The Dissertation Committee for Jennifer Christa Holling
Certifies that this is the approved version of the following dissertation:

Evaluating the Impact of Errors Made by English Language Learners on a

High-Stakes, Holistically Scored Writing Assessment

Committee:

Davida Charney, Supervisor

Colleen Fairbanks, Co-Supervisor

Elaine K. Horwitz

Zena Moore

Janis Schiller

Evaluating the Impact of Errors Made by English Language Learners on a High-Stakes, Holistically Scored Writing Assessment

by

Jennifer Christa Holling, B.A.I.S., M.A.

Dissertation
Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirement

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2004

Dedication

To Dr. Davida Charney, whose mentorship kept me going,
who was there through the good and the bad,
and who encouraged me every step of the way.

To my sisters, Katherine Elizabeth and Stephanie Marie, who clarified my thinking,
who put everything in perspective, who inspired me, and who leant their time
and support during each step of the long, long journey.

To my family, Doris Keesee, Larry, Constance, and James Daniel Holling, who
encouraged me, who never let me give up,
and whose confidence in me meant and means everything.

To my friends and colleagues, Adelita Acosta, Dr. Janis Schiller, Dr. Karina Garcia,
and Janet O'Keeffe, all of whom encouraged and supported me
and who kept me on track.

Acknowledgements

First I would like to give my sincere thanks and gratitude to Davida Charney, Ph.D., without whose confidence in me and encouragement this dissertation would not have come to fruition. Dr. Charney is a consummate professional and an excellent mentor.

I would also like to especially thank Dr. Colleen Fairbanks for her thorough, thoughtful, and honest feedback. My thanks are extended as well to Dr. Janis Schiller, Dr. Elaine K. Horwitz and Dr. Zena Moore for their guidance, input and support. Additionally, I would like to acknowledge the superb technical advice of Dr. Edmund Emmer.

Special thanks go to my sisters, Katherine Elizabeth and Stephanie Marie Holling, who helped me think clearly at times of frustration. They inspired and motivated me at every step of the journey. Their endless support and assistance made this dissertation become a reality.

I would like to thank the rest of my family, including Doris Keesee, Larry, Constance, and James Daniel Holling, friends, and colleagues for their steadfast support and encouragement.

Finally, I would like to thank Adelita Acosta for encouraging and motivating me.

Thank you to all for your belief in my success.

Evaluating the Impact of Errors Made by English Language Learners on a High-Stakes, Holistically Scored Writing Assessment

Publication No. _____

Jennifer Christa Holling, Ph.D.

The University of Texas at Austin, 2004

Supervisors: Davida Charney and Colleen Fairbanks

Due to the ever-increasing number of English language learner (ELL) students in public schools and the increased public demand for school accountability, it is more important than ever before to uncover potential bias among high-stakes assessments. Texas is one state with an annual high-stakes assessment, formerly known as the Texas Assessment of Academic Skills (TAAS), which includes a direct

assessment of writing. The writing portion is scored holistically, and ELL students must meet the same standard as their proficient English-speaking, or non-ELL, counterparts. Prior research has demonstrated that holistic writing assessment raters are open to bias in appearance and irritation due to an overwhelming number of certain kinds of errors. Furthermore, previous research has shown that ELL students are apt to make particular surface errors that may both irritate raters and stigmatize themselves.

The equity issues underlying these findings led to the following research questions: 1) What is the nature of naturally occurring surface errors made by 8th grade ELL writers compared to those made by their proficient English-speaking peers on a high-stakes writing exam? 2) What is the nature of naturally occurring surface errors made by 8th grade writers who received a high score compared to those made by their peers who received a low score? 3) Is there an interaction between superficial errors and ELL status in the scoring of 8th grade TAAS writing exams? In order to discover if, in fact, raters of the state's writing exam are unduly influenced by the presence of surface errors in the writing of 8th grade ELL students, a random stratified sample of 50 ELL essays and 50 non-ELL essays was drawn from the 2002 administration. The essays were then parsed into t-units and errors were coded into 15 categories that were inductively determined from the sample and a review of the relevant literature. A 2 (ELL and non-ELL) X 2 (High Score and Low Score) MANOVA was performed. Main effects were found for ELL status and for scoring

status. Interactions were found for the following dependent variables: number of paragraphs, total number of errors, number of error-free t-units, and number of lexis errors per t-unit.

Table of Contents

List of Tables

List of Figures

Chapter One: Introduction

Background of the Problem

One afternoon about four years ago, my 6th, 7th, and 8th grade English as a second language (ESL) class had written entries in their dialogue journals, and I was reading and responding. When I came to the journal of Carlos, a 7th grade student from El Salvador, I was struck by a sentence he had written: **"**I leeve whit my mader, a sister" [I live with my mother and sister]. *With over four years of experience working with English language learners at the time, I thought I was fairly adept at reading learner-generated writing. I easily deciphered what Carlos had intended to write, but I stopped to think about what was actually written on the page. I imagined what any of my regular education colleagues might think if they had read that sentence. The seven-word sentence contained no less than four spelling errors and a punctuation error. I began to wonder how a general education English language arts or social studies teacher might assess Carlos's writing. Then I began to wonder how the presence of such errors might affect their grades, their scores on high-stakes assessments, and their overall academic success. It was this experience four years ago that served as the impetus for this study.*

Due to the ever-increasing number of English language learner (ELL) students in our schools and the recent focus of policy-makers on high-stakes assessments, the

issue of how ELL students' writing is rated has become more important than ever before. From 1985 to 1991, the ESL student population in K-12 schools increased by 51.3% in the United States. At present the ESL student population across the nation is increasing at a rate of more than two times that of the general student population (Clair, 1994). The Texas Education Agency (TEA) reports that in the 2003-2004 school year, 660,707 ELL students were enrolled in Texas public schools. Of those, Spanish-speakers made up the vast majority (Texas Education Agency, 2004). Educators, students, parents, administrators, policy-makers, and researchers have an interest in discovering what factors influence the scoring of ELL student writing on high-stakes assessments. Because these high-stakes assessments are linked to promotion, graduation, and school accountability, care must be taken to ensure that all students are fairly assessed and that no group is discriminated against. ELL students and their development of English language skills are often misunderstood by mainstream educators and policy-makers, as a result, their educational needs may go unmet. High-stakes assessments designed for proficient English speakers may, in fact, not be valid or appropriate for special populations like ELL students.

In Texas each year, students in grades 3-11 take the Texas Assessment of Knowledge and Skills (TAKS).[*] Each year, students sit for this multiple-choice

[*] Prior to the spring of 2003, the statewide criterion-referenced assessment of academic achievement was the Texas Assessment of Academic Skills (TAAS). The writing exam was given in grades 3, 8, and 10. The sample of essays for this study was drawn from the last TAAS general administration in the spring of 2002.

assessment in various subject areas, including: reading; mathematics; science; and social studies, depending on their grade level. Additionally, in grades 4, 7, and 10-11 (exit level) students must take an additional writing assessment. At the exit level, writing is combined with reading into one English language arts exam. The writing exam consists of an indirect measure, a multiple-choice section, and a direct measure, a writing sample. The writing sample is elicited as a response to readings on a particular theme. The students are allowed to select their own genre for response.[**] The writing samples are scored holistically on a scale of 0 to 4 (4 is the highest, and 0 is reserved for answers that are unreadable, completely off topic, or completely absent) by two raters, with a third rater employed if there is disagreement.

All students are required to take the TAKS, including students with limited English proficiency. ELL students in Texas public schools are identified by a home language survey given at the time of initial enrollment. Any home language survey indicating a language other than English either spoken in the home or by the student initiates a series of language assessments to determine English proficiency. The language proficiency assessments cover oral language, reading, and language arts/writing. If the student fails to pass all assessments, the student is considered limited English proficient (LEP). New immigrants are granted an exemption from

[**] For the TAAS, the previous state assessment, the topic elicited a particular genre—narrative, persuasive, how-to, or classificatory.

TAKS for 1 to 3 years if they meet a variety of criteria; however, ELL students who have been in U.S. schools for more than 3 years must take the high-stakes exams.[***]

Important research in second language acquisition has shown that most students can acquire basic interpersonal communicative skills (BICS) within this three-year exemption window. However, it takes considerably longer, 4 to 10 years, for students to acquire cognitive academic language proficiency (CALP), which is the academic language demanded in schools, particularly in writing (Cummins, 1980; Ovando et al., 2003). Knowing that many ELL students are still in the process of acquiring CALP when they take the TAKS, one could reasonably question whether or not it is a fair and impartial evaluation of their writing skills. In order to answer that obvious question, we need to know more about how ELL student writing compares with that of their more proficient English-speaking counterparts and how raters view the imperfect language usage of ELL students. Yet there is a marked paucity of empirical data about how ELL student writing samples are scored and whether they are treated differently by raters than those created by proficient English students.

Rationale

Previous research has pointed to ongoing questions about the reliability and validity of holistic writing assessment, such as TAAS and TAKS, in general (Charney, 1984; Huot, 1990; Hayes et al., 2000) and specifically in evaluating the

[***] In 2001, the Texas state legislature reduced the new immigrant state assessment exemption to one year; however, at the last minute it was extended to a maximum of three years.

writing of second language (L2) students (Haswell & Wyche-Smith, 1994; Hamp-Lyons, 1995). Of particular concern is that the process of holistic scoring necessitates an unsophisticated, cursory reading, in which surface errors may become more noticeable, irritating, or prominent to the reader. Additionally, researchers have questioned whether one writing sample can reflect the breadth of writing ability. For L2 writing assessment, most criticism has focused on the use of holistic scoring for placement purposes. A holistic score, by its nature, cannot diagnose specific writers' problems or effectively provide information about what courses are best suited for students.

As a result of these criticisms, improved rater training for holistic scoring has led to increases in interrater reliability for large scale writing assessments like TAKS; nevertheless, studies have shown that different raters are influenced by differing levels of background knowledge about the topic (Mosenthal et al., 1987), and/or focus on varying text features, and/or interpret scoring rubrics in divergent ways (Wolfe et al., 1998). Raters seem to react to texts more positively when the organization and topic are within their own realm of experience. Thus, ELL students who adopt a discourse pattern from their first language (L1) may unwittingly be setting themselves up for a negative reaction from raters. This research suggests overwhelmingly that holistic scoring criteria remain idiosyncratic. This does not mean that high interrater reliability cannot be reached. It simply means that how raters arrive at the scores they give differs with experience, knowledge, and other

rater variables. Which aspects of the rubric or text become salient for a particular rater and why are still largely unknown.

Other studies have found that surface features of texts, such as handwriting and mechanical errors, unduly figure into holistic scoring (Sloan & McGinnis, 1978; Freedman, 1979; Sweedler-Brown, 1992; Graham et al., 1997). Typically, texts with good penmanship and relatively few mechanical errors receive higher scores, even when other factors, such as organization and coherence, are held constant. Similarly, longer texts generally receive higher scores than shorter texts. Length, however, may be considered more than just a surface feature. It may reflect content far better than other superficial features, such as handwriting and spelling, found to correlate with holistic scores. Furthermore, there remains a discrepancy in scoring between novice and experienced raters (Breland & Jones, 1984; Ruth & Murphy, 1988). These findings persist despite training and despite rubrics that do not specifically focus on surface errors (Charney, 1984; Sweedler-Brown, 1992; Wolfe et al., 1998).

Coupled with questions about the reliability and validity of holistic writing assessment, there is another body of research that suggests that L2 writers may make specific kinds of surface errors. One particular area in which ELL writers make errors due to interference from the L1 is in mechanics. Specifically, adult L2 writers seem to make spelling errors similar in nature to those made by L1 children (Cook, 1997). If ELL students make these kinds of spelling errors as well, as a natural developmental stage of acquiring conventional English spelling, then they will seem

more out of place as the students become older. Furthermore, L2 writers from specific language communities make specific spelling errors because of the influence of the L1 graphophonemic system (Ibrahim, 1978; Bebout, 1985; Zutell & Allen, 1988; Simich-Dudgeon, 1989; Cook, 1997). For example, native Spanish speakers often spell /θ/ with a *d* because that particular phoneme does not exist in Spanish, and the letter *d* in Spanish most closely corresponds to that English phoneme. Similarly, many Asian students make spelling errors in English by confusing the use of *r* and *l*.

It seems likely, too, that L2 writers may have a tendency to rely on their L1 for punctuation rules, specifically for the use of commas and periods that impact what constitutes a sentence. Run-on sentences and sentence fragments that result from improper marking of sentence boundaries have been shown to be particularly bothersome to readers (Hairston, 1991). Although there has been no empirical data to test this particular hypothesis, ELL students may create errors in marking sentence boundaries due to interference from their L1. For instance, in Spanish, if two independent clauses are closely related to one another, it is acceptable to use a comma to separate them as semi-colons are often used in English. However, if ELL students use commas instead of periods based on rules from their L1 (as is the case with Spanish), it may well cause a negative reaction in a native English-speaking audience.

Nor are mechanical errors the only errors made by ELL writers. A great deal of previous research has been conducted to determine which kinds of errors L2 writers make in English (Sheorey, 1986; Ferris, 1992; Schairer, 1992; Dordick, 1996;

Porte, 1999; see Silva, 1993 for a review of these studies). Although it is not clear whether all or even most errors made by ELL writers are caused by interference from the L1 or are simply a function of learning English, several categories of errors are consistently found in L2 writing. These include, but are not limited to, errors in verb tense, lexis, prepositions, punctuation, spelling, word order, post-verb construction, and it-deletion. Most of these errors are common among English language learners from a wide range of native language communities, but some are especially predominant among specific native language groups. For instance, Spanish is a pro-drop language, which means that subject pronouns are not essential to sentences; they may be dropped. For example, in Spanish one can create a sentence such as: *Está lloviendo.* [Is raining.]. The subject pronoun is not expressed in Spanish, but understood from the verb conjugation and context. Therefore, it-deletion is often found in the English writing of native Spanish speakers. Also, some Asian languages do not mark verbs for tense. This causes some ELL students to make errors with verb inflections in English.

Some additional studies in contrastive rhetoric have sought to determine whether and to what extent L2 readers and writers rely on their L1 linguistic knowledge when managing reading and writing tasks in English for which they are imperfectly prepared (Nagy, et al. [1997] for reading transfer from Spanish; Cronnell [1985] for writing transfer from Spanish; Montaño-Hartman [1991] for discourse pattern transfer from Spanish).

Yet another strand of research has shown that certain errors that L2 writers consistently make are considered more serious than others. Since the mid 1980s, much research has been devoted to establishing a hierarchy of error gravity in speaking and writing (Sheorey, 1986; Ferris, 1992; Schairer, 1992; Dordick, 1996; Porte, 1999. See Silva, 1993 for a review of these studies). This research has succeeded in establishing a fair hierarchy of error gravity and in showing differences in perceptions among various audience groups (naïve native English speakers, native English speaker teachers, and non-native English speaker teachers). Overall, two findings from this strand of research are relevant to this discussion. First, verb and lexical errors appear to receive the most negative reactions. The explanation of this finding is that these errors reduce comprehensibility most. The criterion of comprehensibility is also comprised of global, as opposed to local, errors. Global errors are those that impact the whole sentence or even beyond the sentence; local errors are those within a single word or phrase. The second finding is that non-native English speakers tend to judge surface, or local, errors more harshly than native English speakers.

Research from the field of sociolinguistics also suggests that non-standard varieties are both unexpected and stigmatized, particularly in formal writing situations (Arthur, Farrar, and Bradford, 1974; Frazer, 1996; Potts and Gingerich, 1988; Wolfram, 1991; Wolfram, Adger, and Christian, 1999). Several non-standard varieties of English, including Chicano English, Vietnamese English, and African

American Vernacular English are not only spoken by students born and raised in this country, but they may also be the conversational English that newcomers to this country acquire. For students who do not yet have full control over standard written English, the influence of their spoken dialect appears to be more directly reflected in their writing (Wolfram, 1991).

The methods used in these studies varied, but frequently the studies presented errors to native and non-native speakers in isolated sentences, instead of using errors in a natural context, which reduces ecological validity, and/or they employed ESL/English as a foreign language (EFL) teachers as judges. The use of trained ESL/EFL teachers as raters or judges is problematic because language teachers often become desensitized to errors over time due to their repeated exposure to those errors (Porte, 1999; Cumming et al., 2002). In high-stakes writing assessments, the raters are not generally teachers with a lot of experience with ELL students. The raters most probably resemble native English speaking teachers. Moreover, previous research has been almost exclusively conducted using college-level L2 writing.

Statement of the Problem

Given that ELL students may make certain kinds of errors in their writing and that previous research has shown that raters may be biased against, or at least react negatively to, some of these errors, it seems logical to hypothesize that ELL student writing may be unfairly rated in holistic assessment. The risk of bias in rating would

be particularly high if the errors that ELL student writers are hypothesized to make are the same errors that are considered especially grave by native speaker, non-ESL teacher raters who score assessments such as TAKS and its predecessor, TAAS. The research indicates that this might well be true, yet there have been no empirical studies to determine if this is, in fact, the case.

Therefore, an analysis of ELL student writings on an actual high-stakes assessment must be conducted to determine the extent to which they contain the particular surface errors that previous research has suggested are particularly bothersome or irritating. Perhaps raters of high-stakes assessments for school-age students react differently to particular surface errors than those groups studied in previous research. Additionally, ELL student writing may differ markedly in the kind and frequency of surface errors it contains compared to college level L2 writing in previous studies. Furthermore, the results must be analyzed to ascertain whether or not the presence of particular surface errors are different for ELL and non-ELL writers and whether or not raters assign differential scores to students in each group based on the presence of such errors.

Research Questions

This study will address some of the questions that previous research has left unanswered. Specifically, this study will use 8th grade writing samples from a high-stakes exam. Furthermore, ratings are the actual scores given, not simply subjective

guesses of which errors would be serious or easily overlooked. Finally this study will also examine errors made by ELL students as they naturally occur; none of the errors will be contrived or manipulated by the researcher.

In particular, this study will address the following questions:

1) What is the nature of naturally occurring surface errors made by 8th grade LEP writers compared to those made by their proficient English-speaking peers on a high-stakes writing exam?

2) What is the nature of naturally occurring surface errors made by 8th grade writers who received a high score compared to those made by their peers who received a low score?

3) Is there an interaction between superficial errors and ELL status in the scoring of 8th grade TAAS writing exams?

Operational Definitions

EFL: English as a foreign language

ELL: English language learners, who are identified as limited English proficient by the laws of the State of Texas

Error-free t-units: T-units with no errors at all (Gaies, 1980)

ESL: English as a second language

Global Error: An error that impacts the sentential or inter-sentential level

High-Scoring Essay: An essay that received a holistic rating of 3 or 4

Local Error: An error that impacts only one word or phrase

Low-Scoring Essay: An essay that received a holistic rating of 1 or 2

Non-ELL: A student who is either a native speaker of English or who has met exit criteria and is now considered English proficient by the laws of the State of Texas

Surface Error: An error that is easily identifiable as an error, which may include both local and global errors

SWE: Standard written English

T-Unit: An independent clause and all of its dependent clauses (Hunt, 1965)

Chapter Two: Literature Review

Introduction

In order to discover how the errors 8th grade ELL students make on high-stakes assessments are treated by raters, one must first look to the body of existing research. Specifically, research in the areas of holistic scoring, L2 writing errors, native speaker reaction to errors, and error gravity provide useful insights. Furthermore, some sociolinguistic research regarding stigmatized varieties of English and oral dialect influence on writing is relevant to this study.

