TESL-EJ Teaching English as a Second or Foreign Language

The Electronic Journal for English as a Second Language

May 2019– Volume 23, Number 1

Reading L1 and L2 Writing: An Eye-tracking Study of TESOL Rater Behavior

Grant Eckstein
Brigham Young University
<grant_eckstein@byu.edu>

Wesley Schramm
Brigham Young University
<wschramm4649@gmail.com>

Madeline Noxon
Brigham Young University
<mnoxon521@gmail.com>

Jenna Snyder
Brigham Young University
<jenna.snyde@gmail.com>

Abstract

Researchers have found numerous differences in the approaches raters take to the complex task of essay rating including differences when rating native (L1) and non-native (L2) English writing. Yet less is known about raters' reading practices while scoring those essays. This small-scale study uses eye-tracking technology and reflective protocols to examine the reading behavior of TESOL teachers who evaluated university-level L1 and L2 writing. Results from the eye-tracking component indicate that the teachers read the rhetorical, organizational, and grammatical features of an L1 text more deliberately while skimming through and then returning to rhetorical features of an L2 text and initially skipping over many L2 grammatical structures. In reflective interviews, the teachers also reported more consensus on their approach to evaluating grammar and organization than word choice and rhetoric. While these findings corroborate prior research comparing the rating of L1 and L2 writing, they promise to expand our understanding of rating processes by reflecting the teachers' reading practices and attentional focus while rating. Moreover, the study demonstrates the potential for using eye-tracking research to unobtrusively investigate the reading behaviors involved in assessing L1 and L2 writing.

Introduction

Essay writing serves as a critical gatekeeping mechanism and subsequent measurement tool of student learning in most colleges. For instance, students are granted or denied acceptance to the university partially on their essay writing abilities. And once enrolled, students regularly perform on essay assessments in required courses. Because writing assessment in educational settings is pervasive, researchers in recent years have taken an increased interest in the reading behaviors associated with essay rating (DeRemer, 1998; Huot, 1993; Wolfe & Ranney; 1996). Wolfe (2005), for instance, has posited an elaborate scoring framework which involves reading a text, creating a mental image of it, and then performing processing actions that compare the image of the text to the criteria of an associated rubric. Edgington (2005) found that teachers utilize numerous strategies, including "evaluating, clarifying, questioning, and inferring" (p. 142) when reading student writing. However, there remain questions of whether raters adopt the same reading behaviors in similar proportions to texts written by native English (L1) and non-native English (L2) writers.

The issue of reading for evaluation has gained relevance recently given that there are now more international students studying in U.S. higher education than ever before (Institute of International Education, 2016). As a result, teachers increasingly encounter L1 and L2 students in their classes, read their essays side-by-side, and ultimately assign grades based partially on that writing. National Council of Teachers of English's (NCTE) Standards for the Assessment of Reading and Writing explains that language assessments must be sensitive to "the length of time students need to become skilled at [basic and academic language]" (2009, para. 3). When L1 and L2 writers take the same writing assessments, there can be large disparities in the linguistic backgrounds of these students making it difficult to rate their writing in identical ways.

Moreover, when reading L1 and L2 texts, Hall (2014) calls for an awareness he terms multilingual equity which encourages raters to be aware of and sensitive to students' language backgrounds and needs, echoing Raimes' (1985) argument that L2 writers need more time to pre-write, draft, revise, and edit their writing. Raters who are sympathetic to language learners, for instance, may change their reading and rating behaviors to accommodate for differences in students' backgrounds, such as showing some leniency for language errors or acknowledging differing rhetorical styles among L2 writers. While some evidence appears to suggest that raters use a similar variety of strategies when rating essays by both L1 and L2 writers (Erdosy, 2004; Zhang, 2016), other studies show that raters appear to read and rate L1 and L2 texts differently (Barkaoui, 2010; Cumming, 1990; Cumming, Kantor, & Powers, 2001; Haswell, 2007; Lindsey & Crusan, 2011; Rubin & Williams-James, 1997; Song & Caruso, 1996; Sweedler-Brown, 1993; Vaughan, 1991). Despite the high quality of these research strands, all information about reading differences to this point have been gleaned from surveys, think-aloud protocols, and interviews (Winke & Lim, 2015). With the development of eye-tracking technology, however, it is now possible to add direct reading behavior measurements to the discussion of differential L1/L2 essay rating processes. The present study provides this additional insight by investigating the reading behavior of several ESL writing teachers as they read and rated L1 and L2 writing. Raters with ESL training were chosen for this study because of their professional familiarity with features of L2 writing (Eckstein, Casper, Chan, & Blackwell, 2018) and a sensitivity to L2 writers' needs. Results demonstrate specific areas of attentional focus which contribute to research of differences in the reading and rating of L1 and L2 writing.

