2018-12-01

The Effects of Teacher Background on How Teachers Assess Native-Like and Nonnative-Like Grammar Errors: An Eye-Tracking Study

Wesley Makoto Schramm
Brigham Young University

The Effects of Teacher Background on How Teachers Assess Native-Like and Nonnative-Like Grammar Errors: An Eye-Tracking Study

Wesley Makoto Schramm

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Grant Taylor Eckstein, Chair
Don William Chapman
David S. Eddington

Department of Linguistics

Brigham Young University

ABSTRACT

The Effects of Teacher Background on How Teachers Assess Native-Like and Nonnative-Like Grammar Errors: An Eye-Tracking Study

Wesley Makoto Schramm
Department of Linguistics, BYU
Master of Arts

Studies have shown that composition and L2 writing teachers give different scores (Golombek, Weigle, Boldt, & Valsecchi, 2003) and focus on different features (Brown, 1991) when assessing student writing, which is assumed to be due to the differences in their background and training (Santos, 1992; Atkinson & Ramanathan, 1995). Error gravity is thought to be one reason why composition and L2 writing teachers give different scores (Rifkin & Roberts, 1995). Common methods for examining error gravity were to analyze scores and responses given by the raters and to have raters reflect on the rating process and analyze their responses. Only one study had used eye-tracking methodology to explore the raters' reading behaviors (Eckstein, Briney, Chan & Blackwell, 2018). The current study built on Eckstein et al.'s study to examine how composition and L2 writing teachers rate grammar errors differently. The researchers identified three native-like errors and three nonnative-like errors and introduced them into eight paragraphs written by students in a first-year composition class. The researchers asked composition and L2 writing teachers to read and assess the eight paragraphs while an eye-tracker measured their eye-movements. We assume that what raters look at while assessing the paragraphs reflects what they are cognitively processing (Rayner, 1998). The results indicate that composition and L2 writing teachers assign significantly different scores to grammar (L2 writing teachers assign higher scores), yet their reading behaviors are similar. This indicates that teachers with different backgrounds do not process grammar errors differently, but rather reach different scores based on other differences.

Keywords: error gravity, composition, L2 writing, grammar, eye-tracking

Table of Contents

List of Tables

Introduction

The number of international students studying in United States universities has been steadily increasing for several decades now (Institute of International Education, 2017), making it more and more common for composition teachers to have mixed classes where both native and nonnative students are present (Harklou, 1994; Ferris, 2011). Furthermore, many of the international students have been taught English by L2 writing teachers at intensive English programs prior to attending university. In response to this phenomenon, researchers have investigated how composition and L2 writing teachers score nonnative writing (Golombek, Weigle, Boldt, & Valsecchi, 2003) and what features of writing the teachers tend to focus on (Brown, 1991). Such research has shown that composition teachers can be harsher than L2 writing teachers and that the teachers often focus on different features.

Researchers have tried to identify the reason for why composition and L2 writing teachers give different scores. They have examined areas such as teachers' attitudes towards nonnative students (Ferris, Brown, Liu, & Stine, 2011; Matsuda, Saenkhum, & Accardi, 2013; Shvidko, 2015), rubrics used for assessment (Song & Caruso, 1996; Barkaoui, 2010; Lindsey & Crusan, 2011), novice and expert teachers (Santos, 1988; Cumming, 1990; Song & Caruso, 1996), ethnolinguistic bias (Janopoulos, 1992; Rubin & Williams-James, 1997; Lindsey & Crusan, 2011), and error gravity, or what kinds of errors are considered more severe (Santos 1988; Brown, 1991; Sweedler-Brown, 1993; Song & Caruso, 1996; Elder, Golombek, Weigle, Boldt, & Valsecchi, 2003). It is believed that differences in these areas would be the manifestations of the ideological differences between the fields of composition and L2 writing (Santos, 1992; Atkinson & Ramanathan, 1995).

It is important to note that research has also pointed at student proficiency as a cause for different scores. Native and nonnative students' writing has been shown to be different in ways such as rhetorical structure (Connor, 2011), cohesive devices and organizational patterns (Leki, Cumming, & Silva, 2008), lexical variation (Crossley & McNamara, 2009), and the number of language errors (Eckstein & Ferris, 2017). However, the current study will focus on teachers' differences after taking these differences into consideration.

In reviewing 28 studies covering 20 years of error gravity research, Rifkin and Roberts (1995) stated that we "have only skimmed the surface" (p 513). Yet, there has been little research examining teachers' reading behaviors as a likely source of difference. Eye-tracking technology is an emerging technology in language research (Rayner, 1998; Conklin & Pellicer-Sanchez, 2016) that would be able to record data, such as where teachers are looking and for how long, which could be analyzed to see if there are differences in what composition and L2 writing teachers look at when scoring writing.

Literature Review

The following section will first explain some of the differences between composition and L2 writing teachers including their ideologies and how they view grammar. It will then summarize what is known about the different possible reasons why composition and L2 writing teachers give different scores. It will finish with an explanation of how eye-tracking can be used to examine reading behaviors.

Differences between composition and L2 writing teachers

The fields of composition and L2 writing split shortly after World War II when there was an increase in international students studying in U.S. universities and there was a demand for teachers who were trained in teaching nonnative students (Matsuda, 1999; Ferris, 2009). Following the split, composition has developed in a focus of rhetoric and creative writing whereas L2 writing has strong ties to applied linguistics (Santos, 1992).

Atkinson and Ramanathan (1995) identified a few of the cultural (beliefs and ideology) differences in their comparison between the composition program and the L2 writing program of a large university. In their comparison, Atkinson and Ramanathan took an ethnographic approach in investigating the cultural (i.e. beliefs, objectives, and practices) differences between the two programs. Major areas they differed were in the assumptions of students' cultural knowledge, the metagoals of the programs, and the form and content of the writing. In more detail, composition assumes a shared Western cultural knowledge among students (i.e. shared understanding what topics such as originality or critical thinking mean), goes beyond students' academic needs and encourages developmental writing, and encourages sophisticated communication (and discourages formulaic writing). On the other hand, L2 writing does not assume a shared Western cultural knowledge, focuses on practical skills that are useful for students' academic needs, and encourages clear and straightforward communication (often in the form of formulaic writing).

English composition teachers' view of grammar

In composition, grammar has been defined in a couple of different ways. Pelosi (1973) explained grammar in terms of "grammatical knowledge" and "grammatical description". Grammatical knowledge is a speaker's observable performance and grammatical description is the description of observable data. In other words, grammatical knowledge shows speakers'

competence in formulating grammatically correct sentences, often without being able to explain the prescribed rules, or grammatical description, associated with the formulated sentence. Other researchers seem to have reached similar conclusions in that this dichotomy is an important part of grammar (Lance; 1977; Hartwell, 1985). This definition of grammar seems appropriate for teachers who expect to teach native speakers and who need to be able to describe what students subconsciously know and can do with English (grammatical knowledge) and what the rules are (grammatical description).

Another way that researchers have defined grammar is in terms of its scope or boundaries. In other words, they try to define "what" and "how much" is included in grammar. Some of the areas that are included in grammar are syntax and morphology (Francis, 1953; Pelosi, 1973; Hartwell, 1985), appropriate usage or etiquette (Francis, 1954; Lance, 1977), and stylistic grammar (Hartwell, 1985). Areas that have been excluded from grammar include spelling (Pelosi, 1973), appropriate usage (Hartwell, 1985), and word choice (Pelosi, 1973). As can be seen, while most researchers agree that syntax and morphology are included in the definition of grammar, the other areas, especially appropriate usage and word choice, are controversial.

Because composition teachers have historically expected to teach students whose native language is English (Atkinson & Ramanathan, 1995), their primary concern for grammar was to investigate whether formal grammar teaching was effective or not (Hartwell, 1985). In 1963, Braddock, Lloyd-Jones, and Schoer boldly declared that:

> In view of the widespread agreement of research studies based upon many types of students and teachers, the conclusion can be stated in strong and unqualified terms: the teaching of formal grammar has a negligible or, because it usually displaces some

instruction and practice in composition, even a harmful effect on improvement in writing (p 37-38).

This statement has received great attention and has shaped composition dramatically (Kolln, 1981). In fact, though some researchers since 1963 have been optimistic for more attention on grammar (Pelosi, 1973; Rifkin & Roberts, 1995), the permeating attitude seems to be a dismissal of grammar instruction (Hartwell, 1985; Ferris, Eckstein, & DeHond, 2017) even when good control of grammar is expected of students (Harrington, Malencyzk, Peckham, Rhodes, & Yancy, 2001; Matsuda, 2012).