Holistic Scoring—The Inherent Dangers

Holistic scoring of writing has become widely used in high-stakes writing assessment due to its efficiencies in terms of cost and time. In holistic scoring, the rater gives each text a fast, impressionistic reading and assigns a single numerical score (Charney, 1984; Hamp-Lyons, 1995; Perkins, 1983). The score does not reflect specific strengths or weaknesses of the writing, such as mechanics or organization, yet it is often used for placement, assessment, and exit testing (Huot, 1990; Hamp-Lyons, 1995, Hayes et al., 2000). When large numbers of samples need to be assessed, as in those settings mentioned above, holistic scoring is the most feasible and cost-effective method of scoring. A handful of raters can rate a large number of essays a relatively short period of time, and often a high degree of inter-rater

reliability can be achieved (Hayes et al., 2000). Furthermore, holistic assessment of writing samples clearly has a high level of construct validity if the construct that is being assessed is overall writing proficiency (Hamp-Lyons, 2001; Perkins, 1983). Raters can take several aspects of the writing sample into account when determining a holistic score, without being required to focus on specific features. For example, the rater can consider the overall effectiveness of the argument, the support and evidence, coherence and organization, as well as writing conventions, including mechanics.

Despite the positive facets of holistic rating, there are still some serious drawbacks that cannot be ignored and may actually outweigh the positives for particular populations (Perkins, 1983; Homburg, 1984). One of the apparent issues that is unresolved regarding holistic scoring is that studies have shown that certain superficial characteristics of writing, which are not meant to factor into the scoring, may impact raters' judgments (Charney, 1984; Homburg, 1984). Specifically, expert handwriting and typing have been positively correlated to high holistic scores (Sloan & McGinnis, 1978; Freedman, 1979; Sweedler-Brown, 1992; Graham et al., 1997), while poor spelling has been found to be negatively correlated to holistic scores (Freedman, 1979; Graham et al., 1997). Kameen (1983) explains the so-called "halo" effect in writing assessment. That is, essays that conform to mechanical conventions, such as punctuation and spelling, receive higher scores. Furthermore, these biases persist despite rater training (Charney, 1984; Sweedler-Brown, 1992).

Knoblauch and Brannon (1984) summarize Diederich's comments on holistic writing assessment biases, which can either help or hinder the writer. Namely, biases may be

> based on the readers' own compositional preferences (say, plain versus ornamented style, or brevity versus length), their political commitments (liberal or conservative), their impressions of students apart from writing samples (energetic versus lazy, bright versus slow, teamplayer versus troublemaker), or their intolerance for particular errors (split infinitive or beginning a sentence with 'but') (Knoblauch & Brannon, 1984, p. 156).

Furthermore, when raters are in an evaluation setting they are more likely to focus in on "errors," which might otherwise be overlooked or which are, in fact, a matter of stylistic preference rather than instances of breaking hard and fast grammar rules. They may feel that as assessors, they have a license to identify all "errors" because of the evaluative stance they are in.

The appearance bias, in particular, in holistic rating has been well documented (Marshall & Powers, 1969; Diedrich, 1974; Sloan & McGinnis, 1978; Sweedler-Brown, 1992). One such study was conducted by Sloan and McGinnis (1978) to determine the effect of handwriting on holistic ratings assigned to essays. The researchers took a random sample of 9th grade student essays and had them rewritten by five experts in the Palmer Handwriting Method. The copies and the original essays were then holistically rated. The results showed that the essays written with expert handwriting were rated significantly higher than the originals. Furthermore,

the best essays had a greater advantage when expert handwriting was used than the essays in the lower third of the scores.

Once the appearance bias was established, researchers tested the impact of training on the bias. Sweedler-Brown (1992) conducted a study in which nine essays written in very poor handwriting, nine essays in very good handwriting, and nine typewritten essays were rewritten exactly as the original in the other two modes. Three raters who had been trained to avoid the appearance bias and another three raters who had been given no specific instruction with regard to appearance then rated the total of 81 essays. Sweedler-Brown found that training had no significant effect on the appearance bias among the raters.

Another area for potential bias in holistic rating is the undue influence of superficial features, such as mechanical errors. It seems that raters make a gross initial categorization of essays based on one feature and then further subdivide essays based on other features to determine ultimate scores. Homburg (1984) refers to this process as The Funnel Model. To investigate this possibility further, Freedman (1979) conducted a study to determine why raters assign particular ratings to college students' writing. Teachers often claim they emphasize organization and content; however, their comments often reveal an emphasis on mechanics. Therefore, Freedman took essays and rewrote them to make them strong or weak in the areas of content, organization, sentence structure, and mechanics. The essays were then given to 12 raters who were asked to holistically rate the essays as strong or weak.

Freedman found that "only if the essay had strong organization did the strength or weakness of the mechanics and sentence structure matter" (Freedman, 1979, p. 333). The interaction between organization and mechanics was found to be stronger than that between organization and sentence structure. Another finding of this study showed that mechanics was perceived to be strong or weak depending on the perceived strength of other features, such as organization and sentence structure.

Holistic rating is clearly highly subjective (Harris, 1969) since raters are susceptible to bias, fatigue, lack of internal consistency, lack of background knowledge of topics, and other factors that may cause them to focus on superficial essay features (Perkins, 1983). One reason raters may focus on surface features, such as spelling, punctuation, usage, grammatical accuracy, and paragraph formation, is that those types of essay features are readily discernible (Knoblauch & Brannon, 1984).

Still other concerns about the reliability of holistic scoring have been raised. One issue is how reliable holistic scores are when given by untrained classroom teachers. One study found that untrained classroom teachers favored students who wrote at the teacher's level of background knowledge of the topic and according to the teacher's ideological perspective of writing (Mosenthal et al., 1987). Mosenthal et al. differentiate between teachers with an academic ideology and those with a cognitive-developmental ideology. The former group prefers to focus on the product of writing and emphasizes accuracy, whereas the latter group emphasizes prior

knowledge, personal experience, and growth in writing. Another study demonstrated that scorers who are less proficient tend to focus on specific essay features, such as mechanics and local errors, rather than overall features. Moreover, they are less able to adopt the language of the rubric when thinking aloud as they holistically rate essays (Wolfe et al., 1998). Taken together, these studies indicate that there is plainly potential for rater bias is holistic scoring.

Yet another study pointed to the inadequacies of a one-time writing sample as a means of evaluating overall writing ability. Hayes et al. (2000) found that, in fact, writing performance of a particular student can vary extremely over several essays; therefore, drawing important conclusions from a single, isolated writing sample is problematic. Most of these studies on holistic scoring were conducted with participants who were writing in their first language (L1). If holistic scoring of a one-time essay is problematic for L1 writing, then it is clearly even more so for L2 writing. Raters in this setting are less likely to have the same background knowledge as the writers and may differ in storytelling styles and ideological perspectives of writing. Furthermore, raters for high-stakes assessments, such as TAAS and TAKS, are often classroom teachers or college students. Research suggests that this type of possibly unsophisticated rater may be susceptible to bias against surface errors.

Errors Made by L2 Writers in English

In addition to questions regarding the reliability and validity of holistic scoring to assess L2 writing, the literature pertaining to surface features in L2 writing, particularly surface errors, is informative. Undoubtedly, as pointed out by Gaies (1980), certain errors that are found in L2 writing either do not appear or appear far less frequently in L1 writing. Therefore, they may appear "foreign," confusing, or irritating to native speaker audiences. This section will discuss surface, or local, errors made by L2 writers and the reaction of native speakers to them.

Cronnell (1985) analyzed the errors made by third- and sixth-grade Mexican-American children that may be influenced by Spanish, interlanguage or learner language, and/or Chicano English. Cronnell found that several errors seem to be influenced by the speech pattern of those languages and by Spanish spelling. Specifically, surface features of writing are particularly prone to influence from oral language among students who do not have full control over standard written English (SWE). Specifically, speech patterns in non-standard forms may influence spelling, grammatical structures, and perhaps overall discourse patterns.

Cronnell identified seven error categories: Spanish spellings, pronunciation-consonants, pronunciation-vowels, verbs, nouns, syntax (excluding verbs and nouns), and vocabulary. The first three error categories reflected interference from Spanish and/or influence from the oral language of the students. Cronnell found that verbs were of particular difficulty for Mexican-American students. Specifically, verb

inflections were troublesome. This category included errors such as subject-verb agreement errors for third-person singular verbs in the present tense and regular simple past verbs.

Cronnell also found that because a subject pronoun is not always necessary in Spanish, several students omitted them in English, as in *is* instead of *it is*. Cronnell's syntax category also included possible interference errors, such as article usage and word order. Another finding of Cronnell's study was that one of the most frequent vocabulary errors for Spanish speakers was the use of prepositions, especially the use of *in* and *on*.

Cronnell concluded that a significant number of surface errors in writing produced by Mexican-American students in third and sixth grade could be attributed to influence from Spanish, learner language, and/or Chicano English. "Moreover, the relative presence or absence of errors was not necessarily related to the overall quality of the writing samples. A paper with many errors could be one in which the student tried harder, took more risks, and had more opportunities for errors" (Cronnell, 1985, p. 172).

Native Speaker Reactions to Errors

The literature regarding native speaker reactions reveals that native speakers generally consider errors that impede comprehension as more egregious than mechanical errors. For example, Hughes and Lascaratou (1982) and Vann, Meyer,

and Lorenz (1984) found that spelling errors do not seem to interfere with comprehensibility. Nevertheless, another strand of research on native speaker reactions focuses on irritation as the primary criterion.

It seems that specific syntactic and discourse errors may cause high levels of irritation because the presence of certain features taxes the affective perception of the native-speaking listener or reader (Ludwig, 1982). Furthermore, comprehension may be jeopardized in the presence of these irritating features since they may divert attention from the meaning and content of the message (Magnan, 1983). The irritation effect of errors cuts across several languages as these studies indicate.

Research in the area of native speaker reaction reveals that particular kinds of errors are more irritating than others. For example, Ludwig (1982) found that, in general, errors in verb forms are singularly irritating to native speakers. Additionally, Hairston (1981) found that run-on sentences, sentence fragments, and comma splices were particularly troublesome to non-academic readers. Santos (1988) found that professors who were asked to read ESL student essays were highly irritated by the presence of double negatives even though these errors did not impede comprehension. Lexical errors were found to be most serious overall in Santos's study. Santos's findings confirmed that non-native teachers are less tolerant of errors and that more experienced faculty found surface errors less irritating than did their less experienced colleagues. Though this research suggests that some errors are more bothersome than

others, there are those who maintain that all errors are equal and that sheer frequency of error occurrence is responsible for varying levels of irritation (Vann et al., 1984).

Error Gravity

Given that L2 writers are likely to make errors, researchers have tried to establish a hierarchy of error gravity. In other words, several studies have attempted to discover which errors that L2 writers make are most irritating or interfere most with comprehension. One such study was conducted by Vann, Meyer, and Lorenz (1984). Vann, Meyer, and Lorenz conducted a study to determine which common ESL writing errors were judged as most serious by academic faculty. They mention that an impetus for this study was that ESL writing teachers always struggle to strike a balance between a focus on structural and mechanical correctness and a focus on other areas of writing, such as organization, coherence, and voice. Obviously ESL writing teachers have to take into account the opinions of the audiences for whom their students will be writing. At the university in which this study was conducted, there was a complaint about the writing of foreign students made by content area faculty that focused on local, mechanical errors, such as spelling and punctuation, in addition to more global errors that interfere with communication. While the concern for global errors seemed valid, the researchers were concerned about the focus of mainstream faculty on mechanical errors.

In this study, the researchers created 36 sentences that contained common ESL writing errors and sent them, along with a demographic questionnaire, to faculty members from several schools to discover the level of gravity of each error. The 36 sentences included the following errors: spelling (British spellings and other spelling errors); articles; comma splice; prepositions; pronoun agreement; subject-verb agreement; word choice; relative clauses; tense; *it*-deletion; and word order. The results showed that respondents accepted British spellings as the least serious mechanical error, while word order was viewed as the most serious. In addition to British spellings, other common errors made by native English speakers, such as comma splice and pronoun reference errors, were accepted more than errors which tended to interfere more with comprehension, such as word order, word choice, tense, relative clause, and *it*-deletion errors.

Sheorey (1986) conducted a similar study in which native and non-native speaker teachers were asked to evaluate 20 sentences containing eight error types. The eight errors were those most frequently found in a sample of college-level ESL student writing, including: tense; agreement; article; preposition; question formation; indirect question formation; lexis; and spelling. This study found that non-native speaker teachers tend to be less accepting of errors. Furthermore, verb errors were found to be most serious by both groups of teachers. The results of this study indicated a hierarchy of error gravity among the native speaker teachers. The most

serious errors were question formation and subject-verb agreement, while the least serious were preposition and spelling errors.

Ten years later, Dordick (1996) conducted a study similar to the Vann, Meyer, and Lorenz study to determine which errors native speakers considered most serious. However, instead of using a questionnaire, he created several essays loaded with poor rhetorical style and one six common error types each. These error types included: articles, lexis, preposition, transitions, verb, and a mixture of all error types. Then he used a test to discover the level of comprehension among readers of those essays. Thus, his criterion for seriousness of errors was comprehensibility, rather than a subjective judgment of an isolated sentence as in previous studies. This study found that verb and lexis errors were the most serious as they interfered most with comprehension. Furthermore, the errors that seemed to interfere with comprehensibility least were prepositions and word order.

Yet another study regarding the level of toleration among teachers and readers for common EFL errors was conducted by Porte (1999). Specifically, Porte set out to discover if there were differences in error gravity perception between native-speaker and non-native speaker faculty. Porte points out that previous research found that errors at the word and sentence level were critical factors in causing English as a foreign language (EFL) essays to fail. The researcher created sentences including errors in eight categories that were inductively developed by analyzing a corpus of EFL student writing. Those eight error categories included: tense; subject-verb

agreement; article; preposition; post-verb construction; pronouns; lexis; and spelling. Faculty who were native speakers or non-native speakers rated the errors in terms of gravity.

The overall findings indicated little difference in the hierarchy of gravity between the two groups of faculty; however, non-native speaker teachers tended to be less tolerant of errors in general than native-speaker teachers, a finding which confirmed Sheorey's earlier study. The finding that is most important of this study is the perception of error gravity among native-speaker teachers since the raters of high-stakes writing exams, such as TAAS and TAKS, are likely to be native speakers. The most serious errors among this group were found to be subject-verb agreement and spelling, while article and preposition errors were less serious.

Stigmatized Errors

In addition to certain errors causing problems with comprehensibility and irritation, some errors are associated with particular non-dominant social groups and may carry a certain stigma. This section reviews the relevant sociolinguistic literature regarding stigmatized errors that may be made by language-minority student writers. A review of the relevant sociolinguistic literature yields the notion of stigmatization of certain errors; in fact, those very errors common to L2 writing are among those that are most frequently stigmatized. Stigmatization occurs when listeners or readers ideologize certain errors by making a connection between certain errors and a

particular marginalized social group that often uses those features. The marginalized group is then essentialized as poor, lazy, uneducated, etc. (Gal and Irvine, 1995). Conversely, Gal and Irvine (1995) show that some nonstandard features may be erased, or made invisible, because the listeners' or readers' linguistic ideology cannot explain the presence of such features. The linguistic ideology that evaluators of writing likely possess is the standard language ideology, which Lippi-Green (1994) says includes "a bias toward an abstracted, idealized, homogeneous spoken language which is imposed from above, and which takes as its model the written language. The most salient feature is the goal of suppression of variation of all kinds" (p. 974). This idealized standard is often referred to as unmarked, whereas non-standard forms are marked. That is, their use suggests that the speaker or writer belongs to a certain language-minority group (Ovando et al., 2003). This research suggests that it is not only a matter of comprehensibility or irritation that causes some errors to be treated more severely than others. This research, in fact, suggests a more insidious social bias against nonstandard variants, which the above discussion demonstrates ELL students may acquire and use in their writing.

Stigmatized Dialectal Errors—ELL and Non-ELL

The specific stigmatized varieties that ELL students may acquire and employ are well documented in the literature. It seems that students acquiring English may acquire a nonstandard dialect, such as Chicano English (Arthur, Farrar, and Bradford,

1974; Frazer, 1996; Potts and Gingerich, 1988; Wolfram, 1991; Wolfram, Adger, and Christian, 1999), Vietnamese English (Wolfram, 1991; Wolfram, Adger, and Christian, 1999), or African American Vernacular English (Wolfram, 1991; Wolfram, Adger, and Christian, 1999). While there is some continuing debate on whether or not Chicano English is a full-fledged American English dialect or just an intermediate stage on the road to acquiring standard English (Frazer, 1996), its features are, nevertheless, potential sources for rater bias.

Wolfram (1991) and Wolfram, Adger, and Christian (1999) list some specific features, which are associated with several nonstandard American English dialects. The modified list below is not comprehensive, but it shows clearly that some features are shared by several nonstandard dialects, while other are indicative of interference from the L1:

I. Stigmatized features common (and grammatical) to many nonstandard dialects of American English.

   A. Double negatives

   B. Lack of 3rd person, singular inflection (s)

   C. Lack of possessive inflection (…'s)

   D. Irregular past forms--overgeneralization of rules (e.g. knowed, goed)

   E. Lack of copula

   F. Pronoun case (e.g. Me and him are going to the store.)

G. Spelling with “d” instead of “th”

H. Irregular comparatives and superlatives (e.g. the most best, gooder, worser, etc.)

II. Stigmatized features more indicative of L2 status (interference from the L1).

A. Irregular comparatives (e.g. He has to stay here three days more.)—Spanish

B. Use of “no” in place of “not” in negatives--Spanish and Vietnamese

C. Lack of past inflection (e.g. I play soccer yesterday.) — Vietnamese

D. Graphophonic transfer errors (spelling)--Spanish

E. Prepositions--at, on and in—Spanish and Vietnamese

F. Use of “of” for possessive constructions--Spanish

Influence of Spoken Dialect on Writing

Wolfram (1991) points out that dialect features in writing “are not reflective of spoken language in a simple one-to-one relationship” (p. 257). In fact, the modes of writing and speaking are different in important ways. Writing is formal and *all* students, native and nonnative speakers alike, have difficulty acquiring skills in standard written English (SWE). However, for those who have been indoctrinated into SWE throughout their schooling and home life, most of the obvious errors (e.g. double negatives, ain’t) have been rooted out and phonological reflection of the spoken dialect is relatively rare at the secondary level. Yet for students who are still

acquiring English and who are relatively new to U.S. schools, phonological reflection of the spoken dialect in writing and use of nonstandard features is more prevalent (Wolfram, 1991). Similarly, Potts and Gingerich (1988) state that bidialectal and bilingual students are at a disadvantage in writing because "they are not print oriented to standard written English...Their intuitive grasp of English syntax is largely oral, hence the codeswitching and interference problems" (p. 111).