Literature Review

It should come as no surprise that L1 writing can differ widely from that of L2 writing [1], even when produced by similarly-skilled writers. Leki, Cumming, and Silva (2008) reviewed twenty years of second language writing research and found differences in cohesive device usage, organizational patterns, discourse modes, grammar issues, and lexical control between L1 and L2 texts. More recent corpus-based research by Ai and Lu (2013) demonstrated that non-native English writers produced less syntactically complex text than L1 English writers, while Crossley and McNamara (2009) reported that L2 writers had less lexical variation, sophistication, and depth of knowledge in their texts. Eckstein and Ferris (2018) showed that L2 writers made more grammatical errors across nine broad error categories than L1 writers. It is perhaps these textual differences that trigger legitimate differential ratings between L1 and L2 essays.

However, not all differences in rating judgements can be explained by textual features alone. Researchers have established that raters judge texts differently based on their backgrounds and biases external to the texts (Ball, 1997; Pula & Huot, 1993). Raters' rating proficiency is one mitigating factor that illustrates this. In his think-aloud study, Wolfe (2005) found evidence of raters' cognitive behaviors on a variety of processing actions and content foci. Ultimately, he argued that less proficient raters tended to make and revise assessment decisions early in the rating task while more proficient raters withheld their assessments until after completing the text.

Disciplinary background has emerged as another discriminating feature (Roberts & Cimasko, 2008). Cumming et al. (2001) found that raters' instructional backgrounds as native-English or ESL teachers affected the way they described their rating processes as did Eckstein et al. (2018) in an eye-tracking study of essay rating behavior. Moreover, raters' biases for or against perceived student ethnicity also play into L1 and L2 scoring. In a matched-guise study, Rubin and Williams-James (1997) applied Thai, Danish, and American guises to a stable set of six essays and collected holistic and analytic scores from 33 writing teachers. Results showed that when they believed a text was written by an L2 writer, raters assigned higher holistic and lower analytic scores to presumed L2 writers. In a more recent replication study of teachers across the curriculum, Lindsey and Crusan (2011) observed similar results and concluded that simply seeing a student's name can trigger differential scoring for L1 and L2 essays.

Haswell (2007), using another matched-guise design, asked teachers to evaluate the same student essay under the supposition that it was written by an L1 and an L2 writer respectively. Raters also reported the importance of 10 criteria on their judgements. Results showed that "on every trait used to judge a particular essay, the presumed L2 writer was rated more highly than the presumed L1 writer" (p. 13). Yet in describing which traits contributed most to raters' perception that the student was ready to exit first-year composition, the two groups of raters disagreed on the relative importance of 7 of the 10 traits. Taken together, these various studies suggest that raters do not only judge L1 and L2 texts differently, but that they approach L1 and L2 texts with different criteria in mind.

Criteria Raters Attend To

Subjective attentional focus appears to inform the criteria raters consider when evaluating writing (see Cumming, 1990; DeRemer, 1998), as does the rubric design (i.e., holistic or analytic) (Knoch, 2009; Jonsson & Svingby, 2007; Walcott & Legg, 1998; Weigle, 2002). Holistic scoring is still used extensively in language testing (Singer & LeMahieu, 2011), and researchers have aimed to identify criteria raters use in assigning their scores (Kuiken & Vedder, 2014; Perkins, 1980; Rafoth & Rubin, 1984; Sparks, 1988; Sweedler-Brown, 1993).