L2 writing teachers' views of grammar

L2 writing teachers have approached formal grammar instruction differently than composition teachers. One reason for this is because L2 writing teachers generally teach nonnative students who do not have the grammar intuition that native speakers have, and so grammar instruction is assumed to be helpful in developing students' competence (Ferris, 2009). This assumption can also be seen in the plethora of research devoted to L2 grammar instruction methodology (Borg, 1998). Though the effectiveness of grammar instruction has been the focus of some debate (Truscott, 1996; Truscott, 1999; Ferris, 1999), it has not been the main focus like it was for compositionists. Instead, a large portion of research has been targeted at error gravity of nonnative students' errors (Rifkin & Roberts, 1995).

L2 writing has adopted a much broader definition of grammar than composition has. It is often referred to using such terms as "language use errors" and "linguistic accuracy" (Barkaoui, 2010; Eckstein & Ferris, 2017). Even when the term "grammar" is used, an examination of the specific grammar analyzed reveal that areas such as "word choice", "spelling", "punctuation",

"redundancy", and "cohesion" are used (Santos, 1988; Sweedler-Brown, 1993). This indicates that L2 writing teachers use the term "grammar" to represent a large variety of language problems, even those not included in the compositionist's definition of grammar.

Using this broader definition of grammar, researchers have used error gravity research to examine the effects that these errors have on native readers. It is important to note that error is commonly defined in this area of research as language issues that interfere with comprehension, that are not considered acceptable or normal, or that irritate or distract a native reader (Santos, 1988; Rifkin & Roberts, 1995). Research in error gravity has investigated questions such as what are the composition teachers' attitudes towards nonnative students (Ferris, 2011; Matsuda, Saenkhum, & Accardi, 2013; Shvidko, 2015), which error types affect comprehension the most or are most irritating for composition and/or L2 writing teachers (Santos 1988; Brown, 1991; Sweedler-Brown, 1993; Song & Caruso, 1996; Elder, Golombek, Weigle, Boldt, & Valsecchi, 2003), the differences in holistic and analytic scores (Song & Caruso, 1996; Barkaoui, 2010; Lindsey & Crusan, 2011), the difference between novice and experienced raters (Santos, 1988; Cumming, 1990; Song & Caruso), and the effects of ethnolinguistic bias (Janopoulos, 1992; Rubin & Williams-James, 1997; Lindsey & Crusan, 2011).

Possible reasons for why teachers give different scores

Composition teachers' attitudes towards nonnative students in their classes seem to be mixed. First, in response to a survey question using a 6-point Likert Scale asking whether it was more challenging to teach nonnative students than native students, about two-thirds agreed to varying degrees, and the majority of the remaining one-third only somewhat disagreed (Shvidko, 2015). Examining another survey study, grammar seems to be a major point of confusion for

teachers when they interact with nonnative student writing. Ferris, Brown, Liu, and Stine (2011) surveyed composition and L2 writing teachers at eight sites concerning how the teachers responded to nonnative students. Of the composition teachers, responses were mixed, indicating confusion on how they should respond to L2 writing. Some teachers expressed negative attitudes towards nonnative students and little concern for their challenge with grammar. Others expressed a desire to help but a lack of knowledge of how to help. Many of the teachers reported that they changed their approach and focused more on grammar than they normally do. Other survey studies have shown similar results (Matsuda, 2013; Shvidko, 2015). These survey studies seem to indicate that for the teachers who want to help, helping nonnative students is another demand of their already busy schedules.

Research on rubric type seems to be more in agreement than research on error-type. Song and Caruso's (1996) research showed that when using a holistic rubric, composition teachers gave significantly higher scores than L2 writing teachers. However, when analytic rubrics were used, there were no differences between the teachers. Barkaoui (2010) used think-aloud protocol research design to investigate how L2 writing teachers use holistic and analytic essays differently. The research indicated that holistic rubrics lead teachers to pay more attention to the essay (focus of assessment) whereas analytic rubrics lead teachers to pay more attention to the rubric (source of assessment criteria). Finally, Lindsey and Crusan (2011) compared how teachers from a variety of disciplines (including composition teachers) use holistic and analytic rubrics when assessing nonnative writing. Their results indicate that the teachers assigned significantly higher scores when using a holistic rubric versus an analytic rubric. Overall, these studies seem to indicate that teachers score nonnative writing higher when holistic scores are used.

Research on teachers' experience seems to be more consistent as well. In his investigation, Santos (1988) showed that older teachers (i.e. more experienced) showed less irritation towards grammar errors than younger teachers (i.e. less experienced). Cumming (1990) asked novice and expert L2 writing teachers to assess 12 essays written by nonnative students with different proficiency levels, he then recorded the strategies that the teachers used in their assessment. The teachers' performance highlighted a number of strategic differences between the novice and expert teachers such as strategies to interpret different features and strategies to judge different features. Expert teachers used a wider variety of strategies in the assessment task. Song and Caruso (1996) also briefly covered the topic of teacher experience in their study. Their results agree with Santos' in that more experienced teachers tended to be more lenient than less experienced teachers, at least when holistic rubrics were used. These studies indicate that more experienced teachers use a wider variety of strategies to help them reach a decision, which may often be more lenient on grammar errors made by nonnative students.

In early ethnolinguistic bias research, researchers hypothesized that teachers may either find greater fault in nonnative students' writing than there is (Land & Whitley, 1989) or make extra allowances because of the difficulties that these students face (Janopoulos, 1992). To test these hypotheses, Janopoulos (1992) asked teachers from a variety of disciplines to rate isolated sentences that each contained an error. Each sentence was introduced to be written by either a native student or a nonnative student to different raters. In other words, some raters thought a native student had written the sentence and others thought that a nonnative did. The results indicated that teachers from the humanities department did not rate the sentences differently based on perceived student ethnolinguistic background. Rubin and Williams-James (1997) conducted similar research in which they had composition teachers rate writing samples that

were labeled as being written by either a U.S. student, a Danish student, or a Thai student. The writing samples had grammar errors implanted in them based on previous error gravity research. After the analysis, the composition teachers were shown to not be harsher on the nonnative students' writing samples; in fact, the Thai students received the highest scores. Lindsey and Crusan (2011) repeated Rubin and Williams-James' method but broadened the participants to include teachers from many disciplines and examined the scores these teachers assigned when using an analytic and a holistic rubric. Their results showed that writers believed to be nonnative English speakers often received lower scores than those believed to be native speakers of English when scored analytically but higher scores when scored holistically. These studies show a wide range of ways teachers can react to nonnative writing. Teachers may score more leniently or more harshly based on students' perceived ethnolinguistic background.

Error gravity research has focused mostly on comparing how composition and L2 writing teachers differed and on comparing how composition teachers differed when grading native and nonnative writing. Santos' (1988) research investigated the ratings of professors across many disciplines (including composition teachers) by asking them to rate four essays, two of which were written by nonnative students and the other two of which were partially revised versions of the other two essays. The researchers compared the scores given to the essays and found that the raters seemed to agree that lexical errors were the most serious errors and that raters gave significantly higher scores to language use than to content even though they considered the language use unacceptable. Brown (1991) had eight professors, who were either in the English department or ESL department, rate 112 compositions written by native or nonnative students at the end of their FYC class. The findings showed that there were no significant differences in the scores assigned to native and nonnative writing. However, a feature analysis showed that the

composition teachers tended to focus more on cohesion and syntax while the L2 writing teachers attended to organization. In yet another study, Sweedler-Brown (1993) collected six nonnative essays and prepared an original and a corrected (sentence-level errors) version of each for assessment. The results showed a correlation between sentence-level features and grammar/mechanics with overall score, but no correlation between rhetorical and organization features with overall score, indicating that sentence-level grammatical errors affected the overall scores of nonnative writing. A similar study (Song & Caruso, 1996) contradicted Sweedler-Brown's in that composition teachers seemed to give greater weight to content and rhetorical features than they did to language use errors. Finally, when investigating how composition teachers respond to field-specific text related writing rather than impromptu writing, Weigle, Boldt, and Valsecchi, (2003) found that composition teachers seemed to score the writing harsher than L2 writing teachers, and that composition teachers focused most on content and grammar while L2 writing teachers focused on a wider variety of features. The results of these five studies show little agreement. Areas of contradiction include whether teachers are more lenient or harsh because of grammar errors and what types of features teachers focus most on.