Even though spoken dialect is not directly reflected in the writing of all students, Wolfram, Adger, and Christian (1999) define three areas of vernacular influence on writing:

1. Organization or progression of an argument or narrative.
2. Mechanical aspects of writing, especially spelling.
3. Grammar.

It seems, then, that stigmatized spoken features, such as those listed in the previous section, may carry over into writing. In particular, these features may be present in the writing of students still acquiring English.

Another important finding in the literature that supports the idea that some students' writing may be more orally-bound is provided by Cummins (1980). Cummins has found that nonnative speakers acquire basic interpersonal communication skills (BICS) within the first year or two of arriving in a country that speaks a different language. BICS are largely oral language, and writing at this stage reflects more closely spoken forms. It is not until approximately 4-10 years after

studying a second language that a student develops cognitive academic language proficiency (CALP), which is more closely associated with SWE (Cummins, 1980; Ovando et al., 2003). Considering that acquisition of SWE takes years even for native speakers to acquire (Wolfram, 1991), it is safe to say that acquiring SWE for nonnative speakers is a long process that requires *at least* 4-10 years.

Conclusions

A careful review of the literature reveals some important findings that suggest that raters of high-stakes tests, like TAAS and TAKS, may be unduly influenced by the presence of certain superficial errors. First, it seems clear that raters of holistically scored writing assessments are subject to bias. Superficial features, such as handwriting, spelling, and other mechanical errors, consistently correlate with holistic scores. Second, ELL students are likely to make certain kinds of mechanical errors in their writing due to the influence of a nonstandard spoken dialect, transfer from the L1, or developmental difficulties associated with acquiring SWE in general, though it is not always possible to determine the exact cause from among these possibilities. Third, the errors ELL students are likely to make are irritating and stigmatized because of their ideological association with particular non-dominant social groups.

Given these findings, it seems likely that ELL students' writing may be rated unfairly by high-stakes assessment raters. Therefore, I propose to carry out a study

that will shed light on the effect of superficial errors on the holistic scoring of ELL students' writing in high-stakes assessments. The main research questions are as follows: 1) What is the nature of naturally occurring surface errors made by 8th grade ELL writers compared to those made by their proficient English-speaking peers on a high-stakes writing exam? 2) What is the nature of naturally occurring surface errors made by 8th grade writers who received a high score on the 2002 TAAS assessment compared to those made by their peers who received a low score? 3) Is there an interaction between superficial errors and ELL status in the scoring of 8th grade TAAS writing exams?

Chapter Three: Methods and Procedures

Research Questions

In light of the findings reported in Chapter 2, the following research questions are asked in this study:

1) What is the nature of naturally occurring surface errors made by 8th grade ELL writers compared to those made by their proficient English-speaking peers on a high-stakes writing exam?

2) What is the nature of naturally occurring surface errors made by 8th grade writers who received a high score compared to those made by their peers who received a low score?

3) Is there an interaction between superficial errors and ELL status in the scoring of 8th grade TAAS writing exams?

The Writing Samples

For this study, a random sample of 100 8th grade TAAS writing samples was drawn from the 2002 exam administration. The TAAS exam was scored using focused holistic scoring, "which take[s] into account the student's developmental capabilities and the constraints of the testing situation" (TEA, 1999, p. 3). The focused holistic process requires raters to evaluate the writing according to four pre-established writing objectives, which are based on the writing curriculum in public schools in the state of Texas:

Objective 1: The student will respond appropriately in a written composition to the purpose/audience specified in a given topic.
Objective 2: The student will organize ideas in a written composition on a given topic.
Objective 3: The student will demonstrate control of the English language in written composition on a given topic.
Objective 4: The student will generate a written composition that develops/supports/elaborates the central idea stated in a given topic (TEA, 1999, p. 3).

For the 2002 exam administration for 8th grade, the writing prompt was as follows:

> Think of a project that you have done.
> Write a composition for your teacher explaining how you did this project. Be sure to include step-by-step instructions so that someone else could complete this project the way you did (TEA, 2002, p. 18).

This topic elicited a how-to essay. Several of the essays, both those written by ELL and non-ELL students, were written in letter form to the students' teachers.

First, the complete group of 2002 8th grade TAAS essays was divided into two groups: ELL and non-ELL and a stratified random sample was drawn. Within each of those groups, 12 essays with a score of 4, 13 essays with a score of 3, 13 essays with a score of 2, and 12 essays with a score of 1 were randomly selected. An effort was made to divide the ELL and non-ELL groups evenly across gender, but for ELL students this was not possible. There was no stratification on the basis of ethnicity, and because the L1 was not reported on the TAAS answer document, there was no opportunity to know with certainty what language individual students speak natively. Nevertheless, those ELL students of Hispanic ethnicity are assumed to speak Spanish

as a primary language. For those ELL students who are ethnically categorized as Asian, the L1 could be a number of common Asian languages spoken within the public school system of the state of Texas, including Vietnamese, Korean, and Mandarin Chinese (TEA, 2004).

The Texas Education Agency's Student Assessment Division conducted the stratified random sampling. Before releasing copies of the writing samples to the researcher, TEA removed all identifying words from the essays, including names of schools, towns, and/or individuals and numbered the essays in a random order. The 100-essay sample was turned over along with information about scores, ELL status, gender, and ethnicity. That information is shown in Table 1 and 2 below.

Table 1. Gender Distribution of Sample

Gender	Non-ELL	ELL	Total
Female	25	32	57
Male	25	18	43
Total	50	50	100

Table 2. Ethnic Distribution of Sample

Ethnicity	Non-ELL	ELL	Total
African American	10	1	11
Asian	2	5	7
Hispanic	19	41	60
Native American, Pacific Islander	0	1	1
White	19	2	21
Total	50	50	100

Procedures for Data Collection

The demographic and descriptive data received from TEA were then entered into a database in SPSS by the researcher and checked by a research assistant, who is a graduate student in psychology. The random order and numbering of the essays provided by TEA was retained. Then all essays were typed into a word processing program exactly as written. If any words were illegible due to poor handwriting and/or poor copy quality, it was noted in the typed text as "[illegible]." The typescripts were checked and double-checked by two assistants to ensure that all errors were typed exactly as originally written. The second assistant is a middle school teacher with 10 years of experience with 8th grade students. In addition, total word, sentence, and paragraph counts were compiled and entered into the database.

Data Coding

T-Units

Next each essay was parsed into t-units, which are minimal terminal units first defined by Hunt in 1965. T-units include an independent clause and all subordinate clauses that go along with it. T-units are preferable to sentences as units of measure, especially in children's writing for two main reasons. First, when determining t-units, the punctuation of the writer is ignored. Novice writers often improperly use end punctuation, creating fragments, comma splices, and run-on sentences. T-units look beyond incorrect punctuation to the syntactic unit of a true sentence. Secondly, the

use of t-units allows for comparison of long and short essays. Instead of examining absolute counts of words and errors, essays of varying lengths can be compared by using counts per t-unit. As Hunt (1965) suggests, any extraneous fragments not related to a t-unit, such as titles, greetings, and/or salutations, were eliminated at this point. Only complete t-units were imported into a software program commonly used for qualitative analysis and coding of text, QSR N6. Once the data files had been imported into QSR N6, the total number of t-units per essay was calculated by the program and the data entered into the SPSS database.

Surface Errors

Coding categories were determined inductively beginning with those categories suggested by a review of the literature: verb tense; lexis; preposition; punctuation; spelling; word order; post-verb construction; and it-deletion. However, these categories did not match the actual errors made by 8th graders in this sample. Hence, other error categories were added and some of the above categories were omitted or compressed in order to more accurately code the writing in this study. The final error sub-categories were organized into four major categories: Sentence Boundary (comma splice, fragment, and run-on); Mechanical (apostrophe, capitalization, punctuation, and spelling); Verbal (subject-verb agreement, verb, and verb tense); and Other Surface (article, lexical, preposition, pronoun reference, and other). All error categories can be considered local as they impact single words or

phrases, except the "other syntactic error" category that includes some errors that can be considered global. However, all are surface errors because they are readily discernible. Table 3 describes the error categories and lists examples of each.

Reliability

Each t-unit was then coded according to the frequency and kind of surface error present. After all errors had been coded by the researcher, a random sample of 10 essays was drawn using a random digit chart. A second coder was then trained in identifying t-units and asked to divide the essays into t-units. In order to take chance agreement into account, a Pearson Product Moment correlation was calculated between the two coders with reference to t-unit identification. A correlation of .97 was found. The second coder was then trained in identifying surface errors and categorizing them according to the established coding scheme (see Table 3). All 10 essays were coded by the second coder, and a correlation of .88 was achieved. Hayes and Hatch (1999) suggest that using a correlation measure to establish interrater reliability is more appropriate than a simple percentage of agreement because it takes chance agreement into account.

Inferential Analysis

Next a 2 (ELL and non-ELL) X 2 (High Score and Low Score) MANOVA was performed to test for differences between groups for general essay features, including: total words; total sentences; number of paragraphs; total errors; total error-

Table 3. Coding Scheme

Error Code		Description	Example
Sentence Boundary	Comma Splice	Two independent clauses separated only by a comma	Dolls are for your models, they shouldn't be to large.
	Fragment	A dependent clause or phrase separated by a period or other end punctuation	While, teacher is High-scoring out compass and rulers. Get out Peice or PaPers.
	Run-on	Two or more independent clauses with no punctuation or coordinating conjunctions to separate them	How you do this is very simple you put some feed in the pen and take the goat out of the pen you then let the goat loose and he will run back to the pen.
Mechanical	Apostrophe	Incorrect usage of an apostrophe either by a) insertion or b) omission	a) To get the goat lot's of muscles you need to walk them everyday for about thirty minutes. b)but it doesnt taste that good.
	Capitalization	Incorrect capitalization either by over- or under-capitalization*	In my Science class we were assigned to do a Science project.
	Punctuation	Incorrect punctuation, including: comma usage, direct speech, the use of colons, semi-colons, hyphens, and dashes	She said is that all I said yes. "--------- did you know that people use diffison every day and don't even know it?
	Spelling	Any non-conventional spelling if the intended word could be surmised; this category includes substituting your for you're and similar errors	So we whent back home live [leave] it like that then after your done with it you need to tape it to the board.
Verbal	Subject-verb Agreement	Incorrect or lack of agreement in number between the subject and verb	Just whatever you decides to make.

* Some capitalization errors appeared to be idiosyncrasies of poor handwriting. For example, some students seemed to capitalize every "r" or every "p" throughout the essay. These errors were counted because they may have contributed to irritation among the raters. For example, *while otheR people pRefer something different.*

Table 3. Coding Scheme

Error Code		Description	Example
	Verb	Any verbal error that was not subject-verb agreement or verb tense, such as omitting part of a complex verb tense	I just sending this letter to explaining how I did my project.
	Verb Tense	Incorrect verb tense if the verb actually written was a valid tense, but simply the wrong tense	Before drawing pictures began with writing.
Other Surface	Article	Incorrect article usage either by omission, insertion, or wrong choice	The first thing you do is, you set up old Bike
	Lexis	Improper word usage or any words that were unrecognizable as an English word	One of my better project is, when I make a world of paper with the 7 countries.
			the first think the I did it was when I put the titulo in the poster
	Pronoun Reference	Use of any pronoun without a referent or with a referent that doesn't match	All of those picture will be glue into the big cloth. You will have to glue it [them] very carefully.
	Other Syntactic	Any other error that could not be classified by the above categories, including word order, word omission, or other syntactic error	is the same thing that Vegas make on the little bulbs turn on and off.
			but also you must make a chart of how much time did the chalk take to dissolve.

free t-units; words per t-unit; error-free t-units per t-unit; sentences per t-unit; and errors per t-unit. Other dependent variables tested were the specific surface error categories outlined above. The use of the MANOVA was necessary to discover if raters treated errors differently across ELL and non-ELL groups and across high-scoring and low-scoring essays. Upon finding main effects and interactions, further post hoc t-tests were run to further describe the data.

Chapter Four: Results

Introduction

In order to address the research questions, a 2 (ELL, non-ELL) X 2 (High Score, Low Score) MANOVA was conducted. After finding the main effects, follow-up t-tests were conducted in order to show specific differences among the essays. This chapter will begin with four examples of entire student essays that will assist the reader in visualizing what the quantitative results might look like in context. Following the four sample essays will be a presentation of the results of the statistical procedures, focusing on the first two research questions and explaining the numerical results further using excerpts from the actual essays. The next section will present an analysis of the interactions revealed by the MANOVA. These sections of the chapter will be presented in order of the research questions they reference. A summary of the findings as they relate to the research questions will conclude this chapter.

Sample Student Essays

The four essays below (see Figures 1, 2, 3, and 4) are included to give the reader a complete view of what typical ELL and non-ELL student writing looks like at a high-scoring and low score level. The first essay is of a non-ELL high-scoring essay. The next sample is an ELL student essay that received a high score. The third essay is that of a non-ELL student that scored low, while the final sample is that of an

ELL student that scored low. The score levels shown are at the high end of the high-scoring or low-scoring levels. A score of 2 is the highest low score, while a score of 4 is the highest high score possible.

Figure 1. **Sample #1**

At summer school my ingeneering teacher told us all to construct a catapolt. Out of all the catapolts mine shot the farthest clearing over forty yards in distance. Since I enjoy sharing my knowledge, I will give you step by step instructiones on how to build this marvelouse wonder.

First gathering your materials. The first item you will be needing is a strong rat trap. You can find this at a local hardwear store. Second go purchase some wire, about a foot, a box of sturdy popsicle stix, and about a yard of kite string. You can find all of these items at a arts and crafts store nearest you. Furthermore, you'll need some sticky wood glue and a pair of powerful pliers, these can be found at a trusty hardware store or your tool box. Another material you will be using is a very sharp saw, you can find in your toolbox, and four tires you can attach to the catapolt. You can buy these tires at a toy store or an arts and crafts store. Last you'll need news paper and a ruler, these items are sold at most H-E-B stores.

Now you may commence building the catapolt, but first here are the steps. First get your rat trap and with the pliers pull out the pull pin, the long rod, and

the pin holder, the loose metal piece that holds the pull pin when trap is armed. All that should be left is the staple that held the pull pin and the trap itself. Next glue for popiscle stix together flat side to flat side, then do it a second time so you have two. Now glue both of them together so it makes a long stick. Since you use wood glue it will all dry fast and strong so try not to mess up. Once the stick has dried, make a very shallow box with the center about the size of the marble. For example, the marble should easily fly out of the box, so it will go farther. When glueing the box to the stick make sure you saw off all extra wood sticking out, so when the box is glued to the stick it looks like a spoon. Next glue about two layers of popsicle stix to the base of the powerful rat trap. Once this is all dried grab your trusty kite string and tie the spoon like stick to the spring of the rat trap. But befor you do this tie down the trap sling, so it doesn't hurt you, and face the arm in the direction of how the trap closes. Do this so the marble won't go backwards. The catapolt should be taking shape, all thats left is two steps. The first is loosen the trap sling so it will stop at about a seventy-five degree angle. If the string week you might want to add extra string to down the sling. Right after you do that tie the arm to the sling so when the sling stops so will the arm. Last get the wheels and attach them to the bottom two layers of popsicle stix. Don't worry the staple you use to tie down the stix, won't come out or loosen. Now just fold the newspaper and throw it away.

Finally your project is complete. You can now shoot marbles all the way across your yard. Be careful this is not a toy you can be seriously hurt, don't shoot marbles at people or windows. I want to thank you for building this cretion, and know you'll have a lot of fun with it.

Essay 15, Hispanic non-ELL Male, Score 4

Figure 2. **Sample #2**

Dear Ms. ---------,

Have you ever done an chalk experiment in your science class? Well I have and If you haven't done one, well then this is the time you to do one just follow my steps and instruction for your can do an chak experiment and you will see how much fun it is to do one.

First you follow the instructions that your teacher gave you to follow in the booklet that she place on the taple and get all of the materials that you need to compleat the project like if you are going to do the chalk experiment then you will need, two beackers, two cylinders, two pieced of chack (the same size), vinegar and a hot plate.

Then when you have gather all of the materials together then you start the experiment.

Next you get the hot plate plug it in and turn it on to 350° and let is get hot and it doesn't take more than two to three minutes.

After the plate is hot full in the two cylinder up to 50 ml with vinegar and check if the two cylinder have the same amount.

Then you get the two beakers and put them on the table, and you get the two cylinder with the vinegar inside and slowly dip the vinegar to the two beakers and check if the two beaker have the same amount of vinegar.

Next you get on beacker put it on top of the hot plate and put the other beaker a side o the hot plate like 10 in away from the hot plate.

After you had done that get a watch ang take track of time when you got the time ready then get the two piece of chalk put them on top of the beaker, and when the errow is on the twelve or hit twelve let the two pieces of clalck go and each beaker must have there an piece of chalk.

When that is happening you must kip on looking at the watch and take track of time to see witch one takes more time to dissolve in matter of minutes, but also you must make a chart of how much time did the chalk take to dissolve.

Then check if the chalk is comeplitly dissolve into liquid or just dost in water as sand in the ocean, and if not then give it more time to dissolve but you must kip track of time.

When that is done you take the beaker from the hot plate and turn or the hot plate and plug it out, Then grabe a paper and a pin and write down how much time did the chalk on top of the hop plate took less time then the one on the taple with out no heat and one with heat.

After all that you get a sheet of pape and you start discribing what did you just do and what happen with details.

Next you get that information and pasted it somewhere in the board and write the problem hypothesis, Data, conclusion and introduction and paste it to in the board but not together seperateded and write on too what is it a problem, hypothesis, data conclusion or the introduction and you have done that the past every thing and give it to your teacher.

In conclusion, see it is not hard to do an experiment you just have to follow the steps and the instructions and you will see it is easy to do.

Sincerely

Essay 29, Hispanic ELL Female, Score 4

Figure 3. **Sample #3**

You need a perfect grade on your project to pass, but you do not know what to do. Make an active volcano that really erupts. Just follow this simple instructions and you're on your way to an A.

Before you start, you need to have your parents permmission. When they say yes, ask them for your materials, which is: 1 box of baking soda, 1 bottle of vineger, newspaper, water, flour, red food die, spray paint, and a cardboard base.

After you have collected all your materials, you can start making your volcano. Mix the flour and water together to make a paste like substance. Take the newspaper and make nine paper balls about the size of a baseball. Use six of

them on the bottom to make a circle, then with the three remaining balls make a triangle on top of the circle. With the rest of the paper tear strips, then take them and let them soak in the paste mixture. Then cleaning off the extra paste, lay the strips over the mound until you can't see the paper balls and there is 7 to 8 layers on the mound. Then let it set to dry a couple of days. After it has dried, cut a hole in the top for the baking soda and vineger to be put in. Then paint it whatever colors you want and glue onto the cardboard base.

Finally, test it and see if it works. If it doesn't work like you want it to, you may need to adjust the amounts of ingredents used. When finished, clean up your mess and take it to school the following day and present it to your teacher.

