accurately. *Journal of Second Language Writing*, 16(4), 255-272.

- Upshur, J. (1968). Four experiments on the relation between foreign language teaching and learning. *Language Learning*, 25, 297-308.
- Vann, R. J. (1979). Oral and written syntactic relationships in second language learning. In C. Yorio, K. Perkins, & J. Schachter (Eds.), *On TESOL '79: The learning in focus* (pp. 322-329). Washington, D.C.: TESOL.
- Wolfe-Quintero, K. Inagaki, S. & Kim, H. (1998). *Second Language Development in Writing: Measures of fluency, accuracy, and complexity*. Hawaii: University of Hawaii at Manoa.
- Wright, B.D. & Linacre, J.M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370.
- Zamel, V. (1976). Teaching composition in the ESL classroom: What we can learn from research in the teaching of English. *TESOL quarterly*, 10, 67-76.
- Zamel, V. (1983). The composing processes of advanced ESL students: Six case studies. *TESOL Quarterly*, 17(2), 165-178.
- Zamel, V. (1985). Responding to student writing. *TESOL Quarterly*, 16, 195-209.
- Zemach, D. (2007). The process of learning process writing. *Essential Teacher*, 4, 12-13.

Appendix A: Examples of Coded Feedback for Error Correction

Error Samples	Correction
1. The climber slowly ascended to ^D top.	<i>A determiner is needed before top.</i>
2. She think ^{SV} he will win the race.	<i>She thinks he will win the race.</i>
3. Eat ^{VF} pizza at parties is fun for us.	<i>Eating pizza at parties is fun for us.</i>
4. He bought pizza ^{ro} she came by they ate it.	<i>These independent clauses need to be separated or combined properly.</i>
5. Because inflation ^{inc} had risen so sharply.	<i>An independent clause is required.</i>
6. Yesterday she drive ^t to Provo.	<i>Yesterday she drove to Provo.</i>
7. He was always studying in ^{PP} 7:00 AM.	<i>He was always studying at 7:00 AM</i>
8. She was exceptional at math ^{SPG} onatics.	<i>She was exceptional at mathematics.</i>
9. He truly was a very dilig ^{WF} ence student.	<i>He truly was a very diligent student.</i>
10. She typed the paper on her calculat ^{WVC} or.	<i>She typed the paper on her computer.</i>
11. He bought five apple ^{S/PL} with the money.	<i>He bought five apples...</i>
12. She breathed in the fresh air ^{C/NC} s.	<i>She breathed in the fresh air.</i>
13. The desk (walked to the eat door.) ?	<i>(requires clarification)</i>
14. My family has 1 bother ^{AWK} and 1 sister.	<i>I have one brother and one sister.</i>
15. She ran two times the marathon.	<i>She ran the marathon two times.</i>
16. then mr. white came home. ^{C C C}	<i>Then Mr. White came home</i>
17. She said I am so happy ^{P P}	<i>She said, "I am so happy."</i>
18. I will very study very hard.	<i>I will study very hard.</i>
19. After class ^A did all my homework.	<i>After class I did all my homework.</i>

Appendix B: Rhetorical Writing Competence Rubric

Writing Rubric Adapted from the iBT TOEFL Test	
ETS Level	Description
5	<p>The essay accomplishes the following:</p> <ul style="list-style-type: none"> effectively addresses the topic and task is well organized and well developed, using clearly appropriate explanations, examples, support or details displays unity, progression, and coherence
4	<p>The essay accomplishes the following:</p> <ul style="list-style-type: none"> addresses the topic and task well, though some points may not be fully elaborated is generally well organized and well developed, using appropriate and sufficient explanations, examples or details displays unity, progression, and coherence, though it may contain redundancy, digression, or unclear connections
3	<p>The essay is marked by one or more of the following:</p> <ul style="list-style-type: none"> addresses the topic and task using somewhat developed explanations, example or details displays unity, progression, and coherence, though connection of ideas may be occasionally obscured
2	<p>The essay may reveal one or more of the following:</p> <ul style="list-style-type: none"> limited development in response to the topic and task inadequate organization or connection of ideas inappropriate or insufficient examples or details to support or illustrate generalizations in response to the task
1	<p>The essay is seriously flawed by one or more of the following:</p> <ul style="list-style-type: none"> serious disorganization or underdevelopment irrelevant specifics or questionable responsiveness to the task little or no detail
0	<p>An essay at this level merely copies words from the topic, rejects the topic, is otherwise unconnected to the topic, or is blank.</p>
<p><i>Directions to Raters:</i> The purpose of this rubric is to measure the <i>rhetorical competence</i> of the writers whose essays you will analyze. While it is understood that problems with linguistic accuracy may affect your ability to understand an essay and follow its organization and development, strive to focus on those features of rhetorical competence included in the rubric without concern linguistic accuracy. Use the benchmark essays carefully to guide your rating.</p>	