Eye-tracking

The methods that researchers have employed in order to examine the cognitive processes of the teachers in error gravity research can be grouped into two general methods. The first type of method is when researchers ask teachers to perform an assessment task and then analyze the product of the assessment task such as scores (Brown, 1991; Lindsey & Crusan, 2011) or teacher comments (Cumming, 1990; Rubin & Williams-James, 1997). The second type of method is when researchers ask teachers to narrate their own thought processes during the assessment task

(Barkaoui, 2010) or ask questions after the task for teachers to reflect on (Song & Caruso, 1996). Although these studies have been informative, they are rather limited in their generalizability. For example, the implications that can be drawn from scores assigned to writing samples cannot be verified. Also, there can be a mismatch between what teachers report they think and do and what they actually do (Montgomery & Baker, 2007).

Eye-tracking technology has been gaining attention in language research (Rayner, 1998; Conklin & Pellicer-Sanchez, 2016) possibly because it does not add an extra task for teachers to perform during the assessment (Paulson, Alexander, & Armstrong, 2007). That is, they can focus on the assessment without performing other tasks that may distract or divert attention away from the assessment.

Eye-tracking provides empirical data by measuring fixations (pauses in eye-movement when the eyes gather information) and saccades (quick eye-movement between fixations) (Huey, 1908/1968; Just & Carpenter, 1987), which are believed to reflect attentional focus (Conklin & Pellicer-Sancez, 2016; Rayner, 1998). Researchers believe that what readers give attention to represents the moment-to-moment cognitive processes of the reader (Rayner, 1997; Rayner 1998). Especially useful is the finding that readers fixate on problem areas, areas hard to understand, more frequently and for longer durations than other areas (Frazier & Rayner, 1982). When it comes to grammar, this means that the longer a teacher looks at a grammar error, the longer he or she is assumed to be processing the error.

The data that an eye-tracker collects can be classified into early and late reading measurements. Early measures generally reflect automatic word recognition and lexical access while late measures generally show strategic comprehension processes (Conkin, Pellicer-Sanchez, & Carrol, 2018). That is, on the first reading, readers are recognizing words and

making general sense of the words, and on subsequent readings, readers are trying to integrate and comprehend the text. Therefore, differences in early reading measures would indicate difficulty in recognizing the words and differences in late reading measures would indicate difficulty in integrating and comprehending the text.

A recent preliminary study using an eye-tracker (Eckstein, Briney, Chan, & Blackwell, 2018) shows great promise for enriching error gravity research. The researchers were able to indicate differences between L2 writing and composition teachers in how they look at four categories commonly found on writing rubrics (rhetoric, organization, word choice, and grammar). The researchers asked five L2 writing and five composition teachers to read a single essay written by a nonnative student and give a holistic score for the essay. Specifically for grammar, the results showed that composition teachers spent more time on the first read and L2 writing teachers spent more time in the later stages of reading. However, because it was a preliminary study, its' results are difficult to generalize. First of all, a total of ten participants, with five of each teacher-type, is too few to represent the teacher populations. Second, the researchers do not define what they mean by grammar error. We do not know what types of grammar error are represented in the essay they used, how the different teachers reacted to the specific types of grammar, or even how they identified what constitutes a grammar error. Finally, the researchers only employed one writing sample. We cannot tell whether the results are representative of nonnative writing or if it is sensitive to this one piece of writing. Though promising, this study is indeed preliminary. Further research is necessary to reap the benefits of eye-tracking technology in our efforts to gain a better understanding of how teachers assess grammar. The current study will build upon the preliminary study in an attempt to address these questions:

1. How do the scores assigned to writing differ between composition and L2 writing teachers?
2. How do the reading measurements of composition and L2 writing teachers compare when they rate grammar in nonnative students' writing?
3. How do the reading measurements of composition teachers compare when assessing native and nonnative writing?
4. How do the reading measurements of L2 writing teachers compare when assessing native and nonnative writing?

We expect that teacher background will have an effect on how long the teachers look at different errors. Specifically, we think that teachers will look at errors that they are not accustomed to longer than those that they are accustomed to during late reading measures. For example, composition teachers see native-like errors more often than nonnative-like errors, and so they will have longer total dwell times and more regression-in counts for nonnative-like errors. We believe the opposite will be true for L2 writing teachers. Because they are accustomed to nonnative-like errors, we believe they will have longer late reading measures for native-like errors. Furthermore, we think that decoding and word recognition will take the same amount of time regardless of teacher background because the text needs to be decoded first before teachers can notice grammar errors. Therefore, we expect early reading measurements to be similar between composition and L2 writing teachers.

Methodology

Participants

A total of 29 teachers participated in the study, with 15 being composition teachers and 14 being L2 writing teachers. The average age was 33 years old and the average experience was 7 semesters of teaching writing.

Among the composition teachers, five were male and ten were female. The average age for composition teachers was 34 years old and average experience was 11 semesters of teaching writing. All composition teachers had graduated or were currently pursuing a Master's degree in an English related field (most common being rhetoric). The teachers reported receiving departmental training at the start of every semester and attending weekly workshops with a group of fellow teachers to discuss any concerns related to their class.

Of the 14 L2 writing teachers, one was male and 13 were female. The average age was 33 years old and average experience was four semesters. There were three L2 writing teachers who were pursuing a bachelor's degree but had completed a teaching practicum course. The rest had or were pursuing a Master's degree in TESOL. The L2 writing teachers reported receiving two hours of training at the start of the semester targeted at writing for each semester that they taught a writing class. They also received training and calibration on writing assessment at least once per semester.

Passages

The researcher collected authentic student essays which were written in a mainstream FYC class at a four-year university in the Western United States. The essays were written during the first week of the semester and asked students to describe themselves as a writer after they had

reflected on one piece of writing that they had produced within the past year. The expected length of the essay was between 500 to 750 words.

The researcher selected eight paragraphs from the essays to use in the research. Of the eight paragraphs, four were written by native students and four by nonnative students. The paragraphs were selected so that those with any information that could easily identify the writer as a nonnative student were discarded. Also, only the introduction paragraph from each respective essay was used so that they would not differ too much in terms of content and organization with each other.

Once the paragraphs were selected, the researcher counted the number of errors each contained in terms of the six error types determined for this study (misused commas, vague pronoun reference, capitalization, determiners, pronouns, and singular/plural forms), and then introduced new errors into the paragraphs so that there was a total of 12 errors of each error type dispersed throughout the paragraphs. The dispersion was controlled so that each paragraph had nine errors, though the quantity of specific types of errors was not equal between each paragraph. The paragraphs were not changed in any way other than the introduction of these errors; other errors such as spelling were not corrected, and wording was not changed more than was necessary to create these errors.

In order to verify that each error the researcher identified was indeed an error, the researcher asked a group of L2 writing teachers to read each paragraph and identify errors they found. Any error that was not identified during this activity was revised. The researcher then asked a linguistics professor and a composition professor to review the errors and state whether they considered it to be an error that would affect their judgement of the quality of the text. Revisions were again made and checked with the two professors again.

The passages are included in Appendix A.

Types of Errors

The grammar errors used in this study have been organized into errors most commonly made by native students and errors most commonly made by nonnative students. The errors were selected by referring to Lunsford and Lunsford's (2008) analysis of the 20 most common types of errors made by native students, and Company's (2012) analysis of the 15 most common types of errors made by nonnative students. The error types were compared between these two lists so that no error was found in the top ten of both lists (see Table 1). This was done to ensure that the error type could be easily assigned to either group with minimal confusion. The three most common types of errors that do not appear on the other were selected from each list. Only three types were selected from each list so that the quantity would be manageable.

The three error types chosen to represent those commonly made by native students are missing commas (#2), vague pronoun reference (#4), and capitalization (#8). The three error types chosen to represent those commonly made by nonnative students are determiners (#3), prepositions (#4), and singular/plural forms (#5). Spelling and wrong word were both in the top five on both lists and were disqualified for use. Also, Table 2 shows how the errors were dispersed throughout the eight paragraphs.

Table 1 – Most common errors

	native-like errors (Lunsford & Lunsford, 2008)	non-native-like errors (Company, 2012)
1	wrong word	spelling
2	**missing comma**	word choice
3	incomplete documentation	**determiner**
4	**vague pronoun reference**	**preposition**
5	spelling	**singular/plural forms**
6	quotation	word form
7	unnecessary comma	punctuation
8	**capitalization**	subject-verb agreement
9	missing word	verb form
10	faulty sentence structure	verb tense

The following are examples of sentences with each type of error:

- missing comma - Prior to starting an essay I fill my head with so many different ideas.
- vague pronoun reference - By the end of it, the reader might not even remember what the thesis of the essay was and clueless as to what message I was trying to convey.
- capitalization - The best words to describe myself as a writer would be "Reluctant" and "Insecure."
- determiner - I struggle tremendously with organizing my thoughts and displaying them into the words.
- preposition - This realist notion is very apparent on my past essay.
- singular/plural - I confuse many word and have a hard time remembering simple grammar rules.