If you have followed all of these instructions right, then you should get the grade you wanted and deserved for making this project a big success

Essay 71, White non-ELL male, Score 2

Figure 4. **Sample #4**

If you are trying to do a book project but you don't know how then no problem! Follow my step and I will tell you how to do a book project.

First, you need some color pancils or markers, a big poster board, a book that you are doing your project on. And be sure to read the whole book because you will be doing a summery for that book and you need eraser and pancil.

Second, you take your poster board and put it on a table. Now think about what your going to draw in your poster board. Know clue! Well, let me give you Some Ideas. Well there has to be something in the book you like. Something like a hunted house or in the woods, jungle, or forest. Now do you know what to draw; then take your pancil and get started! OK now draw it neatly because you want to get to get and 100 on this. right? After your finish drawing it then start coloring it and do it neatly so when you present your poster student in the back of your class could see it better. Now you have to do your summery. Sorry I can't help you with it because I don't know what book you are doing but I can tell you that the semmery needs a beganing, middle, and end.

Last, how you have your project all done. Good job! but who will clean up you mess. I am not cleaning your mass but I can tell how to do it. now take your markers or color pancils and pancil and put it on your pancil bag or box then take your poster board be careful and don't fold it because it will mass your poster up but you could role it up and put a plastic band on it.

Now you have your project and your semmery ready to go and your room looks clean too. You don't have to worry about your project anymore. Just relax, have some fun or do your other homework.

Essay 84, Asian ELL Female, Score 2

These samples show typical writing of non-ELL and ELL students at different score levels. It is important to remember that the assessment raters read whole

essays. These sample essays further reflect the reality of the following descriptive, comparative, and inferential statistics and provide a better indication of what the raters saw.

Description of the Sample

The overall sample used in this study can be described according to general features of the essays (see Table 4), including: total words, total sentences, total

Table 4. Overall Frequencies of the Sample

Variable	Complete Sample Mean	Non-ELL Mean	ELL Mean	
Total Words	350.35	353.02	347.68	
	(156.17)	(149.64)	(163.92)	
Total Sentences	22.26	22.84	21.68	
	(12.13)	(11.32)	(12.99)	
Total Paragraphs	4.8	5.02	4.58	
	(2.45)	(2.51)	(2.38)	
Total Errors	41.15	29.30	53.00	*
	(27.93)	(17.98)	(31.08)	
Total T-Units	30.47	29.52	31.42	
	(14.38)	(12.96)	(15.75)	
Total Error-free t-units	10.94	13.82	8.06	*
	(9.01)	(10.03)	(6.83)	
Words per T-unit	11.8	12.23	11.37	§
	(2.38)	(2.56)	(2.12)	
T-units per Sentence	1.70	1.43	1.98	
	(1.71)	(0.73)	(2.29)	
Errors per T-unit	1.47	1.11	1.85	*
	(0.97)	(0.73)	(1.05)	
Error-free t-units per T-unit	0.33	0.44	0.23	*
	(0.22)	(0.22)	(0.15)	

* $p<.01$ § $p<.10$

Note: Standard deviations are provided in parentheses. This table provides measures of the mean length and composition of the entire 100-essay sample and the ELL and Non-ELL groups separately.

paragraphs, total errors, total t-units, total error-free t-units, words per t-unit, t-units per sentence, errors per t-unit, and error-free t-units per t-unit.

In addition, it is useful to describe the overall frequency of specific surface errors throughout the entire 100-essay sample (see Table 5). In general, we can see that mechanical errors, such as capitalization, punctuation, and spelling, were most frequent. On the other hand, verbal errors, such as subject-verb agreement and verb tense errors, were rarer.

This study revealed several critical differences among those essays written by ELL and non-ELL students as well as between those essays that scored high and those that did not. In general the differences were those anticipated. Namely, the ELL student essays typically had more surface errors than non-ELL student essays. Likewise, those essays that received a low score, not surprisingly, had more surface errors in general than those that received a high score. A closer examination of the means reveals the magnitude of the observed differences.

Table 5. Overall Error Frequencies of the Sample

Error Category	Error Type	Mean per Essay	Mean per T-unit
Mechanical	Spelling	10.67	0.39
		(11.41)	(0.45)
	Punctuation	6.92	0.25
		(5.18)	(0.17)
	Capitalization	6.8	0.25
		(10.52)	(0.39)
	Apostrophe	0.85	0.037
		(1.17)	(0.07)
	Total Mechanical	25.24	0.92
		(19.09)	(0.74)
Other Surface	Other Syntactic	4.77	0.18
		(3.88)	(0.15)
	Lexis	1.11	0.04
		(2.32)	(0.06)
	Preposition	1.10	0.04
		(1.46)	(0.06)
	Article	1.08	0.03
		(3.62)	(0.07)
	Pronoun Reference	0.28	0.01
		(0.55)	(0.03)
	Total Other Surface	8.34	0.29
		(8.54)	(0.23)
Sentence Boundary	Run-on	2.77	0.10
		(3.68)	(0.14)
	Comma Splice	1.08	0.04
		(1.70)	(0.05)
	Fragment	1.05	0.04
		(1.68)	(0.06)
	Total Sentence Boundary	4.90	0.18
		(4.26)	(0.16)
Verbal	Verb	1.10	0.04
		(1.71)	(0.05)
	Verb Tense	0.97	0.03
		(1.72)	(0.06)
	Subject-verb Agreement	0.61	0.02
		(1.29)	(0.05)
	Total Verbal	2.68	0.09
		(3.25)	(0.11)

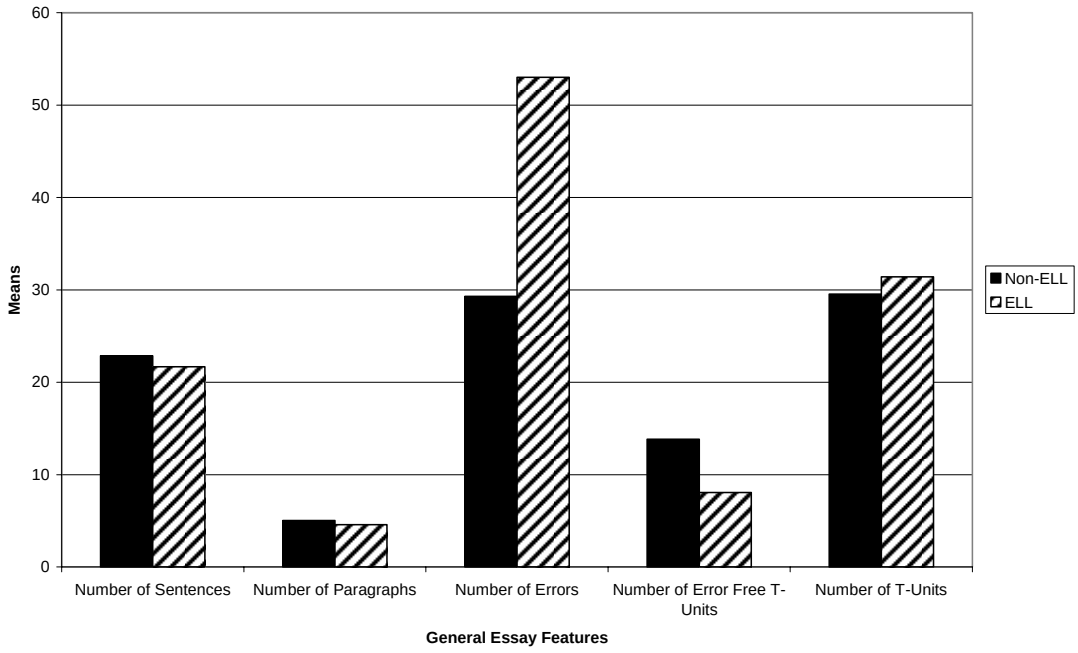
Note: Standard deviations are provided in parentheses. This table shows mean surface errors per t-unit for the entire 100-essay sample. The categories and sub-categories are arranged from most to least frequent.

Question #1: Differences between ELL and Non-ELL Essays

General Essay Features

In terms of overall length, there were no differences between ELL and non-ELL essays. The number of words, sentences, paragraphs, and t-units found in the essays of both groups were quite similar (see Table 4 and Figure 5). The number of words per t-unit was marginally different between ELL essays (Mean = 11.37, *SD* = 2.12, *F* (1, 96) = 3.402, *p* = .068) and non-ELL essays (Mean = 12.23, *SD* = 2.56).

Figure 5. General Essay Features of ELL and Non-ELL Essays



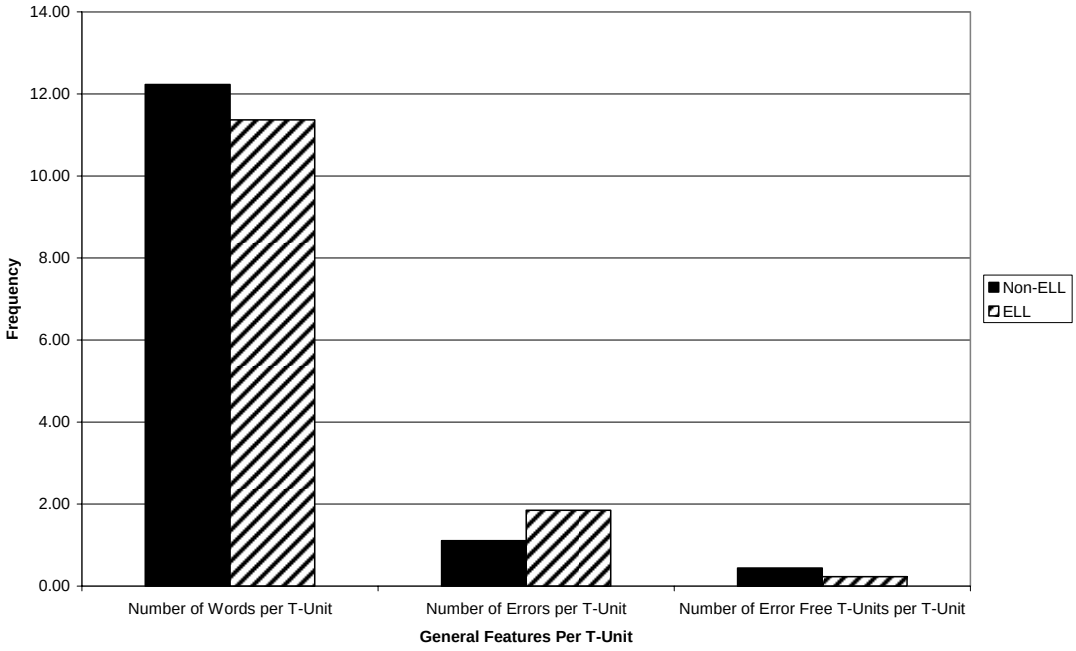
However, when looking at differences in surface errors, significant differences become apparent (see Figure 6). For example, ELL essays contained nearly twice as many errors (Mean = 53.00, SD = 31.08, F (1, 96) = 23.103, p = .000) as non-ELL essays (Mean = 29.52, SD = 12.96). Furthermore, the ELL essays contained more errors per t-unit (Mean = 1.85, SD = 1.05, F (1, 96) = 18.395, p = .000) than did their non-ELL counterparts (Mean = 1.11, SD = 0.73). The errors in the ELL student essays were also more dispersed throughout the entire essays. The number of error-free t-units for ELL student essays (Mean = 8.06, SD 6.83, F (1, 96) = 17.491, p = .000) was approximately half the number of error-free t-units in non-ELL essays (Mean = 13.82, SD = 10.03). The proportion held true for error-free t-units per total t-units (F (1, 96) = 34.203, p = 000) (see Figure 6).

Most of the observed differences between ELL and non-ELL essays are consistent with previous L2 writing research. As expected, ELL student essays had significantly more of certain surface errors per t-unit than non-ELL student essays, including spelling, preposition, punctuation, verb tense, and other syntactic errors (see Appendix A for means and standard deviations). These are some of the errors that English language learners commonly make when writing in English (Sheorey, 1986; Ferris, 1992; Schairer, 1992; Dordick, 1996; Porte, 1999).

Sentence Boundary Errors

An unexpected difference emerged for the sentence boundary errors per t-unit. That is, non-ELL essays had roughly seven times as many run-on errors (Mean =

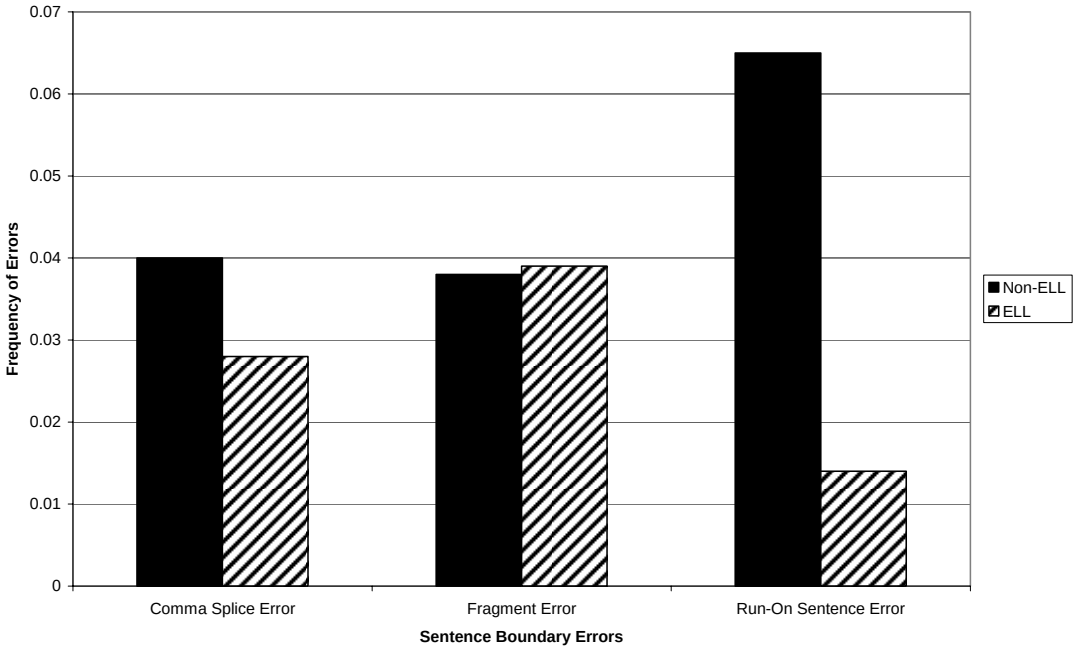
Figure 6. General Features Per T-Unit of ELL and Non-ELL Essays



0.07, SD = 0.11, F (1, 96) = 9.786, p = .002) as ELL essays (Mean 0.01, SD = 0.15). Comma splice and fragment errors were virtually equal for both groups (see Figure 7).

The fact that ELL writers have fewer run-on errors per t-unit than non-ELL writers does not necessarily mean ELL writers are better at end punctuation. They may, in fact, have difficulty marking the end of a sentence appropriately by creating sentence fragments or comma splices, errors that were coded separately. In fact, there is reason to expect native Spanish speakers to make comma splice errors. In Spanish, when two independent clauses are closely related, it is acceptable to divide them with

Figure 7. Sentence Boundary Errors in ELL and Non-ELL Essays



only a comma, much as we use semi-colons in English writing (Montaño-Harmon, 1991). Asian ELL students may also make comma splice errors although whether this is due to possible interference from the L1 or simply a by-product of acquiring English is not known. However, differences in the frequency of comma splice errors were not found.

Furthermore, a comparison of sentences with comma splices shows great similarity between ELL and non-ELL essays. ELL students produced comma splice errors, such as those below:

Dolls are for your models, they shouldn't be to large. (Essay 83; Asian)

Now add a peice sticking out to the audience, don't make it like the one you did before. (Essay 88; Hispanic)

start ripting off heads and peeling of ther skin, leave the heads and the skins in the bag and the body inside the bowl. (Essay 98; Asian)

Yet ELL students made no more comma splice errors per t-unit than non-ELL students, which suggests that English proficient students have just as much difficulty marking sentence boundaries as their ELL peers. The following examples of non-ELL student comma splice errors appear very similar to those made by ELL students above:

and you will not have to look for it, after that you write down your information on your poster board. (Essay 4; African American)

Next, start "brainstorming", think anything up you can on that particular topic and wrinng it down, this will help in your writing. (Essay 95; White)

Another contributing factor to this counterintuitive result may be that non-ELL students have command of more vocabulary and language and are thus more fluent writers. A by-product of writing fluency, or perhaps oral fluency, might be creating run-on sentences. ELL students, on the other hand, are more likely to be less fluent writers, which may tend to cause them to write short, simple sentences. This conclusion is supported by the general essay features found in this study, which indicate that non-ELL essays contain slightly more words per t-unit than ELL essays.

This measure has been used to demonstrate writing fluency (Polio, 2001). To illustrate their differences, here are some examples:

> *How you do this is very simple you put some feed in the pen and take the goat out of the pen you then let the goat loose and he will run back to the pen*. (Essay 6; White)

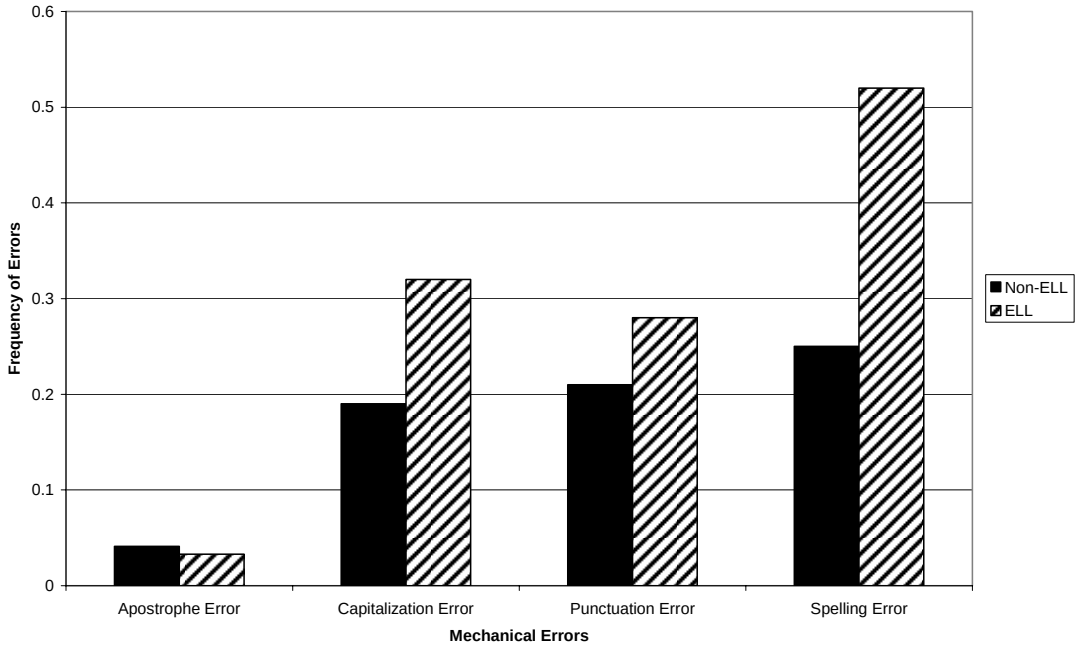
> *You need scissors and glue. You also need alumminium, and colors paint*. (Essay 12; Hispanic)

The excerpt from Essay 6 is written by a non-ELL student who is clearly not struggling for words and knows what he wants to convey. The excerpt from Essay 12 is written by an ELL student. It is evident that the ELL student has written two parallel simple sentences with correct end punctuation, rather than creating a more complex sentence.