Raters tend to consider multiple textual qualities such as content, organization, and grammar when assessing both L1 and L2 texts (Rafoth & Rubin, 1984; Song & Caruso, 1996; Sweedler-Brown, 1993). For instance, Cumming (1990), who explored rater judgments of L2 texts, found that rater's decision-making comments could be classified under four major dimensions he termed content, language, organization, and self-control. Song and Caruso (1996) observed that English faculty were influenced by content and rhetorical features in texts even while ESL faculty attended more to language use. Cumming et al., (2001) found further that raters with experience teaching English composition attended to ideas, argumentation, and language in L2 compositions while ESL-trained raters were more attentive to language overall.

Other researchers, however, have reported conflicting results to those above with raters appearing to make evaluations that were linked to vocabulary and grammar issues, particularly for L2 texts (Homburg, 1984; Kuiken & Veder, 2014; McDaniel, 1985; Sweedler-Brown, 1993). Homburg (1984), for instance, compared several textual measures to holistic scores of 30 L2 essays by trained raters and found that sentence-level grammar accounted for 84% of score variance; these included error-free T-units, dependent clauses and coordinating conjunctions per essay, and words per sentence. Additionally, Vaughan (1991) recorded nine experienced raters as they evaluated two L1 and four L2 essays. Raters made more frequent comments on grammar (including punctuation and language choice) than content. Kuiken and Vedder (2014) demonstrated that lexical diversity and accuracy correlated with raters' scores of linguistic complexity for L2 text. Other studies, such as those by Rafoth and Rubin (1984) and Sweedler-Brown (1993), illustrate that raters were shown to pay more attention to grammar and sentence-level errors than content or rhetorical features in both L1 and L2 texts.

Overall, prior research is complex and contradictory. It suggests that in some cases raters evaluate L1 and L2 writing on a wide variety of features encompassing grammar, vocabulary, organization, and rhetorical features. Yet in some cases, raters fixate on grammar errors or overlook them as a way to compensate for L2 language development (viz-a-viz multilingual equity) (Sakyi, 2000; Song & Caruso, 1996; Sweedler-Brown, 1993). Although various studies have shown glimpses into rater cognition, especially relative to grammar, vocabulary, organization, content, and rhetorical elements (Eckstein et al., 2018; Crossley, et. al., 2014; Kuiken & Veder, 2014; Sparks, 1988; Sweedler-Brown, 1993), there is still a lack of consensus on what raters read, and thereby attend to, when rating L1 and L2 essays. This lack of consensus suggests that we have much to learn about the reading practices associated with evaluating L1 and L2 texts.

When identifying the textual elements raters attend to or read during assessment tasks, research has largely been carried out using surveys, interviews, and think-aloud protocols (Winke & Lim, 2015). Think-aloud protocols in particular have been useful in describing raters' in-the-moment decision-making processes (Barkaoui, 2011). For instance, studies by Edgington

(2005) and Wolfe (2005) have illustrated reading practices of writing raters and differences between more and less experienced raters overall, and other researchers have shown that raters attend to different qualities of L1 and L2 texts when making their rating decisions (Barkaoui, 2010; Cumming, 1990; Cumming et al., 2001; Vaughan, 1991). Despite providing valuable data, these protocol approaches have also been criticized for not fully representing a rater's actual thinking and for potentially interfering with a rater's evaluation process (Barkaoui, 2011; Godfroid & Spino, 2015; Lumley 2005; Winke & Lim, 2015). For instance, Barkaoui (2011) found that think-aloud protocols incompletely reflected raters' complex thought processes and that rater experience and rating scale type led to variability in rater verbalizations. Additionally, reading and articulating thoughts aloud also appears to affect rater evaluations by focusing raters' attention on local errors and reducing text comprehensibility and vocabulary recognition (Barkaoui, 2011; Godfroid & Spino, 2015; Stratman & Hamp-Lyons, 1994).