Table 2 - Number of errors in each paragraph

		capitalization	determiner	missing comma	preposition	singular plural	vague pronoun	Total
paragraph	1	0	2	1	1	2	3	9
	2	1	1	1	1	2	3	9
	3	3	2	2	0	1	1	9
	4	2	1	0	2	2	2	9
	5	2	2	1	2	1	1	9
	6	0	3	2	2	1	1	9
	7	3	0	2	2	2	0	9
	8	1	1	3	2	1	1	9
Total		12	12	12	12	12	12	72

Rubric

The rubric used by the participants when assessing the paragraphs was adapted from the one used by Connor-Linton and Polio (2014) which is itself an adaptation of Jacobs, Zinkgrap, Wormuth, Hartfiel, and Hughey's (1981) rubric. This rubric was used because it has been recently validated (Polio, 2013). Changes made in the adaptation include some changes to the organization section because my study uses paragraphs instead of full essays, and the combination of the language use and mechanics section into one section I labeled as grammar to better reflect the focus of this study. The rubric consists of four categories, which are content, organization, vocabulary, and grammar. Each category could be scored from a 0 to a 7. The rubric is included in Appendix B.

Apparatus

The machine used in this study was an SR Research EyeLink 1000 Plus (spatial resolution of 0.01°) which sampled at 1000 Hz. This eye-tracker required participants to rest their head in a mounted headrest to ensure accurate measurements. A computer screen with a display resolution of 1600 x 900 (approximately 3.5 characters subtended 1° of visual angle) displayed the paragraphs and rubric and was positioned 63 centimeters from the participants.

Areas of Interest

The areas of interest represent the areas of the text that the eye-tracking software collects data for. For example, first fixation duration is the duration of the first fixation that occurs in a certain area of interest. Areas of interest were established for each error in the eight paragraphs by encompassing the word that the error is found in. In the case of a comma, which isn't part of a word, the area of interest encompassed the word preceding where the comma should occur. Bigger areas of interest would be too ambiguous and smaller areas of interest may not capture fixations on the word.

Measurements

The following reading measurements will be used to answer the research questions. Definitions were taken from an eye-tracking manual designed for second language research (Conklin, Pellicer-Sanchez, & Carrol, 2018).

- First fixation duration – The duration of the first fixation on the word.
- First run dwell time – The total duration of all fixations on the word during the first read.
- Skip count – An indication of whether there were any fixations on the word during the first read.
- Second fixation duration – The duration of the second fixation on the word.
- Second run dwell time – The total duration of all fixations on the word during the second read.
- Total dwell time – The total duration of the fixations on the word.

- Regression-in count – The number of times the reader came back to the word from somewhere after the word.
- Run count – The total number of times the reader looked at the word and left.
- Fixation count – The total number of fixations on the word.

The following are the early reading measures and what reading process they indicate (Conklin, Pellicer-Sanchez, & Carrol, 2016).

- First fixation duration – decoding, word recognition
- First run dwell time – word recognition, general understanding of text
- Skip count – estimation of text predictability, skimming

The following are the late reading measures and what reading process they indicate (Conklin, Pellicer-Sanchez, & Carrol, 2016).

- Second fixation duration – word integration, syntactic processing
- Second run dwell time – word integration, syntactic processing
- Total dwell time – late word processing, syntactic processing
- Regression-in count – confusion, syntax ambiguity
- Run count – rereading, text integration
- Fixation count – rereading, text integration

Exit-Interview

The exit-interview was used to gather more information on what participants were thinking during the procedure. There was a total of six questions and the questions were concerned with matters such as what was the general approach, how did they look at the grammar in the paragraphs, and which errors were the most distracting. The information gained

from the exit-interview is used to supplement the eye-tracking data. The exit-interview is included in Appendix C.

Procedure

At the beginning of each session, the participant was seated at the eye-tracking apparatus in front of the computer screen used for displaying the paragraphs. The participant was asked to rest his or her head in the headrest and to not move his or her head until the end of the eye-tracking portion of the session. The researcher then performed a calibration and validation so that the camera could accurately measure the participant's eye-movements. A calibration and validation were performed after every third paragraph to ensure accurate measurements throughout the session.

After the calibration was completed, the procedure was introduced to the participant during a practice paragraph. The practice paragraph was a paragraph of similar length to the other eight paragraphs that was prepared so that the participant could get accustomed to the procedures prior to the study. During the practice paragraph, the participant was instructed to read the paragraph and prepare for scoring the paragraph while the paragraph was displayed on the computer screen. When ready, the display changed from the paragraph to the rubric, and the participant was instructed to verbally give the scores for each of the four subcategories as well as an overall score. When the researcher confirmed that the participant understood the procedure, the participant was given control of the pacing and completed the assessment of the remaining eight paragraphs. The ordering of the remaining eight paragraphs was randomized so that each participant read the paragraphs in a different sequence. After the participant was finished with the paragraphs, he or she was asked to complete an exit-interview.

Data analyses

The current study provided two types of data. First, the teachers provided five assigned scores (content, organization, vocabulary, grammar, and overall) for each paragraph. The scores ranged from 0 to 7. Second, the eye-tracker provided eye-measurement data in the form of the early and late reading measurements discussed above.

The assigned scores were analyzed in two ways. First, a descriptive mean was calculated for each teacher to give a general picture of how the teachers scored the paragraphs. Then, a mixed-effects analysis was performed to see if there was an interaction between composition and L2 writing teachers and the five scores (content, organization, vocabulary, grammar, and overall) they gave to the paragraphs they rated.

In order to analyze the reading measurement data (first fixation duration, first run dwell time, skip count, second fixation duration, second run dwell time, total dwell time, regression-in count, run count, and fixation count), the data was organized in a few ways. In order to compare composition and L2 writing teachers, a subset of the data only including measurements for nonnative-like errors was created. For comparing how composition teachers look at native-like and nonnative-like errors, a subset of only composition teachers' measurements was taken. A similar subset was also taken in order to compare how L2 writing teachers look at the errors. In other words, the data for L2 writing teachers was not included in the analysis for composition teachers and vice versa. Skip count is nominal data while the remaining eight reading measurements are interval data. A mixed-effects analysis was done for each reading measure in each of these subsets, totaling 27 different mixed-effects analyses. The random effects for the mixed-effects analysis are which teacher was performing the experiment and which paragraph

the error was in. A mixed-effects analysis was performed because the data consists of repeated measures, which means that we needed to take into account individual idiosyncrasies. The natural log of the dwell time measurements (first fixation duration, first run dwell time, second fixation duration, second run dwell time, and total dwell time) were used in order to normalize the data so the assumptions of a mixed-effects analysis would be met (Whelan, 2008).

Results

The assigned scores will be given first, which are concerned with the first research question. The reading measurements will then be given in three sections that each reflect one of the remaining three research questions

Differences in assigned scores

We wanted to see how the scores assigned by composition and L2 writing teachers compared. In completing the research task, each participant gave five scores for each paragraph (content, organization, vocabulary, grammar, and overall). Table 3 shows the average scores that each participant assigned to the eight paragraphs. On average, L2 writing teachers had a higher score for each of the five scores assigned, with their grammar scores (4.7) being 1-point higher than composition teachers' scores (3.7).

We then performed a mixed-effects analysis to see if the difference in scores was statistically significant. The analysis showed that there is an interaction between teacher background and the category scored ($F(29)= 5.949$, $p< .001$). L2 writing teachers gave significantly higher scores for grammar and vocabulary scores (Table 4). Also, the overall scores are not significant by standard procedure, but since $p= .052$ is very close, we will treat it as

significant as well. Of the three categories that are significant, grammar has the largest mean difference (-1.016). Table 4 summarizes the results of the mixed-effects analysis.

Eye-tracking measurements of nonnative-like grammar errors

We wanted to see how composition and L2 writing teachers compare when they look at nonnative-like grammar errors. After performing a mixed-effects analysis for each of the reading measurements in the subset of data only including nonnative-like errors, we found that none of the measurements showed a significant difference between how composition and L2 writing teachers look at nonnative-like errors (see Table 5 for details). However, total dwell time was close to significant (p= .08). On average, L2 writing teachers spent 66.5 milliseconds longer looking at nonnative-like errors than composition teachers.