Mechanical Errors

Contrary to the unexpected results in the sentence boundary category, the mechanical errors category yielded few surprises. In this category, ELL essays included more errors per t-unit on average than non-ELL essays in terms of spelling and punctuation. ELL essays contained twice as many spelling errors (Mean = 0.52, SD = 0.55, F (1, 96) = 9.956, p = .002) as non-ELL essays (Mean = 0.25, SD = 0.27). Also significantly, punctuation errors were more frequent by one fourth in ELL essays (Mean = 0.28, SD = 0.16, F (1, 96) = 5.863, p = .017) than in non-ELL essays (Mean = 0.21, SD = 0.17) (see Figure 8).

Figure 8. Mechanical Errors in ELL and Non-ELL Essays



The difference in spelling errors is to be expected because the graphophonemic system of English is not consistent as it is for the native language of most ELL students in this study, Spanish (Cronnell, 1985; Zutell & Allen, 1988). In addition to inconsistent sound-letter correspondence, English has many irregularly spelled words, called sight words, that are not only spelled irregularly but are also used most frequently. Words such as *their*, *where, were,* and *know* are examples of these words. ELL students who command these common words orally may rely on their native language (L1) graphophonemic system to write them (Ibrahim, 1978;

Bebout, 1985; Cook, 1997). Others may try to spell them phonetically using the English graphophonic system. In so doing, they will typically make spelling errors. This is easily seen in the kinds of errors made by ELL writers in this study below:

> *Main things that the art teacher told students were, introducing about this Project, Start doing things on cardboard, and things Students no* [know] *after finishing Painting.* (Essay 45; Asian)

> *Like if you are going to do the chak* [chalk] *experiment then you will need, two beackers,*[beakers,] *two cylinders, two piece of clack* [chalk] *(the same size), vinegar and a hot plate.* (Essay 29; Hispanic)

> *live* [leave] *it like that.* (Essay 93; Hispanic)

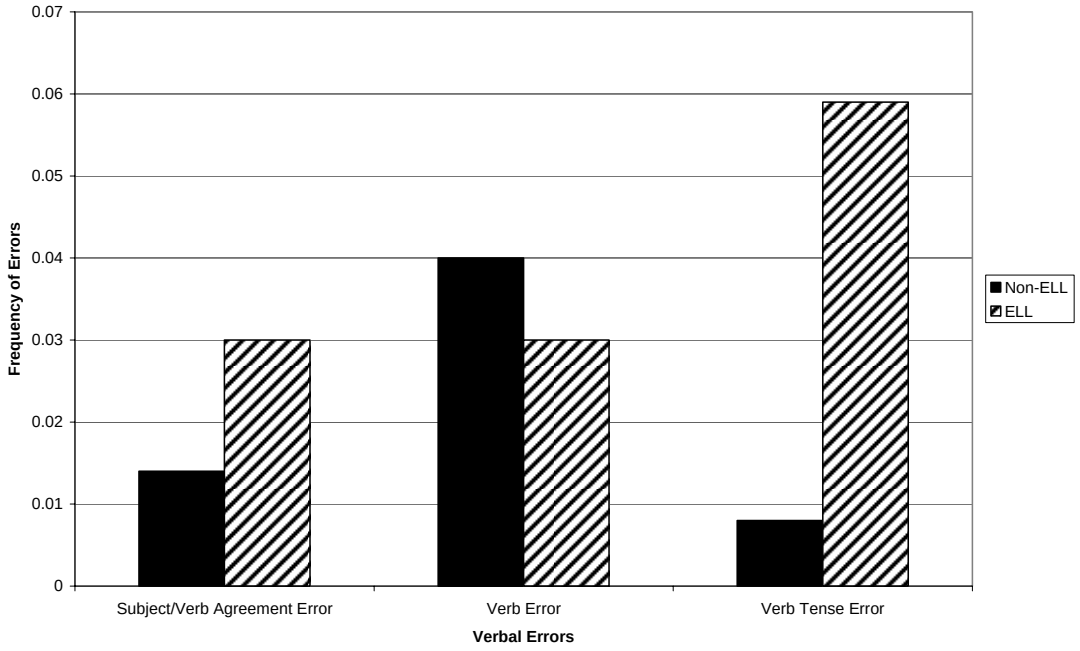
> *So we whent* [went] *back home* (Essay 5; Hispanic)

> *When I was in front of the all class I was nervos* [nervous] *because when somebody do not now inglish* [English] *and they have to talke* [talk] *inglish* [English] *is to hart* [hard] *to the person explaing* [explaining] *something or talk inglish* [English]. (Essay 54; Hispanic)

Verbal Errors

The results for verbal errors per t-unit revealed only one difference between ELL and non-ELL essays. ELL essays contained approximately six times as many verb tense errors (Mean = 0.06, SD = 0.07, F (1, 96) = 21.014, p = 000) as non-ELL essays (Mean = 0.01, SD = 0.03) (see Figure 9). The rate of subject-verb agreement and verb errors was essentially the same for ELL and non-ELL essays.

Figure 9. Verbal Errors in ELL and Non-ELL Essays



This is not surprising as verb tense errors are common errors for ELL students. Often English language learners use complex verb tenses incompletely; that is, they may leave out auxiliary verbs (Freeman & Freeman, 2001). Another reason ELL students make errors in verb tense is interference from their L1. For instance, in future time clauses, English uses the simple present tense, but in Spanish, the future tense is used. The following are several such examples found in this sample of essays:

> *I* [am] *just sending this letter to explaining how I did my project*. (Essay 65; Hispanic)

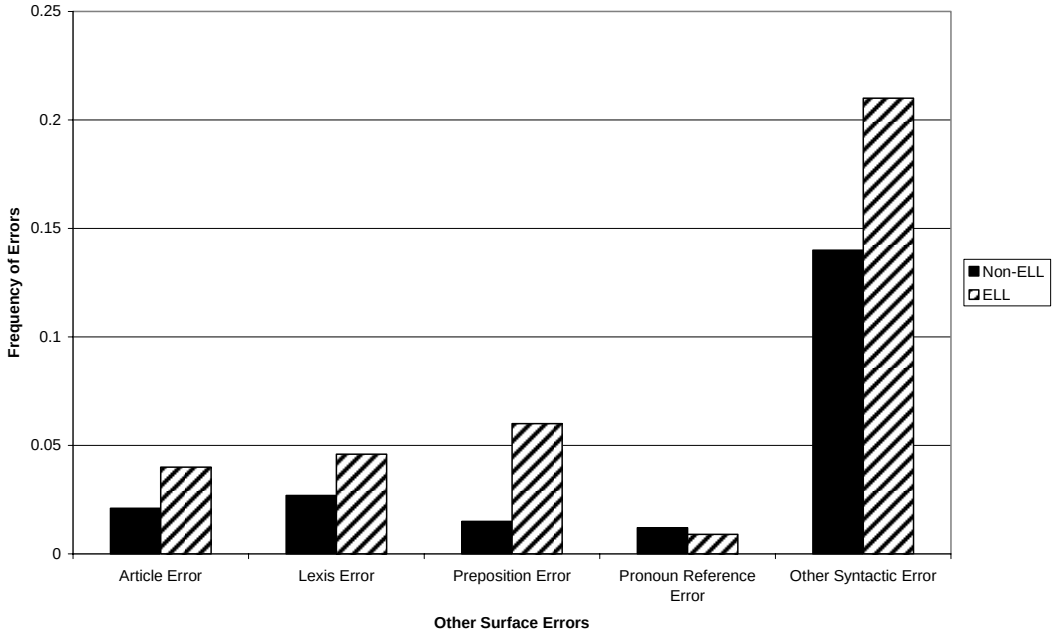
When you <u>will</u> finish of make these activities, take the ballon (Essay 12; Hispanic)

Other Surface Errors

The final category of errors, other surface errors, included errors that had to do with word use, word form, word order, semantics, and global sentence errors. In this category, ELL essays contained more of two kinds of error per t-unit than non-ELL essays (see Figure 10). ELL essays contained three times as many preposition errors (Mean = 0.06, SD = 0.07, F (1, 96) = 19.969, p = .000) as non-ELL essays (Mean = 0.02, SD = 0.03). With regard to other syntactic errors, ELL essays included one and one-third times as many errors per t-unit (Mean = 0.21, SD = 0.14, F (1, 96) = 6.619, p = .012) as non-ELL essays (Mean = 0.14, SD = 0.15).

These findings support previous research findings that preposition errors are a type of surface error that ELL students frequently make (Sheorey, 1986; Ferris, 1992; Shairer, 1992; Dordick, 1996; Porte, 1999). However, previous research has not really addressed why preposition errors are frequent among English language learners. One reason could be that English has several phrasal verbs that contain a verb and a preposition. The prepositions may have a meaning in the phrasal verb that does not necessarily match the meaning of the preposition alone. Other phrasal verbs have non-phrasal verb opposites, for example, *plug in* and *unplug*. In addition to general difficulties English language learners have with prepositions, there are some

Figure 10. Other Surface Errors in ELL and Non-ELL Essays



e

errors that were found in this sample of essays that are specific to Spanish speakers. In Spanish one preposition, *en*, is used in several situations, whereas in English three separate prepositions, *in, at,* or *on*, would be used in those same contexts (Cronnell, 1985). Furthermore, the rules for when to use *in, at,* or *on* are quite complicated. A Spanish speaker might fall prey to interference in the use of these particular prepositions and thus make more errors.

Nor are Spanish speakers the only ELLs who have difficulty acquiring conventional usage of prepositions. In fact, errors in the usage of prepositions are common for all learners of English as an additional language. Some examples below from this study sample show ELL students making just such errors:

lay it down <u>*in*</u> [on] *your table*. (Essay 87; Hispanic)

Then I put the three plant <u>*in*</u> [on] *a table that was aside of the window*. (Essay 79; Hispanic)

When that is done you take the beaker from the hot plate and turn of the hot plate and <u>*plug it out*</u> [unplug it]. (Essay 29; Hispanic)

After washed and rinsing dump all the water <u>*out from*</u> [out of] *the shrimps bowl.* (Essay 98; Asian)

Overall Differences between ELL and Non-ELL Essays

Apart from these errors of major interest, this study revealed that ELL essays contained more other syntactic errors per t-unit than non-ELL essays. Several of the error categories that previous research indicated were common among English language learners were compressed into the category other syntactic errors for this study due to low individual frequencies. Those errors include word order, it-deletion, and post-verb constructions. In addition, some global errors that impacted the entire sentence or even beyond the sentence were coded in this category. These are errors common to L2 writing in English (Dordick, 1996). Yet again, ELL students' writing may include interference errors from their L1, which would cause word order errors. In English noun clauses, question word order is not used although they may be introduced by a question pronoun, such as *how, what*, or *who.* One such example follows:

but also you must make a Figure of how much <u>*time did the chalk take to dissolve*</u> [time the chalk took to dissolve]. (Essay 29; Hispanic)

In addition, because Spanish, the L1 of most of the ELL students in this sample, is a pro-drop language, subject pronouns can be dropped, and the subject will be understood (Chomsky, 1988). Moreover, Spanish does not include a stative subject to describe situations as English does. For example, Spanish does not include the subjects *there* and *it* as in *There is/are…* or *it is*. Possible interference from Spanish can be seen in essays in this study below:

> [It] *is the same thing that Vegas make on the little bulbs turn on and off*. (Essay 42; Hispanic)
>
> *It's when the emotional part of the project start, because* [it] *is when you paint the ballon* (Essay 12; Hispanic)

Finally because ELL students can be beginning, intermediate, or advanced, they have differential levels of control of the basic sentence structure of English. Those who are at the beginning level of English writing may not be able to produce complete and comprehensible sentences. An example of this kind of other error follows:

> *The project it was about respiration system* (Essay 54; Hispanic)
>
> *the first think the I did it was when I put the titulo in the poster* (Essay 54; Hispanic)

Question #2: Differences between High-scoring and Low-scoring Essays

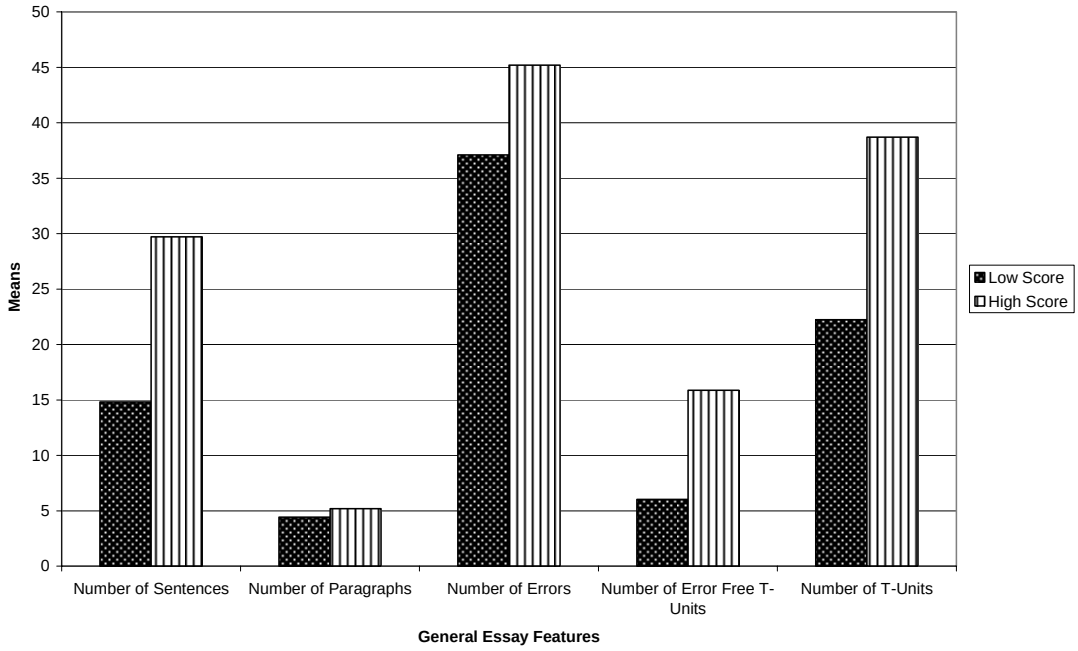
An analysis of the descriptive statistics for those essays that received a high score and those that received a low score reveals differences in overall length and

general features as well as in specific surface error types. As expected, essays that received a high score had significantly more words and t-units on average than those that received a low score. Further, the low-scoring essays had more errors per t-unit than the high-scoring essays. In addition to differences in general essay features, the low-scoring essays had more surface feature errors than the high-scoring essays in reference to punctuation, run on, spelling, and other syntactic errors (see Appendix A for means and standard deviations). These error types cut across ELL and non-ELL essays, indicating that they are problems for all students at this age.

General Essay Features

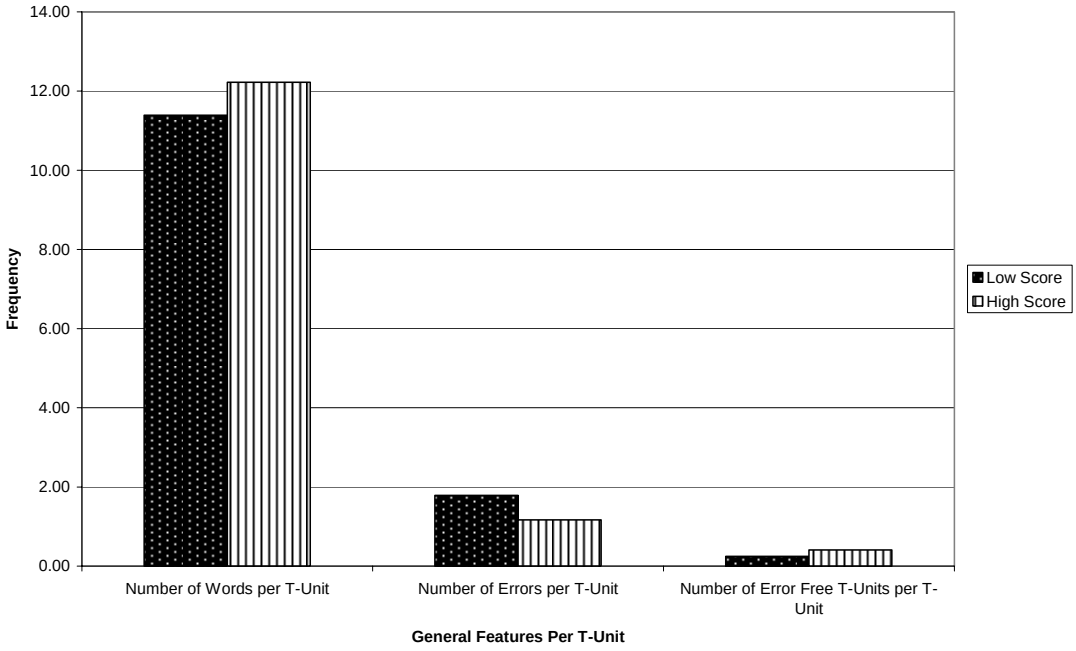
As shown in Figure 11, high-scoring essays had on average 457 (SD = 134.59) words overall, which is nearly twice as many as low-scoring essays (Mean = 243.7, SD = 89.16, F (1, 96) = 86.089, p = .000). Likewise, high-scoring essays had approximately twice as many sentences (Mean = 29.70, SD = 11.01, F (1, 96) = 59.273, p = .000) as low-scoring essays (Mean = 14.82, SD = 7.96). These differences translated into a slight difference in the number of words per t-unit (F (1,96) = 3.156, p = .079), with the high-scoring essays containing approximately one more word per t-unit than the low-scoring essays. Moreover, low-scoring essays contained about two-thirds the number of t-units as high-scoring essays (F (1, 96) = 47.956, p = 000).

Figure 11. General Essay Features of High- and Low-Scoring Essays



In addition to differences in length, when the measure of errors per t-unit is inspected, it shows that the low-scoring essays had a significantly greater number (Mean = 1.79, SD = 1.11, F (1, 96) = 12.900, p = .001) than high-scoring essays (Mean = 1.17, SD = 0.70). Furthermore, high-scoring essays contained less than half as many error-free t-units (Mean = 6.02, SD = 4.86, F (1, 96) = 51.046, p = 000) as low-scoring essays (Mean = 15.86, SD = 9.54). Similarly, about one quarter of the t-units in the low-scoring essays were error free, while more than two-fifths of the high-scoring t-units were error free (F (1, 96) = 20.363, p = .000) (see Figure 12).