An alternative approach to investigating reading behavior while rating is the use of eye-tracking technology (Conklin & Pellicer-Sánchez, 2016). Although eye-tracking research has a long history of applications in psychological research generally (see van Gompel, 2007) and reading studies more particularly (see Rayner, 1998, 2009) it is relatively new to writing and assessment research (Anson & Schwegler, 2012; Brunfaut & McCray, 2015) and TESOL scholarship specifically (Godfried, 2018). Nevertheless, eye-movement measurements can add to an understanding of rater reading behavior because these measurements occur during uninterrupted reading and thus are unobtrusive, temporally and chronologically precise, and ecologically valid (Godfroid & Spino, 2015; Paulson, Alexander, & Armstrong, 2007). Of course eye-tracking has its own limitations as well inasmuch as raters must read while located at a computer in a laboratory setting, and often with their heads stabilized. Also, eye-tracking alone only measures eye movements not actual thinking, though Rayner (1998, 2009) describes a link between eye-movements, attentional focus, and cognition which allows for limited interpretation of rater thinking which can be confirmed through additional means such as reflective interviews or protocols. Thus, eye-tracking, when coupled with rater reflection, presents a novel and valid additional research tool for investigating reading behavior during essay rating.

Eye-tracking Measures

It is commonly assumed that during reading, the eyes glide smoothly over written text taking in information word-by-word. However, this conceptualization is inaccurate; instead, the eyes make short jerking movements, called saccades, that Rayner (2009) explains cover on average about 8 letter spaces. Saccades occur in between periods when eye movements stabilize, called fixations, which are about 225 milliseconds long on average (Rayner, 2009). It is during fixations that visual information is thought to be considered by the reader (Conklin & Pellicer-Sanchez, 2016). The eyes also often look over different parts of a word, return to previously-read words in order to process them further, and linger on unfamiliar, complex, or grammatically incorrect words (Conklin & Pellicer-Sanchez, 2016; Rayner, 1998). Readers also skip up to 30% of words on first-pass reading, though these are usually short words and sometimes highly predictable ones (Rayner, 2009).

Eye-trackers measure the duration of fixations and saccades in milliseconds with the premise that readers pay attention to and cognitively process only that information limited to their foveal vision (the tiny window of sight that remains in focus) as opposed to their parafoveal or peripheral vision (Paulson, Alexander, & Armstrong, 2007; Rayner, 2009; Reichle, Pollatsek, & Rayner, 2006). When readers initially fixate on a particular word, they are decoding text, which refers to the transferring of a written form to a recognizable phonological form with an understandable meaning (Nation, 2009). Subsequent fixations of the same area relate to text comprehension, meaning that readers decode text prior to comprehending it. Repeated readings or returning to a previously fixated word as well as longer fixations of a given word at any time signal increased cognitive processioning and indicate textual elements or words that readers have difficulty processing, such as misspellings, grammatical inconsistencies, or less-frequent words in general (Ehrlich & Rayner, 1981; Frazier & Rayner, 1982; Rayner & Pollatsek, 1989).

Researchers have further distinguished various measures of eye movement within reading tasks that signal particular reading processes, such as first pass reading time, total reading time, regression counts, and so on (see Conklin & Pellicer-Sánchez, 2016 and methods section below for additional details and definitions). First pass reading time measures initial eye fixations of a given interest area in milliseconds and is connected to decoding processes. Total reading time measures total fixations for a given area of interest (AOI) which combines early reading measures and the later reading measures associated with comprehension. Run counts and regressions indicate instances of reading, re-reading, and further processing or comprehension break-downs. Figure 1 below illustrates several common eye-tracking measures and Table 1 explains the measures in more detail.

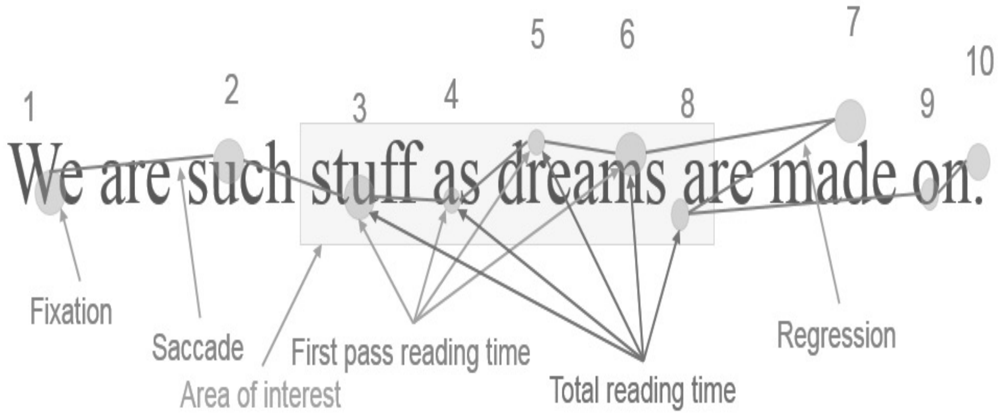

Figure 1. Common Eye-Movement Measures Used in Eye-Tracking Research.