Table 3 - Average Scores each teacher assigned

Composition Teachers	Content	Organization	Vocabulary	Grammar	Overall
1	5.0	4.9	3.1	2.8	4.3
2	4.0	4.1	5.0	3.3	4.3
3	5.4	5.5	4.6	5.1	5.1
4	5.8	5.0	5.6	5.1	5.3
5	4.6	4.6	4.1	4.4	4.1
6	5.5	5.3	4.6	4.8	5.0
7	3.3	3.6	3.3	2.5	3.3
8	5.4	4.9	4.9	4.8	5.3
9	4.5	4.1	4.3	3.4	3.8
10	3.9	4.4	4.3	3.6	3.9
11	5.3	4.8	4.4	4.0	4.8
12	3.3	2.8	3.1	2.1	2.8
13	4.3	4.3	3.9	3.6	4.0
14	3.8	3.5	3.6	2.5	3.4
15	5.3	4.6	4.1	3.5	4.5
Composition Average	4.6	4.4	4.2	3.7	4.2
L2 Writing Teachers	Content	Organization	Vocabulary	Grammar	Overall
1	4.3	4.4	4.9	4.3	4.5
2	5.9	6.1	5.9	5.1	5.8
3	4.1	4.1	4.5	4.0	4.3
4	5.6	5.5	5.8	5.9	5.6
5	4.3	3.6	4.0	4.0	3.8
6	4.5	4.1	3.6	3.8	3.8
7	3.1	4.3	4.5	4.5	4.1
8	5.3	4.8	4.9	4.1	4.9
9	5.4	5.5	5.5	6.0	5.6
10	5.5	4.8	5.0	4.4	4.9
11	4.3	4.6	4.3	4.9	4.6
12	4.5	4.3	5.0	5.4	4.9
13	5.9	5.6	5.4	5.3	5.5
14	5.9	5.6	5.4	5.3	5.5
15	4.9	4.6	4.3	4.3	4.8
L2 Writing Average	4.9	4.8	4.9	4.7	4.8

Table 4 - Average scores by category given by composition and L2 writing teachers

Category	Teacher type	Mean	Mean Difference	Std. Error	df	Sig.
Content	COMP	4.59	-0.224	0.277	40.568	0.424
	ESL	4.81				
Organization	COMP	4.41	-0.327	0.277	40.568	0.246
	ESL	4.73				
Vocabulary	COMP	4.18	-0.632	0.277	40.568	0.028
	ESL	4.81				
Grammar	COMP	3.68	-1.016	0.277	40.568	0.001
	ESL	4.70				
Overall	COMP	4.22	-0.555	0.277	40.568	0.052
	ESL	4.78				

Table 5 - Comparison of composition and L2 writing teachers when looking at nonnative like errors

		Mean	Mean	Mean Difference			
	TEACHER	(log)	ms or count	(COMP-L2) in ms	Std. Error	df	Sig.
First Fixation Duration	COMP	5.33	205.6	-1.24	0.06	26.42	0.92
	L2	5.33	206.9				
First Run Dwell Time	COMP	5.41	223.2	-7.95	0.06	25.85	0.54
	L2	5.44	231.1				
Skip Count	COMP		0.545	-0.09	0.06	27.14	0.14
	L2		0.632		0.06	27.14	0.14
Second Fixation Duration	COMP	5.38	215.9	3.64	0.06	27.94	0.76
	L2	5.36	212.3				
Second Run Dwell Time	COMP	5.49	242.3	7.86	0.07	32.13	0.64
	L2	5.46	234.4				
Total Dwell Time	COMP	5.92	373.2	-66.50	0.09	27.17	0.08
	L2	6.09	439.7				
Regression-in Count	COMP		0.44	-0.15	0.10	27.18	0.15
	L2		0.59				
Run Count	COMP		1.47	-0.27	0.24	27.07	0.27
	L2		1.74				
Fixation Count	COMP		1.69	-0.34	0.30	27.07	0.27
	L2		2.03				

Composition teachers' reading behavior

In this section we will present the results that are related to how composition teachers look at native-like and nonnative-like errors. A mixed-effects analysis was done for each of the reading measurements in the subset dealing with only composition teachers. First fixation duration (p= .034) and first run dwell time (p= .001) were found to be significantly different between native-like and nonnative-like errors. First fixation duration measures how long the teacher looked at the word the first time they looked at it, and first run dwell time is the total time the teachers looked at the word during the first reading. For both measurements, composition teachers looked at native-like errors longer than non-native like errors. Regression-in count was also close to being significant (p= .09), indicating that composition teachers look back at nonnative-like errors more often than at native-like errors.

Table 6 - Comparison of how composition teachers look at native-like and nonnative-like errors

		Mean	Mean	Mean Difference			
	ERROR	(log)	ms or count	(COMP-L2) in ms	Std. Error	df	Sig.
First Fixation Duration	native	5.39	218.5	13.55	0.03	776.8	0.03
	non-native	5.32	205.0				
First Run Dwell Time	native	5.52	250.6	27.90	0.04	783.9	0.00
	non-native	5.41	222.7				
Skip Count	native		0.549	0.00	0.03	1015.3	0.93
	non-native		0.546				
Second Fixation Duration	native	5.33	207.1	-10.18	0.04	432.2	0.24
	non-native	5.38	217.2				
Second Run Dwell Time	native	5.44	231.1	-11.12	0.05	400.9	0.31
	non-native	5.49	242.3				
Total Dwell Time	native	5.99	397.8	22.04	0.05	787.3	0.24
	non-native	5.93	375.8				
Regression-in Count	native		0.37	-0.08	0.05	1034.2	0.09
	non-native		0.45				
Run Count	native		1.35	-0.12	0.08	1036.4	0.10
	non-native		1.47				
Fixation Count	native		1.59	-0.10	0.10	1036.9	0.30
	non-native		1.69				

L2 writing teachers' reading behaviors

Finally, we wanted to see how L2 writing teachers looked at native-like and nonnative-like errors. To do so, a mixed-effects analysis was run for each of the reading measurements in the subset dealing with only L2 writing teachers. In this subset, skip count (p= .01) and regression-in count (p< .001) were significant. L2 writing teachers tended to skip native-like errors on first pass more often and look back at nonnative-like errors more often. First run dwell time (p= .09) and second fixation duration (p= .06) are close to significant as well. L2 writing teachers spend more time looking at native-like errors on the first reading, but then spend longer on nonnative-like errors on the second reading.

Table 7 - Comparison of how L2 writing teachers look at native-like and nonnative-like errors

		Mean	Mean	Mean Difference			
	ERROR	(log)	ms or count	(COMP-L2) in ms	Std. Error	df	Sig.
First Fixation Duration	native	5.37	215.3	8.23	0.03	757.0	0.23
	non-native	5.33	207.1				
First Run Dwell Time	native	5.51	246.9	15.31	0.04	762.5	0.09
	non-native	5.45	231.6				
Skip Count	native		0.552	-0.08	0.03	990.1	0.01
	non-native		0.631				
Second Fixation Duration	native	5.28	197.2	-16.85	0.04	461.2	0.06
	non-native	5.37	214.0				
Second Run Dwell Time	native	5.48	238.9	3.79	0.05	430.7	0.76
	non-native	5.46	235.1				
Total Dwell Time	native	6.04	419.1	-22.37	0.05	761.4	0.30
	non-native	6.09	441.4				
Regression-in Count	native		0.41	-0.17	0.05	982.8	0.00
	non-native		0.59				
Run Count	native		1.61	-0.12	0.09	992.6	0.16
	non-native		1.73				
Fixation Count	native		1.97	-0.05	0.12	992.9	0.71
	non-native		2.02				

Discussion

How do the scores assigned to writing differ between composition and L2 writing teachers?

Based on previous research, we expected that composition teachers would give lower scores than L2 writing teachers in general (Weigle, Boldt, & Valsecchi, 2003) and especially in grammar (Sweedler-Brown, 1993; Weigle, Boldt, & Valsecchi, 2003). The results seem to match our expectations; the average scores were lower for composition teachers than for L2 writing teachers for each category, with grammar having the largest mean difference between composition and L2 writing teachers' scores.

The mixed-effects analysis also supports the previous finding through the significant interaction found in it. The interaction shows that L2 writing teachers tend to give significantly higher scores on vocabulary, grammar, and overall than composition teachers. Though we can't say much about the vocabulary scores because we did not control for it in our design, it would make sense that the scores would be different because nonnative students have been shown to have smaller vocabulary than native students (Crossley & McNamara, 2009). Differences in vocabulary and grammar, and especially in overall scores could be disconcerting for students transitioning from ESL to mainstream FYC classes who may become accustomed to the higher scores when being taught by L2 writing teachers.