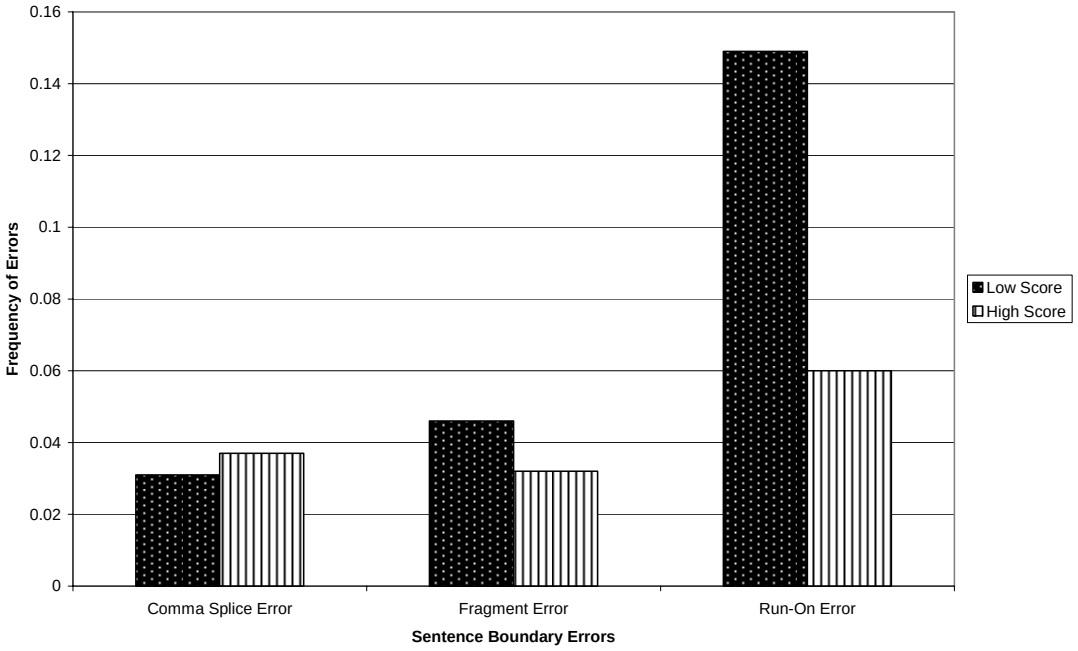
Figure 12. General Features Per T-Unit of High- and Low-Scoring Essays



Sentence Boundary Errors

The sentence boundary category of errors included some intuitive results. Namely, low-scoring essays had more than twice as many run-on errors per t-unit (Mean = 0.15, SD = 0.17) as high-scoring essays (Mean = 0.06, SD = 0.08, F (1, 96) = 12.441, p = .001) (see Figure 13). Interestingly, there were no differences in the frequency of comma splice and fragment errors per t-unit. That is, high-scoring essays contained the same proportion of these kinds of errors as low-scoring essays. This suggests that raters do not use the presence of these kinds of errors to distinguish poor writing from proficient writing.

Figure 13. Sentence Boundary Errors in High- and Low-Scoring Essays



The most significant result here is that the low-scoring essays contained far more run-on errors per t-unit than did the high-scoring essays. Again run-on errors most likely do not interfere with communication as in the following examples:

> *Get you square eggroll wrapping place it on top of the plate make sure the eggrool wrapping is facing vertical so It can look like a diamond*. (Essay 98; Asian)

> *Step 2 this project will take some time to complete you will need time to relax in between writing you have to prepare yourself.* (Essay 4; African American)

However, run-on errors may either correlate with overall poor writing and/or irritate raters (Hairston, 1981), which would explain why low-scoring essays

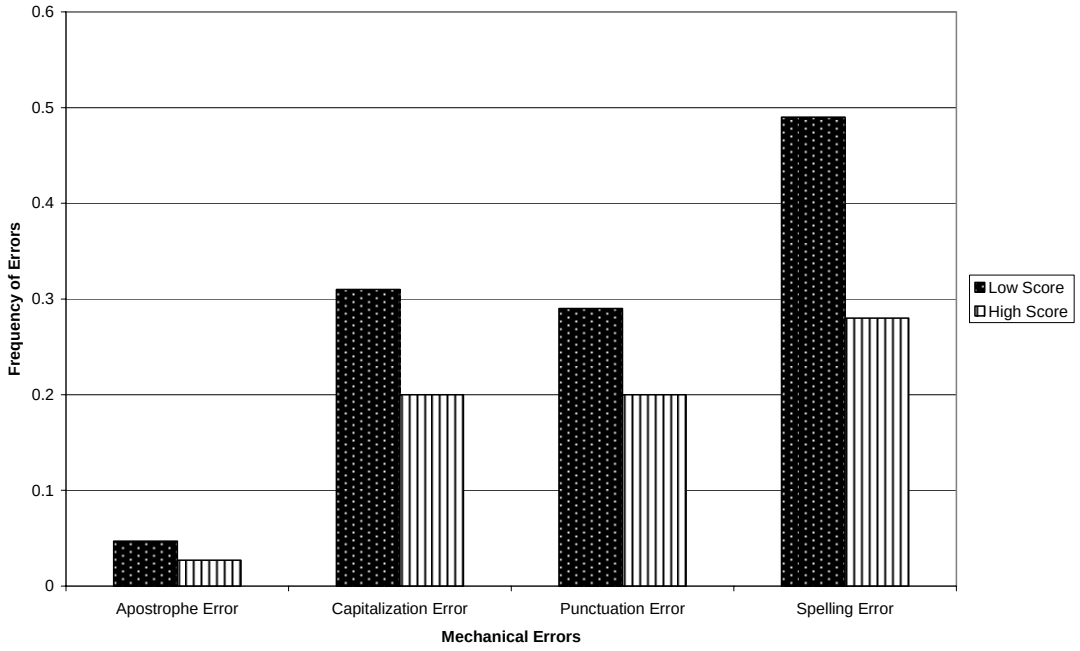
contained far more run-on errors than the high-scoring essays. By considering these results along with the differences found between ELL and non-ELL essays, it could reasonably be surmised that the non-ELL essays that included a great deal of run-on errors were most likely those that scored low.

Mechanical Errors

Not surprisingly within the category of mechanical errors, the essays that scored low contained more nearly twice as many spelling errors as the high-scoring essays (F (1, 96) = 6.316, p = .014) (see Figure 14). The low-scoring essays also contained about one third more punctuation errors than the high-scoring essays (F (1, 96) = 8.141, p = .005). High- and low-scoring essays contained basically the same number of apostrophe and capitalization errors per t-unit.

The spelling errors in several of the essays were often words that an 8th grade student should have command of the conventional spelling of, such as *sticks* and *does*. In addition, many of the spelling errors were those that may be particularly irritating to raters, such as *your* vs. *you're* or *there* vs. *their* (Hairston, 1981; Dordick, 1996). There are two possible explanations for the fact that essays with many spelling errors received low scores. One may be that the presence of these kinds of spelling errors unduly influenced the raters to score essays containing them

Figure 14. Mechanical Errors in High- and Low-Scoring Essays



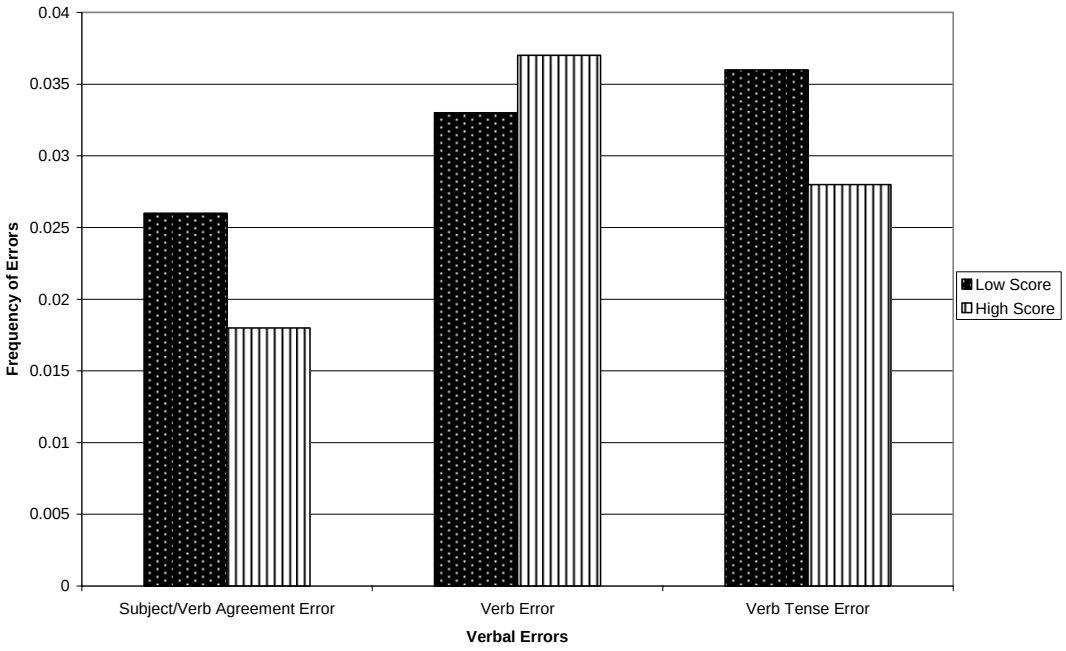
lower. The second may be that the presence of these kinds of errors correlates to other writing quality weaknesses, such as lack of development, lack of coherence, etc. In the absence of an interaction for spelling (discussed in the next section), one must assume that in either case, poor spelling is treated equally for ELL and non-ELL students.

As with spelling errors, frequent punctuation errors may either be an irritant or a feature that correlates to poor composing skills. Regardless of what may be the case, it is not surprising that the essays that received a low score had more punctuation errors than those that received a high score.

Verbal Errors

Figure 15 shows that in the category of verbal errors, the low-scoring essays did not differ significantly from the high-scoring essays in spite of apparent differences in the means. This indicates that raters either ignore these kinds of errors or do not emphasize accuracy in verb forms.

Figure 15. Verbal Errors in High- and Low-Scoring Essays

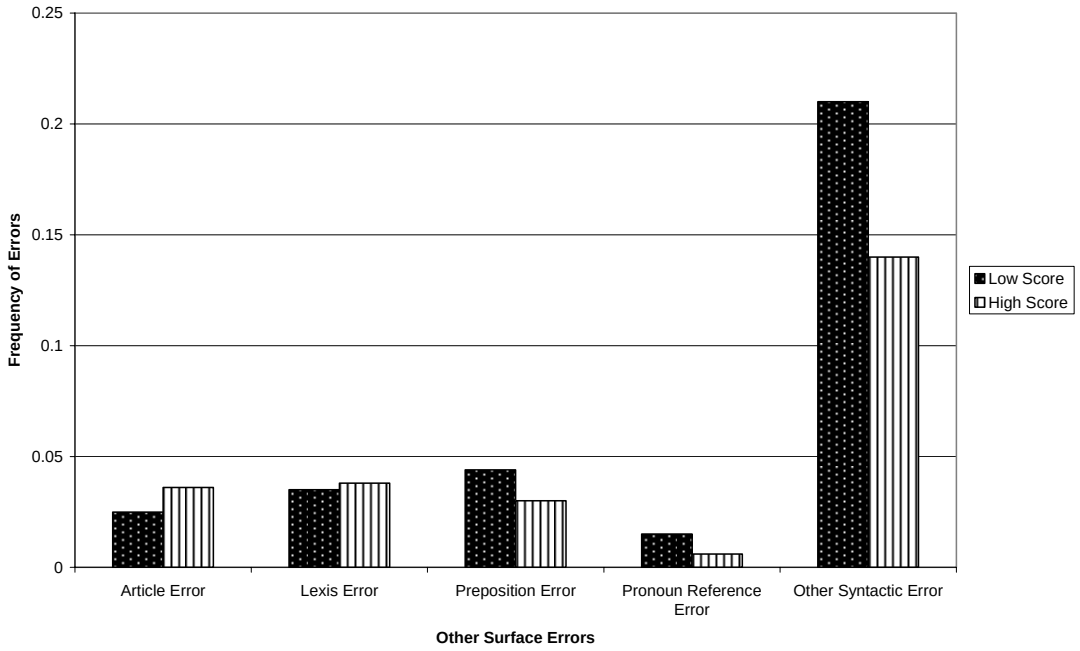


Other Surface Errors

The final category of errors, other surface errors, showed a difference only in terms of other syntactic errors. Low-scoring essays contained more other syntactic

errors (Mean = 0.21, *SD* =0.17, *F* (1, 96) = 6.503, *p* = .012) than high-scoring essays (Mean = 0.14, *SD* = 0.14) (see Figure 16).

Figure 16. Other Surface Errors in High- and Low-Scoring Essays



The only significant error type in this category was other syntactic errors, which included some more serious errors that impacted overall comprehensibility. Other syntactic errors might have included errors in word order, it-deletion, or other global errors. Some examples can be seen below:

> *Today I will tell you how to make a dull with corn life* [leaf as in husk or possibly "live corn" or fresh corn]. (Essay 93; Hispanic)
>
> *and you gat you a play house* (Essay 63; Hispanic)

The first example above was written by an ELL writer, while the example from essay 63 was found in a non-ELL essay. The presence of a large number of

such other syntactic errors per t-unit would greatly impact comprehensibility; therefore, it is no surprise that the essays containing more of these kinds of errors also received low scores.

Overall Differences between High- and Low-Scoring Essays

Given that ELLs make some surface errors more frequently than proficient English speakers, one must determine whether those are also errors that raters deem important in distinguishing low-scoring writing from high-scoring writing. The data suggest that raters distinguish between poor and proficient writing in terms of the frequency of punctuation, run-on, spelling, and other syntactic errors. Since all of these error sub-categories also differentiate ELL and non-ELL writing, it is difficult to establish whether ELL writing is simply poor writing or whether raters are treating ELL essays differently and perhaps unfairly. Nevertheless, ELL writing contains two additional error types, preposition and verb tense, that are not frequently found in non-ELL writing; furthermore, these sub-categories are not used by raters to differentiate poor from proficient writing. This fact allows for the possibility that, in fact, raters perceive differences between ELL and non-ELL writing and rate them differently.

Question #3: Interactions between ELL Status and High-scoring Status

If raters of high-stakes writing exams perceive a certain "foreignness" or difference between ELL and non-ELL writing, they may consciously or

subconsciously treat those essays differently. In other words, if raters encounter features in ELL essays that they expect to be present in 8th grade writing, they may ignore them or treat them as they normally would. On the other hand, if raters encounter features in ELL essays that they do not expect, they may be alerted to the fact that the author is not a proficient English speaker. If they then rate those essays differently, then bias exists. The only manner by which such a bias can be revealed is to seek any interactions between ELL status and score level.

The data revealed an interaction between ELL status and high-scoring status with regard to the number of paragraphs, F (1, 96) = 4.757, p = .032, the number of total errors, F (1, 96) = 5.215, p = .025, the number of error-free t-units, F (1, 96) = 5.001, p = .028, and the number of lexis errors per t-units, F (1, 96) = 7.500, p = .007 (see Appendix B for complete MANOVA results).

Paragraphs

As Table 6 demonstrates, among non-ELL essays, the average number of paragraphs was not significantly different for ELL and non-ELL essays. However, among the ELL essays, high-scoring essays contained significantly more paragraphs (Mean = 5.48, SD = 2.18, t = -2.86, p = .006) than the low-scoring essays (Mean = 3.68, SD = 2.27).

Table 6. Interaction for Paragraphs

	Non-ELL	ELL	Mean
High Score	4.88 (1.33)	5.48 (2.18	5.18
Low Score	5.16	3.68	4.42

	(3.33)	(2.27)
Mean	5.02	4.58

This suggests that the ELL essays that scored high resembled more closely non-ELL writing in terms of the number of paragraphs written. Conversely, the ELL essays that received a low score included significantly fewer paragraphs than any essays written by non-ELL students altogether.

This is most likely a reflection of formulaic writing instruction, i.e. the five-paragraph essay, which is a backwash effect of high-stakes writing exams (Cumming, 2001). The non-ELL essays, both those that scored high and those that scored low, had approximately five paragraphs. This suggests that the five-paragraph essay model has become inculcated in proficient English writers. Furthermore, the results suggest that raters might also favor this format for essays since the ELL essays that scored high had approximately five paragraphs also. However, the ELL essays that scored low had significantly fewer paragraphs, less than four on average. This hints that ELL students who can approximate non-ELL organization patterns may be able to overcome the potentially negative impact of the presence of a high number of surface errors.

Total Errors

The interaction with regard to total number of errors (see Table 7) showed that raters treat errors differently in ELL writing than in non-ELL writing. Furthermore, the results show that ELL essays that scored high actually had far more errors (Mean = 62.68, *SD* = 32.52, *t* = -2.297, *p* = .026) overall than ELL essays that scored low (Mean = 43.32, *SD* = 26.81), whereas among the non-ELL essays, there was no difference in the total number of errors. This difference in the sheer number of errors per essay among ELL student essays is probably due to the fact that ELL essays that scored high were longer overall than those that scored low (*t* = -7.033, *p* = .000).

Table 7. Interaction for Total Errors

	Non-ELL	ELL	Mean
High Score	27.72 (18.26)	62.68 (32.52)	45.20
Low Score	30.88 (17.93)	43.32 (26.81)	37.10
Mean	29.30	53.00	

Error-Free T-Units

In reference to error-free t-units, the interaction showed that the largest disparity in scores was for high-scoring essays (see Table 8). Those non-ELL essays that scored high had a far greater number of error-free t-units (Mean = 20.28, *SD* = 9.76, *t* = 3.669, *p* = .001) than those ELL essays that scored high (Mean = 11.44, *SD* = 7.07). However, among the low-scoring essays, there was no difference between non-ELL and ELL essays in the number of error-free t-units. This interaction

suggests that raters may actually overlook the high frequency and broad dispersal of errors in ELL essays.

Table 8. Interaction for Error-Free T-Units

	Non-ELL	ELL	Mean
High Score	20.28 (9.76)	11.44 (7.07)	15.86
Low Score	7.36 (4.81)	4.68 (4.63)	6.02
Mean	13.82	8.06	

Lexis Errors

In terms of lexis errors, high-scoring essays produced by ELL students had significantly more lexis errors per t-unit (Mean = 0.06, SD = 0.09, t = -2.490, p = .016) than those non-ELL student essays that scored high (Mean = 0.01, SD = 0.02). On the other hand, there was no significant difference found among the number of lexis errors per t-unit in ELL (Mean = 0.03, SD = 0.05) and non-ELL (Mean = 0.04, SD = 0.05) student essays that scored low (see Table 9).