Table 1. Eye-Tracking Measures and Descriptions.

Processing time	Eye-tracking measure	Measurement description	Measure purpose
Early reading processes	**First pass reading time**	The total time, in milliseconds, of all fixations within a specific area of interest for just the first pass	Decoding processes: letter- and word-recognition and lexical access
	Skip count	Indication that no fixations occurred in an interest area during first-pass reading	Strategic or expeditious early reading: previewing, skimming, scanning, or inferring words
Later reading processes	**Total reading time**	The total time, in milliseconds, of all fixations within a specific area of interest across all passes	Comprehension-based reading: meaning integration and attention to confusing, novel, unusual words
	Run count	The number of times a participant's gaze entered and left a specific area of interest irrespective of which direction the gaze originated	Comprehension-based re-reading: later attention to words or phrases for comprehension, confirmation or resolving confusion; includes later skimming and scanning
	Regression in count	Indication that an interest area was entered into from a later part of the sentence (e.g., rater looked back at a previous interest area)	Meaning integration: resolving confusion or confirmation check within the reading stream

Although substantial research has investigated eye-tracking and reading (see reviews by Rayner, 1998, 2009), very little has examined eye-tracking applications to writing assessment (Winke & Lim, 2015), and none that we know of have used eye-tracking to investigate how raters differentially interact with the writing of L1 and L2 writers. Yet raters may interact with L1 and L2 texts differently based on salient textual differences, and TESOL-trained raters may be especially sensitive to language-related differences on account of a disciplinary emphasis on language development (Eckstein et al, 2018). As a result, we embarked on a small-scale study to determine how TESOL-trained raters differentially read elements of an L1 and L2 text using an eye-tracking machine to capture quantitative measures of reading dwell time and regression and skip counts. We also used reflective protocols to elicit participant reflections on their reading behaviors. We anticipated that raters would be more attuned to the unique characteristics of L2 texts, meaning that they would demonstrate more fixations and re-reading of all features of the L2 text, especially grammar and word choice features, when compared to the L1 text since features of the L2 text would require longer processing and evaluation time and because these observations would be consistent with earlier think-aloud and observational studies (Cumming et al., 2001; Homburg, 1984; Kuiken & Veder, 2014; McDaniel, 1985; Sweedler-Brown, 1993). We use our results to provide preliminary insights into raters' reading practices when rating L1 and L2 essays.

Research Questions

The present small-scale study aimed to investigate the ways in which TESOL-trained writing teachers interacted with university-level L1 and L2 student writing. We did this by measuring the eye movement of writing teachers as they scored an L1 and an L2 text from a first-year composition assignment on four textual criteria: grammar, word choice, organization, and rhetorical content. We then asked for their reflective thinking on the rating experience. We were guided in our research design by these questions:

1. What differences in dwell time, regression, and skip count emerge when TESOL-trained writing teachers read L1 and L2 texts in areas of grammar, word choice, organization, and rhetoric?
2. What features do TESOL-trained writing teachers report considering when reading L1 and L2 student texts?

Methods

Participants

Two male and three female ESL writing teachers participated as essay raters following internal review board approval; they had between four and 16 years of teaching experience. All were native-English speakers; three had completed an MA in TESOL, and the others were finishing their final year of the same degree. All teachers were trained essay raters in the ESL writing program where they taught and participated in hour-long calibration and norming sessions three times yearly as part of their employment.