The comparison of grammar scores suggests that composition and L2 writing teachers assess grammar differently. Though we do not yet know for certain what is causing the difference in scores, it may reflect the findings of Sweedler-Brown (1993) who said that composition teachers are harsh on grammar errors even when other features are strong. On the other hand, it could be that L2 writing teachers are being lenient because of their experience with L2 writing (Eckstein, Briney, Chan & Blackwell, 2018).

How do the reading measurements of composition and L2 writing teachers compare when they rate grammar in nonnative students' writing?

To reiterate our expectations, we thought that there would be no significant differences in early reading measures because we expect both composition and L2 writing teachers to take the same amount of time to decode and recognize words. We did, however, expect differences to appear in late reading measures when they are trying to process and integrate the text. We believed that teachers' different backgrounds and training would result in differences in how they process the errors. Our discussion on the assigned scores would seem to match these expectations. Since they are reaching different scores, we would expect them to process the text differently as well.

However, after performing a mixed-effects analysis on each of the reading measurements, we found no statistically significant differences between composition and L2 writing teachers in how they look at nonnative-like errors (see Table 5). This was surprising for us because it contradicts our expectations, and also the findings of Eckstein, Briney, Chan, and Blackwell (2018), which showed that L2 writing teachers spent more time on errors than composition teachers. Although we see no significant differences in how composition and L2 writing teachers decode and recognize nonnative-like grammar errors, we also see that they do not significantly differ in how they integrate and process the text afterwards. However, if we were to consider total dwell time to be significantly different (p= .08), then it brings us closer in agreement with Eckstein et al. The L2 writing teachers in our study spent more time overall looking at nonnative-like errors than the composition teachers, suggesting that the L2 writing teachers took more time to process the errors.

Returning to the discussion on the assigned scores, we expected that composition and L2 writing teachers read and interact with errors differently based on the fact that they reach significantly different scores. However, this does not seem to be the case; only total dwell time was close to being significant. There must be something other than reading behaviors that is causing teachers with different backgrounds to reach different scores. One possible explanation could be that teachers reach scores in diverse ways (Brown, 1991) based on which features of writing they think are most important. Another explanation could be that teachers implement different strategies, strategies for interpreting the text and strategies for judging the text, to reach the scores (Cumming, 1990). Rater bias (Lindsey & Crusan, 2011) is another possible explanation. Although teachers read the text in a similar way, they may reach different scores based on biases they may have towards certain ethnolinguistic backgrounds. A final likely reason is that of error gravity. Teachers place different weight on error types based on personal perceptions of what is a severe error or what is distracting for them (Hartwell, 1985).

How do the reading measurements of composition teachers compare when assessing native and nonnative writing?

We expected that composition teachers would not show differences in early reading measurements because we believed they would decode and recognize words similarly regardless of whether it contained native-like or nonnative-like errors. We did believe, however, that they would take longer in the late reading measures for nonnative-like errors because it is more cognitively demanding for them to integrate and process. We were once again surprised with the results.

The mixed-effects analyses showed statistically significant differences in first run dwell time and first fixation duration, both early reading measurements, but not in any other measurements (see Table 6). Furthermore, raters spent longer on native-like errors than nonnative-like errors. This indicates that composition teachers took longer to decode and recognize native-like errors, the errors they are most accustomed to. Perhaps because composition teachers are most accustomed to native-like errors, it is easier for them to recognize native-like errors and then correct them in their mind in the moment they recognize the errors. This would explain the slightly longer dwell times for native-like errors. Whereas because composition teachers are not as accustomed to nonnative-like errors, they choose to return to or process the errors later. This would result in shorter dwell times because the teacher did not correct the errors the moment they recognized them. This explanation would also be supported by the significant difference in regression-in count (see Table 6). Composition teachers tended to look back at nonnative-like errors more than native-like errors. If composition teachers already corrected native-like errors after recognizing them, they would not need to return to them. If composition teachers were choosing to leave nonnative-like errors for later then they would need to look back at them.

One more point of interest is that once initially processed, composition teachers do not find it necessarily difficult to integrate and process text with nonnative-like errors, as shown by no significant differences in later reading measures.

How do the reading measurements of L2 writing teachers compare when assessing native and nonnative writing?

Our expectations for L2 writing teachers were similar to those for composition teachers in that we did not think there would be any differences in early reading measurements and there would be differences in late reading measurements accounting for their familiarity with nonnative-like errors. Furthermore, after seeing that composition teachers looked at native-like errors longer and speculating that this was because they were more familiar with the errors, we thought that the same would be true for L2 writing teachers and nonnative-like errors.

The mixed-effects analyses showed statistically significant differences in skip count and regression-in count (see Table 7). The L2 writing teachers skipped nonnative-like errors more often than native-like errors, and they returned to nonnative-like errors more often than to native-like errors. The skip count may be due to the L2 writing teachers' familiarity with the errors. Since they are familiar with the errors, they strategically choose to skip over the errors. Another explanation is that the errors were highly predictable, which allowed raters to skip them initially. For the regression-in count, it may be that although L2 writing teachers are familiar with nonnative-like errors, they still needed to go back to reread them and process the ambiguity in them, meaning that nonnative-like errors may be inherently harder to comprehend than native-like errors. The fact that composition teachers also looked back at nonnative-like errors more often seems to support this explanation.

Also, first run dwell time was close to being significant (see Table 7). L2 writing teachers looked at native-like errors longer than nonnative-like errors, which is similar to what composition teachers did and opposite of what we expected. Since both composition and L2 writing teachers are looking at native-like errors longer in the early reading measures, the reason for this may not be due to familiarity. Instead, a possible explanation is that native-like errors are harder to decode. This could be because missing commas and vague pronouns are included as a

native-like error. Missing commas and vague pronouns tend to cause ambiguity on the sentence level rather than on the word level, which is why it may have taken longer to decode.

Conclusion

Composition teachers were found to assign significantly lower scores to grammar than L2 writing teachers. This suggests that composition and L2 writing teachers assess grammar differently. Though we do not yet know why, we speculate that error gravity is a likely explanation for the difference. That is, teachers place different weight on different types of grammar errors. We think that the difference in background and training affects teachers' perceptions of error gravity, which is why the composition teachers and L2 writing teachers reached different scores.

The data also shows that there were no significant differences between errors that composition and L2 writing teachers look at when reading student writing, as can be shown by the similar early and late reading measurements. We interpret this to mean that teacher background does not influence teachers' reading behaviors when looking at grammar errors. We must, however, still consider that they reach different scores even though they read it similarly, which indicates that there are cognitive processes that we were not able to measure in this study. Such processes may be strategic processes, rater bias, or error gravity.

We also noticed that composition and L2 writing teachers both took longer to decode and recognize native-like errors than nonnative-like errors in the initial stages of reading. Initially, we thought it was because of familiarity. However, since both composition and L2 writing teachers took longer on native-like errors, we decided that this is likely not the case. Instead, it

may be that native-like errors used in the current study caused ambiguity on the sentence level, which would be harder to decode than ambiguity on the word level.

Overall, teacher background does not seem to affect how teachers read student writing, whether it has native-like or nonnative-like errors. However, teachers do assign scores differently based on their background, with L2 writing teachers assigning higher scores than composition teachers. Thus, the difference in score must reflect different cognitive processes not measurable by what teachers look at and attend to. We speculate that the most likely process is that of error gravity. Teachers place different weight on errors based on their individual experiences and preferences rather than on their background.

Limitations and Future Research

We recognize that the current research has several limitations that make it hard to generalize. The first limitation is concerning the participants. The researcher did not control for native and nonnative teachers, a variable that has been shown to affect grammar assessment (Kobayashi, 1992). Though this was a consideration, difficulty in finding participants made it impractical to control for. Future research can improve on this. Implications of such research will be applicable in ESL vs EFL settings. The current study briefly discussed the implications of international students who prepared at ESL schools within the U.S. who then transitioned to mainstream composition classes. This transition experience will likely be different for international students who prepare in their home countries before coming straight to U.S. universities if they are taught by nonnative teachers within their country.

Another limitation concerns how the eye-tracker can be programmed. In the current study, the paragraphs and the rubric were always displayed separately due to screen space

limitations. However, teachers may prefer to see rubrics and writing side-by-side in order to move back and forth throughout the assessment task. This limitation could have majorly impacted late reading measurements, which may become more salient after referring to the rubric. Further research may show that there are differences in late reading measures when this limitation is taken into consideration. One possible solution may be to use a TOBII tracker apparatus, which does not require a headrest so that participants can have a paper rubric to consult.