Table 9. Interaction for Lexis Errors

	Non-ELL	ELL	Mean
High Score	0.01 (0.02)	0.06 (0.09)	0.035
Low Score	0.04 (0.05)	0.03 (0.05)	0.035
Mean	0.03	0.05	

This might suggest that raters rewarded attempts by ELL writers to include vocabulary they did not yet have full control over. It appears that ELL students' lexis errors were forgiven as long as other features of good writing were present, such as

acceptable paragraphing. Some of the lexis errors that were overlooked in the high-scoring ELL essays included some words that were incorrect, yet still comprehensible. Others were clearly cases of interference from the native language, or incorrect translations from the L1. Particular instances of these kinds of lexis errors can be seen in excerpts from ELL student essays that scored high below:

> *Remember <u>to put water</u>* [to water] *every single day and take notes* (Essay 65; Hispanic)
>
> *I readed many books about how do a project, it's very easy to <u>make</u>* [do]. (Essay 12; Hispanic)
>
> *I am sure you will get a good calification* [grade; Spanish *calificación*] (Essay 12; Hispanic)
>
> *After you had done that get a watch ang <u>take</u>* [keep] *track of time*. (Essay 29; Hispanic)

Pattern of ELL-Status X Score-Level Interactions

The interactions reveal surprising results. It seems that raters do, in fact, treat ELL and non-ELL essays differently, but it appears that in some cases these differences may benefit the ELL students. For example, raters appear to forgive lexis errors and may look beyond a high number of t-units containing errors in ELL essays, while they do not seem do so for non-ELL essays. On the other hand, ELLs appear to be at a disadvantage if they do not produce essays with acceptable paragraphing. These results indicate that whether in favor of ELL students or not, rater biases do exist in the scoring of the 8th grade TAAS writing exam.

Conclusions

The first research question asked in this study was: What is the nature of naturally occurring surface errors made by 8th grade ELL writers compared to those made by their proficient English-speaking peers on a high-stakes writing exam? It appears that ELL student writing at the 8th grade level indeed has more surface errors per t-unit than non-ELL student writing. In particular, ELL student essays had significantly more preposition, punctuation, spelling, verb tense, and other syntactic errors per t-unit than the essays of their English proficient counterparts. Other error types, while present in ELL essays, were not significantly different from non-ELL 8th grade essays. This may be due to the fact that many of the ELL students in Texas were born in the United States. That means that their Spanish itself may be anglicized. Thus, when they write in English, they may have internalized some features of English discourse and sentence structure.

The second research question posed in this study was: What is the nature of naturally occurring surface errors made by 8th grade writers who received a high score compared to those made by their peers who received a low score? In general, the high-scoring essays had significantly more words and more t-units than the low-scoring essays. Furthermore, the high-scoring essays had fewer errors per t-unit than did the low-scoring essays. These results confirm expectations from previous research. Specifically, the 8th grade TAAS essays that received low scores had more punctuation, run-on, spelling, and other syntactic errors than those essays that

received a high score. On the other hand, raters did not appear to use comma splice or fragment errors or errors in verb form to distinguish poor from proficient writing.

Finally, these results offer some answers to the final research question: Is there an interaction between superficial errors and ELL status in the scoring of 8th grade TAAS writing exams? The results of the MANOVA clearly show that there is a significant interaction for number of paragraphs and number of lexis errors per t-unit. Interactions found for the total number of errors and the number of error-free t-units are not as important because they are more a reflection of differences in length. That is, those interactions do not remain once they are divided by the total number of t-units. Raters apparently penalize ELL essays that do not conform to the five-paragraph format. Yet, surprisingly, raters appear to reward ELLs who experiment with lexical items even if they do not use them appropriately. However, this rewarding does not apply to non-ELL writing.

Chapter Five: Discussion

Introduction

The purpose of this study was to determine the extent to which naturally occurring surface errors in 8th grade ELL writing differ from those found in the writing of non-ELL students and how these errors impact holistic writing scores. In order to accomplish this goal, writing samples from 8th grade ELL and non-ELL students were analyzed to determine the nature of naturally occurring surface errors on a high-stakes writing assessment and to establish how and if they differ from other ELL populations that have been examined by previous studies. Then the essays of both ELL and non-ELL 8th graders were statistically analyzed to discover any potential bias due to the number and kind of surface errors they contained.

Findings of the study

The significance of this study begins with its value in describing what 8th grade ELL writing looks like and how it compares with non-ELL 8th grade writing. Looking at features, such as total words, total number of errors, and number of error-free t-units is important because previous research has shown that these features correlate with scores. In fact, measures that take into account the presence or absence of errors are particularly relevant in distinguishing poor from good quality writing

(Perkins, 1980). In this study, the number of total words, error-free t-units, and error-free t-units per total t-units together accounted for 67% of the variance in scores (see Appendix C). This indicates that examining errors in student writing is worthwhile and differences in error frequency and dispersal are meaningful.

Nonetheless, knowing that ELL and non-ELL and high-scoring and low-scoring essays do, in fact, differ in terms of surface errors is not enough evidence to claim a bias on the part of raters. It may well be that ELL writing is simply poorer overall and that surface errors and high frequencies of errors merely serve to distinguish good from poor writing. Yet the interactions found between ELL status and score status suggest otherwise for some dependent variables.

One surprise finding of this study is that non-ELL essays actually contained significantly more run-on errors per t-unit than the ELL essays. This is worth discussing because it is counter-intuitive. There are a variety of possible explanations for this finding. First, ELL students might have had difficulty with marking sentence boundaries appropriately, too, but they may have done so using a comma, which would have been coded as a comma splice error. Also, perhaps this difference in the number of run-on errors per t-unit is due to lack of oral fluency among the ELL writers, which may limit their written fluency. Another possible explanation is that ELL students avoid run-on sentences as a result of instruction. That is, perhaps teachers encourage them to write short, simple sentences that they have control over in order to avoid making sentence boundary errors.

Two of the most important findings in this study have to do with the impact that paragraphing and the use of vocabulary have on ELL student scores. The interactions found for these two variables indicate that raters may treat ELL student writing differently than proficient English student writing. Specifically, ELL student essays that contain significantly fewer than five paragraphs received low scores, while ELL student essays that contain approximately five paragraphs received high scores. This is important because the English proficient student essays in this sample, both those that scored well and those that scored poorly, contained approximately five paragraphs. Why there is such an extreme difference in the number of paragraphs among high- and low-scoring ELL essays is not clear. However, it could be that those with relatively fewer paragraphs were more recent immigrants and, therefore, perhaps had not yet been taught paragraphing skills. Regardless of why the number of paragraphs varies so dramatically, this finding suggests that raters have a bias in favor of the five-paragraph essay. ELL students who can approximate American-English proficient paragraphing appear to benefit in terms of their score.

Another interesting and unique finding of this study is the interaction with regard to lexis errors. ELL student essays that scored well had significantly more errors in lexis than did their non-ELL counterparts, while there was no significant difference in the number of lexis errors among all essays that received a low score. The relatively great number of lexis errors per t-units among the ELL essays that received a high score is difficult to explain. One possible explanation is that Spanish

writers take pride in using a broad and elaborate vocabulary. In fact, Spanish writing emphasizes the use of elaboration and advanced vocabulary (Montaño-Hartman, 1991). This may transfer to their writing in English.

The fact that despite the large number of lexis errors, these essays received a high score suggests one of two possible realities: raters forgave incorrect vocabulary usage among ELL students or raters rewarded ELL writers for attempting to use vocabulary that they did not have complete control over. In either case the critical point is that raters treated the errors in lexis made by ELL students differently than those made by non-ELL students. This might indicate a bias in favor of ELL students.

What makes this finding even more interesting is that Santos (1988) and Dordick (1996) found that native-speaking audiences found errors in lexis to be the most serious and to interfere with communication the most. Santos concluded that ESL teachers should focus on teaching vocabulary in order to prepare ELL students for a mainstream audience and in order to help them succeed in writing. Yet for 8th grade writing on a high-stakes assessment, it seems that lexical errors may not hurt ELL writers as much, and, in fact, using vocabulary when the writer is not completely in control of the lexical item may actually be rewarded, or at least forgiven, by raters.

Interactions for total number of errors and total number of error-free t-units, though present, cannot offer much explanation because once those measures were equalized by dividing by total number of t-units, the interaction disappeared. In other

words, it is most likely due to the fact that ELL essays that received a low score were particularly short compared to all other groups that these interactions were found.

Implications of the Findings

By including the writing of 8th grade students who are not ELLs, this study shows that the specific errors that are most serious in high-stakes writing assessments for this population are not necessarily those that are most serious among college-level students. In fact, it seems that errors that are expected by raters as typical of 8th grade students are not as serious as those that appear more "foreign" or unexpected. For example, apostrophe, article, capitalization, comma splice, fragment, lexis, subject-verb agreement, and verb errors, which have been found to be common among L2 writing in English are also common among all 8th grade writing. But the question that most teachers of ELL students want answered remains: how can I focus my instruction to maximize the benefit to my students?

One area for instructional intervention should be sentence combining activities. This kind of instruction will increase the length and complexity of ELL student writing, both of which correlate positively with writing quality. However, since run-on errors were found with significantly greater frequency among low-scoring essays, this instruction should help student avoid run-ons.

In addition, the lack of interactions for those errors that are especially frequent among 8th grade ELL but not non-ELL writing, such as preposition and verb tense errors, indicates that these errors do not cause any bias on the part of raters. Thus,

while teachers of ELL students may want to focus on these persistent errors, which are difficult for non-native English speakers from a variety of native language backgrounds to master, they are not the most important area of instruction focus.

In fact, these results clearly suggest that teachers should focus on teaching ELL students organization and paragraphing skills. The point is not necessarily to teach the five-paragraph essay format, but to teach ELLs American understanding of the purpose and form of paragraphs. This instruction should not be postponed until a certain level of proficiency is attained, but should be taught right away as an essential component of effective writing.

Furthermore, teachers should assist students in building rich vocabulary and encourage their ELL students to use vocabulary even when they are not sure they know how to spell it or if they are not using it completely accurately. In other words, teachers should encourage risk taking in the use of vocabulary among their ELL students from the very beginning. As long as other aspects of good writing are present in the essay, this risk-taking will not hurt and may actually help the student.

Ultimately, this study suggests that the state writing assessment encourages good writing processes, such as getting ideas on paper and editing for errors later. ELL students should be encouraged to include as many ideas as possible in an organized manner in their initial drafts and worry about correcting surface errors only secondarily. This is supported by previous research that demonstrated that during the editing and revising process, it is far more likely that ELL writers (Zamel, 1983) will

address surface level corrections rather than including more ideas. In other words, once a first draft is written, it is not likely that new ideas will be incorporated during the editing and revising process.

Beyond the instructional implications of this study, there are some implications for high-stakes writing assessment. Specifically, there are two options: 1) increase rater training to avoid bias due to ELL status or 2) acknowledge and even embrace differences in ELL writing. Historically, the former option has been the answer to biases in holistic writing assessment; however, since those biases still exist, it is clearly not the best option. There is evidence, though, that acknowledging and embracing differences in writing among certain sub-populations is an appropriate option. For instance, Smitherman (1993) found in her study of NAEP writing samples that the presence of certain African American Vernacular English features enhanced the overall writing and contributed positively to scores.

Suggestions for Future Research

This study examines ELL writing on a high-stakes assessment among school-age students. According to the review of literature, the lack of robust research in this area means this study in and of itself cannot completely answer all possible questions that teachers, parents, students, administrators, policy-makers, and researchers have. Thus, this research suggests several obvious follow-up studies. First, because the state writing assessment in the state of Texas recently changed from TAAS to TAKS and the analytical tool for holistic scoring also changed, this study should be

replicated using scores the essays would have received had they been rated using the TAKS analytical tool in order to determine whether or not these results would hold true.

Other areas touched on by this study that warrant a closer examination are the raters and their actual reactions to the writing samples of 8th grade ELL students while rating them. Future research should describe more fully who the raters of the Texas state writing assessment are, what their level of experience is with holistic scoring and with ELL students, whether or not they are themselves native speakers of English, etc. Furthermore, recently Cumming (2002) has introduced think aloud protocols as a means of describing rater behaviors. A think aloud protocol of raters while they are holistically rating student essays would reveal whether or not raters could identify which essays were written by ELL students and what their reaction to specific essay features would be. This would shed more light on the ultimate questions underlying this study: is ELL and non-ELL student writing treated equally by raters or do potential biases rise to the surface? Furthermore, do raters determine whether the author is an ELL or not, and are they conscious of forgiving or penalizing certain errors?

In order to answer these questions, future research might look not only at surface errors, but the presence of surface errors in conjunction with high- or poor-quality writing along other dimensions, such as organization, coherence, and content. This would allow analyses to show what portion of the final score is attributable to

surface errors. Two additional further analyses would be to consider the actual appearance of the essays and the coherence of the paragraphs. The appearance bias is well-documented and perhaps poor appearance coupled with a high rate of surface errors might lead to bias. Also, it is not clear if those essays that scored well did so because they fit a five-paragraph model or because they included coherent paragraphs.

This study has some limitations of which the researcher is aware. One of the limitations is that students within the non-ELL group may be former ELLs. The state of Texas has no way of indicating former ELL status in its assessment system. Some surface errors that have become fossilized may persist in the writing of former ELLs. Secondly, schools report only the ethnicity of students on the state assessment, not their native language group. While it is a strong likelihood that Hispanic ELLs speak Spanish natively, among the Asian ELLs it is not possible to determine whether the native language is Vietnamese, Korean, Chinese, or other. Future research in this area should take these limitations into account.

Conclusions

Research about how ELL student writing is treated by raters is crucial given the current political and demographic realities in the State of Texas and the nation. Studies like this one are important because there is a real lack of empirical data for educators to consult when making critical instructional decisions. ELL students are faced with many challenges as they attempt to simultaneously acquire English

language skills and navigate the high academic expectations of public schools. If there is any manner in which their teachers can focus instruction on areas that will impact their ultimate success in academic writing, then it is our duty to discover what those areas might be.

This study is significant because it examines a group of ELL students that have not been the subject of previous research. Furthermore, it reveals biases in holistic rating of ELL student writing that have real implications for the classroom instruction. The findings of this study, though not all-encompassing, provide a source of empirical data from which teachers can draw to inform their daily practice. It is essential that more research in the field of second language writing be focused on the school-age population who face serious consequences if they do not succeed on high-stakes assessments.

Appendix A

Overall Error Frequencies of the Sample

Error Category	Error Type	Mean per T-unit					
		Non-ELL	ELL		Low Score	High Score	
Mechanical	Spelling	0.250	0.520	*	0.490	0.280	**
		(0.27)	(0.55)		(0.57)	(0.26)	
	Punctuation	0.210	0.280	**	0.290	0.200	*
		(0.17)	(0.16)		(0.19)	(0.13)	
	Capitalization	0.190	0.320		0.310	0.200	
		(0.39)	(0.38)		(0.41)	(0.36)	
	Apostrophe	0.041	0.033		0.047	0.027	
		(0.058)	(0.087)		(0.096)	(0.039)	
Other Surface	Other Syntactic	0.140	0.210	**	0.210	0.140	**
		(0.15)	(0.14)		(0.17)	(0.14)	
	Lexis	0.027	0.046		0.035	0.038	
		(0.042)	(0.074)		(0.051)	(0.069)	
	Preposition	0.015	0.060	*	0.044	0.030	
		(0.029)	(0.067)		(0.072)	(0.033)	
	Article	0.021	0.040		0.025	0.036	
		(0.043)	(0.080)		(0.048)	0.086	
	Pronoun Reference	0.012	0.009		0.015	0.006	
		(0.033)	(0.020)		(0.037)	(0.012)	
Sentence Boundary	Run-on	0.065	0.014	*	0.149	0.060	*
		(0.11)	(0.15)		(0.168)	(0.084)	
	Comma Splice	0.040	0.028		0.031	0.037	
		(0.055)	(0.050)		(0.052)	(0.054)	
	Fragment	0.038	0.039		0.046	0.032	
		(0.070)	(0.056)		(0.070)	(0.054)	
Verbal	Verb	0.040	0.030		0.033	0.037	
		(0.055)	(0.051)		(0.053)	(0.054)	
	Verb Tense	0.008	0.059	*	0.036	0.028	
		(0.027)	(0.068)		(0.067)	(0.044)	
	Subject-verb Agreement	0.014	0.030		0.026	0.018	
		(0.037)	(0.058)		(0.057)	(0.040)	

*$p < .01$ **$p < .05$

Note: Standard deviations are provided in parentheses. This table includes all mean surface errors per t-unit for all four groups: Non-ELL, ELL, Low Score, and High Score. The categories and sub-categories are arranged from most to least frequent.

Appendix B

TESTS OF BETWEEN-SUBJECTS EFFECTS					
Source	**Dependent Variable**	**Type III Sum of Squares**	***df*** **Mean Square**	**F**	**Sig.**
Model	TOTAL WORDS	1146145.390	382048.463	28.916*	.000
	TOTAL PARAGRAPHS	46.320	15.440	2.716*	.049
	TOTAL ERRORS	18852.190	6284.063	10.339*	.000
	TOTAL T-UNITS	6919.790	2306.597	16.331*	.000
	WORDS PER T-UNIT	37.092	12.364	2.263	.086
	ERRORS PER T-UNIT	22.956	7.652	10.463*	.000
	APOSTROPHE ERRORS PER T-UNIT	1.234E-02	4.113E-03	.750	.525
	ARTICLE ERRORS PER T-UNIT	1.269E-02	4.231E-03	.869	.460
	CAPITALIZATION ERRORS PER T-UNIT	.743	.248	1.673	.178
	COMMA SPLICE ERRORS PER T-UNIT	6.292E-03	2.097E-03	.747	.527
	FRAGMENT ERRORS PER T-UNIT	5.443E-03	1.814E-03	.451	.717
	LEXIS ERRORS PER T-UNIT	2.782E-02	9.272E-03	2.665	.052
	PREPOSITION ERRORS PER T-UNIT	6.090E-02	2.030E-02	7.678*	.000
	PRONOUN REFERENCE ERRORS PER T-UNIT	2.727E-03	9.089E-04	1.211	.310
	PUNCTUATION REFERENCE ERRORS PER T-UNIT	.354	.118	4.725*	.004
	RUN-ON ERRORS PER T-UNIT	.388	.129	8.085*	.000
	SPELLING ERRORS PER T-UNIT	2.921	.974	5.471*	.002
	SUBJECT-VERB AGREEMENT ERRORS PER T-UNIT	8.566E-03	2.855E-03	1.177	.323
	VERB ERRORS PER T-UNIT	3.933E-03	1.311E-03	.454	.715
	VERB TENSE ERRORS PER T-UNIT	5.858E-02	1.953E-02	7.207*	.000
	OTHER SYNTACTIC ERRORS PER T-UNIT	.283	9.443E-02	4.970*	.003
ELL_STAT * PASSFAIL	TOTAL WORDS	8010.250	8010.250	.606	.438
	TOTAL PARAGRAPHS	27.040	27.040	4.757*	.032
	TOTAL ERRORS	3169.690	3169.690	5.215*	.025
	TOTAL T-UNITS	56.250	56.250	.398	.529
	WORDS PER T-UNIT	1.260	1.260	.231	.632