Appendix C: Partially Nested Design for Estimating Interrater Reliability

	Pretest Essay			Posttest Essay		
	R1	R2	R3	R1	R2	R3
S1	X	X		X		X
S2	X	X	X	X	X	X
S3	X	X		X		X
S4	X	X	X	X	X	X
S5	X	X		X		X
S6	X	X	X	X	X	X
S7	X	X		X		X
S8	X	X	X	X	X	X
S9	X	X		X		X
S10	X	X	X	X	X	X
S11	X	X		X		X
S12	X	X	X	X	X	X
S13	X	X		X		X
S14	X	X	X	X	X	X
S15	X	X		X		X
S16	X	X	X	X	X	X
S17	X	X		X		X
S18	X	X	X	X	X	X
S19	X	X		X		X
S20	X	X	X	X	X	X
S21	X	X		X		X
S22	X	X	X	X	X	X
S23	X	X		X		X
S24	X	X	X	X	X	X
S25	X	X		X		X
S26	X	X	X	X	X	X
S27	X	X		X		X
S28	X	X	X	X	X	X
S29	X	X		X		X
S30	X	X	X	X	X	X
S31	X	X		X		X
S32	X	X	X	X	X	X
S33	X	X		X		X
S34	X	X	X	X	X	X
S35	X	X		X		X
S36	X	X	X	X	X	X
S37	X	X		X		X
S38	X	X	X	X	X	X
S39	X	X		X		X
S40	X	X	X	X	X	X
S41	X	X		X		X
S42	X	X	X	X	X	X
S43	X	X		X		X
S44	X	X	X	X	X	X
S45	X	X		X		X
S46	X	X	X	X	X	X
S47	X	X		X		X

X = Rating Obtained

= No Rating Obtained

S_n = Student

R1 = First Rater

R2 = Second Rater

R3 = Third Rater

Appendix D: Error Tally Sheet

	<i>Too Much Freedom</i>	<i>Friendship</i>	<i>Solving Problems</i>										<i>Total</i>
D	3	4	2										9
SV	1	1											2
VF	1	1	1										3
RO													
inc		1											1
VT	1	1											2
PP	3	4	3										10
SPG	3	2	3										8
WF	2	1	2										5
WC	3	1	1										5
S/PL	1	2	2										5
C/NC		1	1										2
?		1	1										2
AWK	1		1										2
WO	1												1
C													
P	1	2	3										6
omit <i>o</i>		1	1										2
Insert	1	1	1										3
¶													
Score	<i>7.3</i>	<i>7.2</i>	<i>7.4</i>										

Appendix E: Edit Log

Topics		Edits				
1	Too Much Freedom	→	→	→	✓	
2	Friendship	→	→	→	→	✓
3	Solving Problems	→	→	✓		
4	Lawyers	→	→			
5	Care for the Elderly	→				
6						
7						

Appendix F: Error List

Error List	
<p>Determinates (D)</p> <ol style="list-style-type: none"> 1. For example, it is unsafe when <i>car</i> drives too fast on urban roads. 2. Too much going on at <i>a</i> same time can cause some stress. 3. Actually, <i>internet</i> is being used by more and more people around the world. 	
<p>Subject Verb Agreement (SV)</p> <ol style="list-style-type: none"> 1. It always <i>need</i> to be for at least one hour. 2. It also <i>increase</i> the student's ability to learn. 3. My sunglasses <i>was</i> my most expensive purchase. 	
<p>Verb Form (VF)</p> <ol style="list-style-type: none"> 1. All of the assignments <i>were been</i> completed by the end of the day. 2. People should always be willing <i>to working</i> together. 3. You must believe in yourself so you do not <i>would be failed</i>. 	

Appendix G: Institutional Review Board Approval Letter

INSTITUTIONAL REVIEW BOARD FOR
HUMAN SUBJECTS



November 28, 2007

James Hartshorn
159 UPC
Campus Mail

Re: The Effects of Manageable Error Feedback on ESL Writing Accuracy

Dear James,

This is to inform you that Brigham Young University's IRB has approved the above research study.

The approval period is from **11/28/2007 to 11/27/2008**. **Your study number is E07-0322. Please be sure to reference this number in any correspondence with the IRB.**

Continued approval is conditional upon your compliance with the following requirements:

- All protocol amendments and changes to approved research must be submitted to the IRB and not be implemented until approved by the IRB.
- A few months before this date we will send out a continuing review form. There will only be two reminders. Please fill this form out in a timely manner to ensure that there is not a lapse in your approval.

If you have any questions, please do not hesitate to call me.

Sincerely,

Christopher Dromey, PhD, Chair /
Sandee M.P. Muñoz, Administrator
Institutional Review Board for Human Subjects
CD/se