A final limitation was discovered after the high skip rate was calculated. One reason for the high skip rate may be due to the predictability of the errors used in the current study, which in hindsight used mostly function words. Future research can improve by using more content words so that the words containing errors are not skipped as often. This can provide more robust data for analysis.

Research on error gravity has been stagnant for the past 20 years, perhaps because of the limitations of technology at the time. The current research applied new eye-tracking technology to answer why composition and L2 writing teachers score grammar differently, a question that has not yet been answered after years of research. Although the current study was still not able to clearly identify why composition and L2 writing teachers reach different scores, it was able to identify that differing reading behaviors is not the reason. Eye-tracking research still has potential to bring us closer to the answer. Future research can combine eye-tracking with reflective protocols to address this question. For example, researchers can initially collect data via an eye-tracker and then have participants reflect on why they regressed, skipped, looked for a long duration, etc. Researchers can also ask about error gravity during the reflection. Combining

the two methods can help us understand the cognitive processes and the behaviors of the teachers.

References

Atkinson, D., & Ramanathan, V. (1995). Cultures of writing: An ethnographic comparison of L1 and L2 university writing/language programs. *TESOL Quarterly*, 29(3), 539-568.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, 7, 54-74.

Borg, S. (1998). Teachers' pedagogical systems and grammar teaching: A qualitative study. *TESOL Quarterly*, 32(1), 9-38.

Braddock, R., Lloyd-Jones, R., & Schoer, L. (1963). Research in written communication. *Urbana, IL: National Council of Teachers of English*.

Braine, G. (1996). ESL students in first-year writing courses: ESL versus mainstream classes. *Journal of Second Language Writing*, 5(2), 91-107.

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly*, 25(4), 587-603.

Company, M. (2012). Error frequencies among ESL writers: A resource guide. *All Theses and Dissertations,* Paper 3420.

Conklin, K., & Pellicer-Sanchez, A. (2016). Using eye-tracking in applied linguistics and second language research. *Second Language Research*, 32(3), 453-467.

Connor, U. (2011). *Intercultural rhetoric in the writing classroom*. Ann Arbor: University of Michigan Press.

Connor-Linton, J., & Polio, C. (2014). Comparing perspectives on L2 writing: Multiple analyses of a common corpus. *Journal of Second Language Writing*, 26, 1-9.

Crossley, S. A., & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, 18(2), 119–135.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, 7(1), 31-51.

Eckstein, G., & Ferris, D. (2017). Comparing L1 and L2 texts and writers in first-year composition. *TESOL Quarterly*, 52(1), 137-162.

Eckstein, G., Caspter, R., Chan, J., & Blackwell, L. (2018). Assessment of L2 student writing: Does teacher disciplinary background matter? *Journal of Writing Research*, 10(1), 1-23.

Elder, C., Golombek, P., Weigle, S. C., Boldt, H., & Valsecchi, M. I. (2003). Effects of task and rater background on the evaluation of ESL student writing: A pilot study. *TESOL Quarterly*, 37(2), 345-354.

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

Ferris, D. R. (2009). *Teaching college writing to diverse student populations*. Ann Arbor: The University of Michigan Press. https://doi.org/10.3998/mpub.263445

Ferris, D., Brown, J., Liu, H., & Stine, M. E. A. (2011). Responding to L2 students in college writing classes: Teacher perspectives. *TESOL Quarterly*, 45(2), 207-234.

Ferris, D., Eckstein, G., & DeHonde, G. (2017). Self-directed language development: A study of first-year college writers. *Research in the Teaching of English*, 51(4), 418-440.

Francis, W. N. (1954). Revolution in grammar. *Quarterly journal of speech*, 40(3), 299-312.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14, 178-210.

Harklau, L. (1994). ESL versus mainstream classes: Contrasting L2 learning environments. *TESOL Quarterly*, 28(2), 241-272.

Harrington, S., Malencyzk, R., Peckham, I., Rhodes, K., & Yancy, K. B. (2001). WAP outcomes statement for first-year composition. *College English*, 63(3), 321-325.

Hartwell, P. (1985). Grammar, grammars, and the teaching of grammar. *College English*, 47(2), 105-127.

Huey, E. B. (1908/1968). *The psychology and pedagogy of reading*. Cambridge, MA: MIT Press.

Institute of International Education, 2017. International student enrollment trends, 1948/49 – 2016/17. *Open Doors Report on International Educational Exchange.* Retrieved from http://www.iie.org/opendoors.

Jacobs, H., Zinkgrap, S., Wormuth, D., Hartfiel, V., & Hughey, J. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.

Janopoulos, M. (1992). University faculty tolerance of NS and NNS writing errors: A comparison. *Journal of Second Language Writing*, 1(2), 109-121.

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Newton, MA: Allyn & Bacon.

Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly*, 26(1), 81-112.

Kolln, M. (1981). Closing the books on alchemy. *College Composition and Communication*, 32(2), 139-151

Lance. D. M. (1977). What is "Grammar"? *English Education*, 9(1), 43-49.

Leki, I., Cumming, A., & Silva, T. (2008). *A synthesis of research on second language writing in English*. London: Routledge.

Lindsey, P., & Crusan, D. J. (2011). How faculty attitudes and expectations toward student nationality affect writing assessment. *Across the Disciplines: A Journal of Language, Learning, and Academic Writing*, 8.

Lunsford, A., & Lunsford, K. (2008). "Mistakes are a fact of life": A national comparative study. *College Composition and Communication*, 59(4), 781-806.

Matsuda, P. K. (1999). Composition studies and ESL writing: A disciplinary division of labor.

Matsuda, P. K. (2012). Let's face it: Language issues and the writing program administrator. *WPA: Writing Program Administration*, 36(1), 141-163.

Matsuda, P., Saenkhum, T., & Accardi, S. (2013). Writing teachers' perceptions of the presence and needs of second language writers: An institutional case study. *Journal of Second Language Writing*, 22, 68-86.

Montgomery, J. L., & Baker, W. (2007). Teacher-written feedback: Student perceptions, teacher self-assessment, and actual teacher performance. *Journal of Second Language Writing*, 16(2), 82-99.

Paulson, E., Alexander, J., & Armstrong, S. (2007). Peer review re-viewed: Investigating the juxtaposition of composition students' eye movements and peer-review processes. *Research in the Teaching of English*, 41(3), 304-335.

Pelosi, A. G. (1973). What is "Grammar"? *The Modern Language Journal*, 57(7), 329-335.

Rayner, K. (1997). Understanding eye movements in reading. *Scientific Studies of Reading*, 1, 317-339.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 372-422.

Rifkin, B., & Roberts, F. D. (1995). Error gravity: A critical review of research and design. *Language Learning*, 45(3), 511-537.

Rubin, D. L., & Williams-James, M. (1997). The impact of writer nationality on mainstream teachers' judgments of composition quality. *Journal of Second Language Writing*, 6(2), 139-154.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly*, 22(1), 69-90.

Santos, T. (1992). Ideology in composition: L1 and ESL. *Journal of Second Language Writing*, 1(1), 1-15.

Shvidko, E. (2015). Assessing preparation of mainstream composition teachers working with multilingual writers. *INTESOL Journal*, 12(2), 37.

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students?. *Journal of Second Language Writing*, 5(2), 163-182.

Sweedler-Brown, C. O. (1993). ESL essay evaluation: The influence of sentence-level and rhetorical features. *Journal of Second Language Writing*, 2(1), 3-17.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning*, 46(2), 327-369.

Truscott, J. (1999). The case for "The case against grammar correction in L2 writing classes": A response to Ferris. *Journal of Second Language Writing*, 8(2), 111-122.

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record*, 58(3), 475-482.

Appendix A – Paragraphs

Paragraph 1

The biggest problem I face in writing is organizing my thoughts. Perhaps due to my strong disliking of reading book (S/P) as a child, I struggle tremendously with organizing my thoughts and displaying them into the (determiner) words; I face this same problem even in verbal communication. Prior to starting an essay (missing comma) I fill my head with so many different idea (S/P) that I consider brilliant, and start to design a (determiner) illustration of it (vague pronoun) into words and paragraphs. Unfortunately, often times this turns out to be nothing but my literary fantasy. When I am actually writing an essay they (vague pronoun) become a mash-up of vague ideas that don't even seem to relate on (preposition) one another. They are just a random collection of ideas. By the end of it (vague pronoun), the reader might not even remember what the thesis of the essay was and clueless as to what message I was trying to convey.