TESTS OF BETWEEN-SUBJECTS EFFECTS					
Source	**Dependent Variable**	**Type III Sum of Squares**	***df*** **Mean Square**	**F**	**Sig.**
	ERRORS PER T-UNIT	6.908E-02	6.908E-02	.094	.759
	APOSTROPHE ERRORS PER T-UNIT	7.114E-04	7.114E-04	.130	.719
	ARTICLE ERRORS PER T-UNIT	6.354E-04	6.354E-04	.131	.719
	CAPITALIZATION ERRORS PER T-UNIT	7.293E-02	7.293E-02	.493	.484
	COMMA SPLICE ERRORS PER T-UNIT	1.692E-03	1.692E-03	.602	.440
	FRAGMENT ERRORS PER T-UNIT	5.424E-04	5.424E-04	.135	.714
	LEXIS ERRORS PER T-UNIT	1.892E-02	1.892E-02	5.438*	.022
	PREPOSITION ERRORS PER T-UNIT	3.171E-03	3.171E-03	1.199	.276
	PRONOUN REFERENCE ERRORS PER T-UNIT	5.800E-04	5.800E-04	.773	.382
	PUNCTUATION REFERENCE ERRORS PER T-UNIT	4.295E-03	4.295E-03	.172	.679
	RUN-ON ERRORS PER T-UNIT	3.242E-02	3.242E-02	2.028	.158
	SPELLING ERRORS PER T-UNIT	2.504E-02	2.504E-02	.141	.708
	SUBJECT-VERB AGREEMENT ERRORS PER T-UNIT	8.818E-04	8.818E-04	.363	.548
	VERB ERRORS PER T-UNIT	9.660E-04	9.660E-04	.334	.564
	VERB TENSE ERRORS PER T-UNIT	6.143E-05	6.143E-05	.023	.881
	OTHER SYNTACTIC ERRORS PER T-UNIT	3.400E-02	3.400E-02	1.790	.184
	TOTAL WORDS	712.890	712.890	.054	.817
ELL_STAT					
	TOTAL PARAGRAPHS	4.840	4.840	.851	.358
	TOTAL ERRORS	14042.250	14042.250	23.103*	.000
	TOTAL T-UNITS	90.250	90.250	.639	.426
	WORDS PER T-UNIT	18.589	18.589	3.402	.068
	ERRORS PER T-UNIT	13.452	13.452	18.395*	.000
	APOSTROPHE ERRORS PER T-UNIT	1.889E-03	1.889E-03	.345	.559
	ARTICLE ERRORS PER T-UNIT	9.023E-03	9.023E-03	1.853	.177
	CAPITALIZATION ERRORS PER T-UNIT	.403	.403	2.726	.102
	COMMA SPLICE ERRORS PER T-UNIT	3.639E-03	3.639E-03	1.296	.258
	FRAGMENT ERRORS PER T-	2.054E-06	2.054E-06	.001	.982

TESTS OF BETWEEN-SUBJECTS EFFECTS					
Source	**Dependent Variable**	**Type III Sum of Squares**	***df*** **Mean Square**	**F**	**Sig.**
	UNIT				
	LEXIS ERRORS PER T-UNIT	8.774E-03	8.774E-03	2.522	.116
	PREPOSITION ERRORS PER T-UNIT	5.280E-02	5.280E-02	19.969*	.000
	PRONOUN REFERENCE ERRORS PER T-UNIT	1.379E-04	1.379E-04	.184	.669
	PUNCTUATION REFERENCE ERRORS PER T-UNIT	.147	.147	5.863*	.017
	RUN-ON ERRORS PER T-UNIT	.156	.156	9.786*	.002
	SPELLING ERRORS PER T-UNIT	1.772	1.772	9.956	.002
	SUBJECT-VERB AGREEMENT ERRORS PER T-UNIT	6.206E-03	6.206E-03	2.558	.113
	VERB ERRORS PER T-UNIT	2.527E-03	2.527E-03	.875	.352
	VERB TENSE ERRORS PER T-UNIT	5.694E-02	5.694E-02	21.014*	.000
	OTHER SYNTACTIC ERRORS PER T-UNIT	.126	.126	6.619*	.012
	TOTAL WORDS	1137422.250	1137422.250	86.089*	.000
PASSFAIL					
	TOTAL PARAGRAPHS	14.440	14.440	2.540	.114
	TOTAL ERRORS	1640.250	1640.250	2.699	.104
	TOTAL T-UNITS	6773.290	6773.290	47.956*	.000
	WORDS PER T-UNIT	17.243	17.243	3.156	.079
	ERRORS PER T-UNIT	9.434	9.434	12.900*	.001
	APOSTROPHE ERRORS PER T-UNIT	9.740E-03	9.740E-03	1.777	.186
	ARTICLE ERRORS PER T-UNIT	3.034E-03	3.034E-03	.623	.432
	CAPITALIZATION ERRORS PER T-UNIT	.266	.266	1.799	.183
	COMMA SPLICE ERRORS PER T-UNIT	9.613E-04	9.613E-04	.342	.560
	FRAGMENT ERRORS PER T-UNIT	4.898E-03	4.898E-03	1.217	.273
	LEXIS ERRORS PER T-UNIT	1.212E-04	1.212E-04	.035	.852
	PREPOSITION ERRORS PER T-UNIT	4.934E-03	4.934E-03	1.866	.175
	PRONOUN REFERENCE ERRORS PER T-UNIT	2.009E-03	2.009E-03	2.677	.105
	PUNCTUATION REFERENCE ERRORS PER T-UNIT	.203	.203	8.141*	.005
	RUN-ON ERRORS PER T-UNIT	.199	.199	12.441*	.001

TESTS OF BETWEEN-SUBJECTS EFFECTS						
Source	**Dependent Variable**	**Type III Sum of Squares**	***df***	**Mean Square**	**F**	**Sig.**
	SPELLING ERRORS PER T-UNIT	1.124		1.124	6.316*	.014
	SUBJECT-VERB AGREEMENT ERRORS PER T-UNIT	1.478E-03		1.478E-03	.609	.437
	VERB ERRORS PER hT-UNIT	4.391E-04		4.391E-04	.152	.697
	VERB TENSE ERRORS PER T-UNIT	1.579E-03		1.579E-03	.583	.447
	OTHER SYNTACTIC ERRORS PER T-UNIT	.124		.124	6.503*	.012
	TOTAL WORDS	1268365.360		13212.139		
Error			6			
	TOTAL PARAGRAPHS	545.680		5.684		
			6			
	TOTAL ERRORS	58350.560		607.818		
			6			
	TOTAL T-UNITS	13559.120		141.241		
			6			
	WORDS PER T-UNIT	524.550		5.464		
			6			
	ERRORS PER T-UNIT	70.207		.731		
			6			
	APOSTROPHE ERRORS PER T-UNIT	.526		5.482E-03		
			6			
	ARTICLE ERRORS PER T-UNIT	.467		4.869E-03		
			6			
	CAPITALIZATION ERRORS PER T-UNIT	14.205		.148		
			6			
	COMMA SPLICE ERRORS PER T-UNIT	.270		2.808E-03		
			6			
	FRAGMENT ERRORS PER T-UNIT	.386		4.026E-03		
			6			
	LEXIS ERRORS PER T-UNIT	.334		3.479E-03		
			6			
	PREPOSITION ERRORS PER T-UNIT	.254		2.644E-03		
			6			
	PRONOUN REFERENCE ERRORS PER T-UNIT	7.205E-02		7.505E-04		
			6			
	PUNCTUATION REFERENCE ERRORS PER T-UNIT	2.399		2.499E-02		
			6			
	RUN-ON ERRORS PER T-UNIT	1.535		1.599E-02		
			6			
	SPELLING ERRORS PER T-UNIT	17.086		.178		
			6			
	SUBJECT-VERB AGREEMENT ERRORS PER	.233		2.426E-03		
			6			

TESTS OF BETWEEN-SUBJECTS EFFECTS						
Source	**Dependent Variable**	**Type III Sum of Squares**	***df***	**Mean Square**	**F**	**Sig.**
	T-UNIT VERB ERRORS PER T-UNIT	.277	6	2.888E-03		
	VERB TENSE ERRORS PER T-UNIT	.260	6	2.710E-03		
	OTHER SYNTACTIC ERRORS PER T-UNIT	1.824	6	1.900E-02		
Total	TOTAL WORDS	2414510.750	9			
	TOTAL PARAGRAPHS	592.000	9			
	TOTAL ERRORS	77202.750	9			
	TOTAL T-UNITS	20478.910	9			
	WORDS PER T-UNIT	561.642	9			
	ERRORS PER T-UNIT	93.163	9			
	APOSTROPHE ERRORS PER T-UNIT	.539	9			
	ARTICLE ERRORS PER T-UNIT	.480	9			
	CAPITALIZATION ERRORS PER T-UNIT	14.948	9			
	COMMA SPLICE ERRORS PER T-UNIT	.276	9			
	FRAGMENT ERRORS PER T-UNIT	.392	9			
	LEXIS ERRORS PER T-UNIT	.362	9			
	PREPOSITION ERRORS PER T-UNIT	.315	9			
	PRONOUN REFERENCE ERRORS PER T-UNIT	7.478E-02	9			
	PUNCTUATION REFERENCE ERRORS PER T-UNIT	2.754	9			
	RUN-ON ERRORS PER T-UNIT	1.922	9			
	SPELLING ERRORS PER T-UNIT	20.007	9			
	SUBJECT-VERB AGREEMENT ERRORS PER T-UNIT	.241	9			
	VERB ERRORS PER T-UNIT	.281	9			

TESTS OF BETWEEN-SUBJECTS EFFECTS					
Source	**Dependent Variable**	**Type III Sum of Squares**	***df***	**Mean Square**	**F**	**Sig.**
	VERB TENSE ERRORS PER T-UNIT	.319	9			
	OTHER SYNTACTIC ERRORS PER T-UNIT	2.107	9			

Appendix C

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.806	.649	.646	.64
2	.816	.665	.658	.62
3	.824	.679	.669	.61

a Predictors: (Constant), WORDS
b Predictors: (Constant), WORDS, ERRFRETU
c Predictors: (Constant), WORDS, ERRFRETU, ERRFREE

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	*t*	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	*VIF*
1	(Constant)	.568	.157		3.625	.000		
	WORDS	5.513E-03	.000	.806	13.474	.000	1.000	.000
2	(Constant)	.461	.162		2.846	.005		
	WORDS	5.192E-03	.000	.759	12.094	.000	.877	.141
	ERRFRETU	.659	.309	.134	2.128	.036	.877	.141
3	(Constant)	.147	.223		.660	.511		
	WORDS	6.331E-03	.001	.925	8.975	.000	.315	.177
	ERRFRETU	1.694	.597	.344	2.838	.006	.228	.380
	ERRFREE	-3.932E-02	.019	-.332	-2.017	.047	.124	.080

a Dependent Variable: SCORE

References

Arthur, B., Farrar, D., & Bradford, G. (1974). Evaluation reactions of college students to dialect differences in the English of Mexican-Americans. *Language and Speech, 17*, 255-70.

Bebout, L. (1985). An error analysis of misspellings made by learners of English as a first and as a second language. *Journal of Psycholinguistic Research, 14,* 569-93.

Breland, H., & Jones, R. (1984). Perceptions of writing skills. *Written Communication, 1,* 101-19.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18,* 65-81.

Chomsky, N. (1988). *Language and problems of knowledge: The Managua lectures.* Cambridge, MA: The MIT Press.

Clair, N. (1994). Mainstream classroom teachers and ESL students. *TESOL Quarterly, 28,* 89-96.

Cook, V.J. (1997). L2 users and English spelling. *Journal of Multilingual and Multicultural Development, 18,* 474-88.

Cronnell, B. (1985). Language influences in the English writing of third- and sixth-grade Mexican-American students. *Journal of Education Research, 78,* 168-73.

Cumming, A. (2001). The difficulty of standards, for example in L2 writing. In T. Silva, & P.K. Matsuda, eds., *On Second Language Writing*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Cumming, A., Kantor, R., & Powers, D.E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86,* 67-96.

Cummins, J. (1980). The entry and exit fallacy in bilingual education. *NABE: The Journal for the National Association for Bilingual Education, 4,* 25-59.

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

Dordick, M. (1996). Testing for hierarchy of the communicative interference value of ESL errors. *System, 24,* 299-308.

Ferris, D.R. (1992). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly, 26,* 414-20.

Frazer, T. (1996). Chicano English and Spanish interference in the midwestern United States. *American Speech, 71,* 72-85.

Freedman, S. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology, 71,* 328-38.

Gaies, S. (1980). T-unit analysis in second language research: applications, problems and limitations. *TESOL Quarterly, 14,* 53-60.

Gal, S. & Irvine, J. (1995). The boundaries of languages and disciplines: How ideologies construct difference. *Social Research,* 967-1001.

Graham, S., Berninger, V.W., Abbott, R.D., Abbott, S.P., & Whitaker, D. (1997). Role of mechanics in composing of elementary school students: A new methodological approach. *Journal of Educational Psychology, 89,* 170-82.

Hairston, M. (1981). Not all errors are created equal: Nonacademic readers in the professions respond to lapses in usage. *College English, 43,* 794-806.

Hamp-Lyons, L. (2001). Fourth generation writing assessment. In *On Second Language Writing*, Silva, T. & Matsuda, P.K., Eds. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly, 4,* 759-62.

Harris, D.P. (1969). *Testing English as a second language.* New York: McGraw-Hill Book Company.

Haswell, R., & Whyche-Smith, S. (1994). Adventures into writing assessment. *College Composition and Communication, 24,* 220-6.

Hayes, J.R., Hatch, J.A., & Silk, C.M. (2000). Does holistic assessment predict writing performance? Estimating the consistency of student performance on holistically scored writing assignments. *Written Communication, 17,* 3-26.

Homburg, T.J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly, 18,* 87-107.

Hughes, A. & Lascaratou, C. (1982). Competing criteria for error gravity. *ELT Journal, 36,* 175-82.

Hunt, K.W. (1965). *Grammatical structures written at three grade levels.* Urbana, IL: National Council of Teachers of English.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41,* 201-13.

Ibrahim, M. (1978). Patterns in spelling errors. *English Language Teaching, 32,* 207-12.

Kameen, P. (1983). Syntactic skill and ESL writing quality. In *Learning to Write: First Language/Second Language,* Freedman, A., Pringle, I, & Yalden, J. Eds. New York: Longman.

Knoblauch, C.H. & Brannon, L. (1984). *Rhetorical traditions and the teaching of writing.* Upper Montclair, NJ: Boynton/Cook Publishers, Inc.

Lippi-Green, R. (1994). Accent, standard language ideology, and discriminatory pretext in the courts. *Language in Society, 23,* 163-98.

Ludwig, J. (1982). Native speaker judgments of second language learners' efforts at communication: A review. *Modern Language Journal, 66,* 274-83.

Magnan, S. (1983). Age and sensitivity to gender in French. *Studies in Second Language Acquisition, 5,* 194-212.

Marshall, J., & Powers, J. (1969). Writing neatness, composition errors, and essay grades. *Journal of Educational Measurement, 6,* 97-101.

Montaño-Hartman, M.R. (1991). Discourse features of written Mexican Spanish: Current research in contrastive rhetoric and its implications. *Hispania, 74,* 417-25.

Mosenthal, P.B., Davidson-Mosenthal, R.L., & Collela, A.K. (1987). Two determinants of teachers' holistic scoring: Prior knowledge and ideology type. *The Elementary School Journal, 88,* 39-49.

Nagy, W., McClure, E., & Mir, M. (1997). Linguistic transfer and the use of context by Spanish-English bilinguals. *Applied Psycholinguistics, 18,* 431-52.

Ovando, C., Collier, V., & Combs, M. (2003). *Bilingual & ESL classrooms: Teaching in multicultural contexts, 3rd edition.* Boston: McGraw Hill.

Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly, 17,* 651-71.

Polio, C. (2001). Research methodology in second language writing research: The case of text-based studies. In T. Silva, & P.K. Matsuda, eds., *On Second Language Writing*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Porte, G. (1999). Where to draw the red line: Error toleration of native and non-native EFL faculty. *Foreign Language Annals, 32,* 426-34.

Potts, M. & Gingerich, W. (1988). The Hispanic American in English composition classes. In D. Bixler- Marquez and J. Ornstein-Galicia (Eds.), *Chicano Speech in the Bilingual Classroom.* New York: Peter Lang.

Ruth, L., & Murphy, S. (1988). *Designing tasks for the assessment of writing.* Norwood, NJ: Ablex.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly, 22,* 69-90.

Schairer, K.E. (1992). Native speaker reaction to non-native speech. *Modern Language Journal, 76,* 309-19.

Sheorey, R. (1986). Error perception of native-speaking and non-native-speaking teachers of ESL. *ELT Journal, 40,* 306-12.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly, 27,* 657-77.

Simich-Dudgeon, C. (1989). *English literacy development: Approaches and strategies that work with limited English proficient children and adults.* Silver Spring, MD: The National Clearinghouse for Bilingual Education.

Sloan, C., & McGinnis, I. (1978). The effect of handwriting on teachers' grading of high school essays. Unpublished.

Smitherman, G. (1993). "The blacker the berry, the sweeter the juice": African American student writers and the national assessment of educational progress. Paper presented at the Annual Meeting of the National Council of Teachers of English (83rd, Pittsburgh, PA, November 17-22, 1993). EDRS 366944.

Sweedler-Brown, C. (1992). The effect of training on the appearance bias of holistic essay graders. *Journal of Research and Development in Education, 26,* 24-29.

Texas Education Agency (2004). 2004 Texas Association of Bilingual Education Conference Bilingual Directors' Meeting. Retrieved November 5, 2004, from www.tea.state.tx.us.

Texas Education Agency (2004). *2003-2004 Regional Home Language Report (Excluding English).* Austin, TX: Texas Education Agency.

Texas Education Agency (2002). *TEKS-based TAAS: Grade 8.* Austin, TX: Texas Education Agency.

Texas Education Agency (1999). *Grade 8 scoring guide for informative writing.* Austin, TX: Texas Education Agency.

Vann, R., Meyer, D., & Lorenz, F. (1984). Error gravity: A study of faculty opinion of ESL errors. *TESOL Quarterly, 18,* 427-40.

Wolfe, E., Kao, C., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication, 15,* 465-92.

Wolfram, W. (1991). *Dialects and American English.* Englewood Cliffs, NJ: Prentice Hall.

Wolfram, W., Adger, C., & Christian, D. (1999). *Dialects in Schools and Communities.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Zamel, V. (1983). The composing processes of advanced ESL students: Six case studies. *TESOL Quarterly, 17,* 165-71.

Zutell, J., & Allen, V. (1988). The English spelling strategies of Spanish-speaking bilingual children. *TESOL Quarterly, 22,* 333-40.

Vita

Jennifer Christa Holling was born in San Angelo, Texas, on September 16, 1971, the daughter of Larry Albert and Constance Geraldine Holling. After graduating from Seguin High School in 1989, she entered Southwest Texas State University in San Marcos, Texas. She received the degree of Bachelors of Arts in International Studies from Southwest Texas State University in 1993. She went on to receive a Master of Arts degree from The George Washington University in May 1995. After graduating from Southwest Texas State University, she worked for The American Nurses Association, Seguin Independent School District, Southwest Texas State University, Pflugerville Independent School District, The University of Texas at Austin, and Region XIII Education Service Center. From June 1996 until December 1998, Holling served as a U.S. Peace Corps Volunteer in Slovakia. In January 2000, she entered graduate school at The University of Texas at Austin.

Permanent Address: 400 Auxiliary Airport Road, Seguin, Texas 78155

This dissertation was typed by the author.