Instruments

Rubric. Raters were provided with a "semi-structured" holistic rubric subdivided into the four areas of rhetoric, word choice, organization, and grammar (see appendix A). Research has shown that raters differ relatively little when evaluating essays analytically (Song & Caruso, 1996). Because we were interested in examining variance in rater behavior between L1 and L2 text, we adopted a semi-holistic scoring approach which asked raters to mentally consider the four sub-components mentioned above but only report a single, holistic score for each essay. The four sub-component categories were articulated in order to provide some structure to rater thinking both during the rating process and in preparation for the reflective protocol used subsequent to the rating task. Thus, although raters did not assign analytical scores, they were encouraged to think about each rubric category. The rubric was drafted by the first researcher for a related research project and resembled rubrics used by the participants at the program where they were employed. It reflected the four critical components discussed in the criteria section of the literature review and was piloted on a small number of sample essays and then revised according to feedback and digitized. The rubric was designed to prime raters to consider the four textual features under consideration in this study. We anticipated that by being sensitized to these features, raters would focus their attention accordingly and their eye movements would reveal differences in how they considered those criteria across the different essays.

Essays. We selected two authentic student writing samples from an upper division writing course as input for this study. The essays were composed by one L1 and one L2 student during the first week of a 10-week first-year composition course at a four-year research university in the western United States. The essay prompt asked students to explain how they would describe themselves as a writer after having reflected on a piece of their own writing [2]. Students were given the prompt in advance and asked to select and read over any piece of writing they had composed in the past year in preparation for the 60-minute in-class writing task. Students were encouraged to write between 500-750 words. The two samples used in this study were selected because they received identical (and low) scores when assessed by two experienced raters using an analytic rubric focused on language proficiency. Our choice of essays was constrained by the limitations of eye-tracking methodologies: for a comprehensive evaluation of eye movement, it was necessary to select essays limited to the length of approximately 350 words in order that each be shown onscreen at a single time while ensuring accurate eye-tracking measurements. We used just two essays in this study to control for length, language features, writer demographics, and essay quality within authentic student writing.

The L1 essay was 346 words long and written by a native-English speaker. She was a sophomore studying agriculture and environmental sciences who had taken two developmental writing courses prior to her first-year composition experience. The L2 essay was 339 words long and composed by a Mandarin-speaking international student in his sophomore year of college with an undeclared major. He had been studying English for approximately 13 years and had taken one semester of a developmental writing class prior to writing the essay used in this study. Essays were displayed without identifying information about the writers so that raters would not be swayed by their perceptions of the writers' names, gender, nationality, or language background.

Procedure

Once the writing samples were selected, the texts were divided into AOIs based on textual characteristics related to the four rubric categories. AOIs function as data collection points in eye-tracking research so that duration and count measurements are taken when a reader's eyes enter each AOI. Selecting meaningful AOIs across the two essays was challenging because of the obvious differences in textual structure between them. Ideally, the two texts should be matched for features, but given that L1 and L2 writing tends to differ in substantive ways, we attempted to code single and multi-word units that represented salient features of rhetoric, organization, word choice, and grammar (see Carrol & Conklin, 2014 for methodological considerations of this approach). Rhetorical phrases were marked when they related to or supported the thesis of each essay; organizational words or phrases were selected which connected ideas or transitioned among them; word choice reflected grammatically acceptable but otherwise non-standard or awkward wording; and grammatical errors were marked which violated standard grammar rules or reflected inaccurate spelling or punctuation. No AOIs overlapped across categories; that is, any word or phrase that was coded as rhetoric was not also coded as organization or any other category. Thus, each category was fully independent of the others.

Four members of the research team independently marked each essay for respective features based on the criteria above and their own judgements of salient semantic units. They then came to a unanimous consensus that resulted in the final interest area selections (see Table 2 for AOI counts and appendix B for AOIs in context). While this approach is understandably subjective, Holmqvist et. al. (2011) describe "expert-defined AOIs" (pg. 218) as a superior and more objective approach compared to a single researcher selection. Both essays were digitized so they could be displayed on the eye-tracking monitor.

Table 2. AOI Counts for L1 And L2 Essays

	L1 Text	L2 Text
Rhetoric	12	15
Organization	7	11
Word Choice	10	12
Grammar	23	17
Total	52	55

The rating rubric was similarly digitized and displayed prior to and following the display of each student's essay. Participants did not receive explicit rubric training or norming, though they did practice utilizing the rubric before the study began. Explicit group norming was omitted partially because we wanted to collect raters' unfiltered processing of the text and ensure that they reacted naturally to the text without attempting to match scores with other raters (Sakyi, 2000). Since the raters had extensive training and experience utilizing similar rubrics outside of the study's rubric, we believed they were sufficiently oriented to the rubric design, though we recognize that rubric design and rater training can always impact rater behaviors.