Paragraph 2

When it comes to writing (missing comma) I feel that I have a love to (preposition) it, but not the knowledge. I strive to create (determiner) piece that shows my creativity and my efficiency as a writer, but the problem is that I don't even know how to start or where to end. Sometimes I think it (vague pronoun) isn't even creative. I enjoy writing, telling stories, and proving points. that (capitalizaton) is why it is so difficult and frustrating for me to know that my writing skill (S/P) are not even close to what I wish they were. I have had problems with it (vague pronoun) since grade school in writing. I confuse many word (S/P), have a hard time

remembering simple grammar rules, I tend to be too wordy, and I can never make it (vague pronoun) clear. I am hoping to obtain greater knowledge and improve my writing abilities.

Paragraph 3

The best words to describe myself as a writer would be "Reluctant" (capitalization) and "Insecure" (capitalizaton). I hate writing paper (S/P) with a passion and I can never seem to get started (missing comma) and once I am done writing, I really don't want anyone to read it. We had to do the (determiner) correcting exercises in High (capitalization) school, and a lot of the really bad papers sounded a lot like mine. I'm not good at grammar. My paragraphs don't flow which makes it (vague pronoun) choppy and hard to read. Even now (missing comma) I continually delete all my ideas that I put down because I am dissatisfied with what I come up with. I think it is even evident in (determiner) way that I speak. I speak very quickly with crazy ideas flowing everywhere.

Paragraph 4

Thoughts and idea (S/P) have no purpose, no existence, without the craft of the Writer (capitalizaton) himself. Expression of feeling through words has always been difficult with (preposition) me simply because throughout my life I have often found myself numb to the many color (S/P) of emotion that give it such purpose. However, this lack of so-called "Zest" (capitalizaton) has perhaps given it (vague pronoun) a beneficial stance compared to others. My writing is numb, almost in the (determiner) sense cold, but I believe that this allows me to see things on (preposition) a different light. I wish to write about things as how they are, not as they should be in an ideal world. This realist notion is very apparent on (preposition) my past essay

that I analyzed from high school. I think it (vague pronoun) is very important because it really shows a part of who I am. While it may make people feel a bit sad and preoccupied, I believe my past and current writing makes people think, both about themselves and the world. I believe that real, raw writing makes people react in such a manner.

Paragraph 5

Writing has always been the (determiner) trivial art in my eyes. I have never thought of myself as being a good writer. Despite what the grade may be in (preposition) the paper after I turn it in, I still feel that my skills are inadequate. this (capitalizaton) being said, I can also respect the fact that despite my dissatisfaction with my current skill, I most definitely can appreciate the development at (preposition) my writing throughout my life. When I wrote pieces in high school I really never found myself wanting to show people who I really was and am. I just wanted to get a good grade and move on. It (vague pronoun) wasn't inspiring others and myself; it was just to move on with my life. Then, during the vast span of time in which I was working on my ap (capitalization) Government essay, I suddenly realized something was different. I cared for that collection of word (S/P) almost more than I would a person. I found myself deeply pursuing every avenue of theory within our government and only selecting the words that I felt were worthy of giving my audience an (determiner) true meaning of my thought. For once in my life as a writer, I really didn't feel that numb. I was passionate (missing comma) determined, and hungry. Hungry to show people that I could give thoughts meaning to an audience, and ultimately, to the world.

Paragraph 6

As (determiner) writer, I feel pretty confident with the standards of writing, such as grammar, punctuation, and usage. I find my pieces satisfactory because I find my writing style rather reflective to (preposition) the way I actually communicate verbally. I often hear myself saying what I'm writing or just read it aloud to question whether I would say what I'm writing or word something the (determiner) certain way. I feel relatively comfortable writing just about anything (missing comma) yet sometimes I feel limited when I find myself in a stump. Often, I encounter situations where I can't express my ideas and thoughts the way I want to (missing comma) or I can't think of anythings (S/P) to write. However, I find it (vague pronoun) to be generally clear, concise, and somewhat affective. Overall, I think I'm pretty decent. My pieces meet with (preposition) my expectations, but I feel the need to go beyond what is expected. I know I'm the (determiner) not best writer, but I think I have the potential to be a really good, effective one.

Paragraph 7

Although there may be a variety of errors and a lack of variety (missing comma) I am content to know that my writing is not completely dull. Every sunday (capitalizaton), I would go back to my church as a youth group leader and teach young childrens (S/P). Kids have a different level of understanding then adults. In order for them to understand what I am trying to show for (preposition) them, I would tell stores about my life or the lives of the prophet (S/P) in the bible (capitalizaton). When I write an essay, my evidence would be given in the same method. Most of my evidence comes among (preposition) my own personal stories and experiences. I try to entertain my audience because if you can not keep a child's attention, then they lose the will to learn. However (missing comma) this means that all of my evidence is typically given in a

roundabout way. the (capitalization) evidence would be interesting to read, but may not be received the same from one person to the next.

Paragraph 8

Throughout my fifteen year (S/P) of education, I never really thought of myself as a good writer. It became apparent to me that writing was one of my weakest subjects as I moved on onto (preposition) high school where writing papers and essays became more relevant. I would always have Writer's Block (capitalization) and I would take the (determiner) very long time to write an essay. I was not able to put my thoughts into words so they (vague pronoun) always seemed shorter. I rarely got A's on any of the papers I wrote in high school. My experiences on (preposition) writing drew me away from it and I saw that my strengths were in different subjects. The reason for this is because I always wanted to have a perfect first draft which never really happens. In order to write a good paper, one must have at least something to work with like a bad first draft (missing comma) and from that you mold it into a piece of work. This I did not learn until my final year of high school (missing comma) and even then I did not get that much better. I struggled with many papers such as my college application essays. They were hard and I often ended up looking at facebook or watching a movie in between, which made me forget exactly where I was. Although I spent many months working on them with many revisions (missing comma) I still felt like they were lacking.

Appendix B - Rubric

	Content		Organization		Vocabulary		Grammar
7 6	• Thorough development of thesis • Substantive and detailed • No irrelevant information • Interesting	7 6	• Excellent paragraph organization • Clear thesis statement or main idea • Excellent use of transition words	7 6	• Very sophisticated vocabulary • Excellent choice of words with no errors • Excellent range of vocabulary • Academic register	7 6	• No spelling errors • No punctuation errors • No major errors in word order or complex structures • Excellent sentence variety • No errors that interfere with comprehension
5 4	• Good development of thesis • Fairly substantive and detailed • Almost no irrelevant information • Somewhat interesting	5 4	• Good paragraph organization • Clear thesis statement or main idea • Good use of transition words	5 4	• Somewhat sophisticated vocabulary • Good choice of words with some errors that don't obscure meaning • Adequate range of vocabulary • Approaching academic register	5 4	• Minor spelling errors in less frequent words • No more than a few punctuation errors • Occasional errors in awkward order or complex structures • Good sentence variety • Almost no errors that interfere with comprehension
3 2	• Some development of thesis • Not much substance or detail • Some irrelevant information • Somewhat uninteresting	3 2	• Some coherent organization • Minimal thesis statement or main idea • Occasional use of transition words	3 2	• Unsophisticated vocabulary • Limited word choice with some errors obscuring meaning • Repetitive choice of words • No resemblance to academic register	3 2	• Some spelling errors with less frequent and more frequent words • Several punctuation errors • Errors in word order or complex sentences • Little sentence variety • Some errors that interfere with comprehension
1 0	• No development of thesis • No substance or details • Substantial amount of irrelevant information • Completely uninteresting	1 0	• No coherent organization • No thesis statement or main idea • No use of transition words	1 0	• Very simple vocabulary • Severe errors in word choice that often obscure meaning • No variety in word choice • No resemblance to academic register	1 0	• Spelling errors even in frequent words • Many punctuation errors • Serious errors in word order or complex structures • No sentence variety • Frequent errors that interfere with comprehension

Appendix C – Exit Interview

1. How much experience have you had in assessing native writing? Nonnative writing? mixed classes?

2. How have you approached assessing native and nonnative writing? And for grammar?

3. Do you think that grammar mistakes should affect the grades of native writing? How about nonnative writing?

4. How did you approach the grammar mistakes in the paragraphs you read today?

5. Which types of grammar mistakes in the paragraphs you looked at today were the most distracting? Least distracting?

6. Which types of grammar mistakes affected your perception of the quality of the writing the most? Affected the least?