Upon arrival to the eye-tracking lab, we asked participants a series of demographic questions including those related to their educational background and experience teaching writing. Afterward, the participants were positioned comfortably in front of the eye-tracking monitor with their chins placed in a headrest to reduce head movement. We used an SR Research EyeLink 1000 Plus eye tracker (spatial resolution of 0.01°) sampling at 1000 Hz. Participants were positioned 63 centimeters from the computer monitor where the student essays and rubric appeared, which had a display resolution of 1600 x 900 so that approximately 3.5 characters subtended 1° of visual angle.

Participants were first calibrated on a nine-point calibration procedure so that the computer could accurately detect their eye movements and fixations. Then, on the first screen, participants viewed general instructions followed by the rubric and then an example text appeared. The example allowed participants to become accustomed to the rating process and was used to ensure further calibration of the eye tracker and limit order bias in which readers tend to be slower on the first essay.

Participants then reviewed the rubric again before reading the students' essays. After reading the text, the participants proceeded to the rubric again and provided a single verbal score for the composition based on the four areas of the rubric. Raters were asked to mentally score all areas of the rubric but report only their cumulative score. After the first essay, the procedure was repeated for the second; the essay order was randomized to further account for possible ordering effects.

Once the participants finished assessing both writing samples, they were audio-recorded during a reflective protocol about their rating and assessment process. Participants were asked what they looked for in each of the four areas on the rubric, how they approached the rating task in general, and if and how they approached the two texts differently.

Data Analysis

In order to compare rating behavior across the two essays, we chose to examine participants' early reading processes (first pass reading time and skip counts) as well as their later reading processes (total reading time and run counts) including an indication of confusion and/or cognitive processing (regression-in counts).

This data was gathered for each AOI and each participant on both essays. However, because AOIs differed in number and length, which can affect gaze duration measures as participants read longer AOIs for more time, we controlled the data by dividing all continuous data by the total number of letters and spaces in each interest area. We examined the descriptive data first and then analyzed the data using independent t-tests for first-pass and total reading time measures, Mann-Whitney U tests for the non-parametric run and regression-in counts, and Chi Square for the skip count, which presented categorical data.

In addition to analyzing the eye-tracking data, we compared raters' scores for each text. We also examined our notes and recordings of the rater interviews looking for patterns of responses using content analysis (Bengtsson, 2016; Krippendorff, 2004) in order to find differences in self-reported procedures for rating L1 and L2 texts.

Results and Discussion

Rater scores

Raters tended to rate the L1 text higher than the L2 text as seen in Table 3. These responses showed a somewhat surprising pattern given that in holistic ratings, L2 writers tend to receive higher scores than L1 writers (Rubin & Williams-James, 1997). However, scores may have been skewed in this study by the categorical structuring of the rubric that made it appear more analytic in nature. Moreover, the raters' backgrounds as trained TESOL instructors likely increased their awareness of salient L2 features of the writing, such as more prominent grammar errors and formulaic essay organization, which they may have penalized in their scoring.

Table 3. Essay Scores Across Raters

	L1 Text	L2 Text
LeDean	9	10
George	8	6
Jenna	8	7
Josh	9	7
Julie	10	8
Total	44	38
Average	8.8	7.6

Rubric Categories

To answer the first research question, we examined eye-tracking data related to participants' assessment of grammar, word choice, organization, and rhetoric of L1 and L2 writing, as well as measures of time-on-task. The time on task measures indicated that raters spent about 3.5 minutes reading and judging each essay irrespective of the writer's language background. However, analysis of the rubric category results showed some sizeable differences in raters' processing of L1 and L2 texts, indicating that raters in this study attended differently to the rhetorical, organizational, and grammatical features they read. These differences highlight behaviors that may have contributed to raters' score discrepancies. Table 4 summarizes all eye-tracking measures in our data for each rubric category [3] by displaying overall ratio data (measurements divided by the number of letters and spaces within each AOI). A further description of each category follows below.

Table 4. Ratio Eye-Tracking Measures by Rubric Category

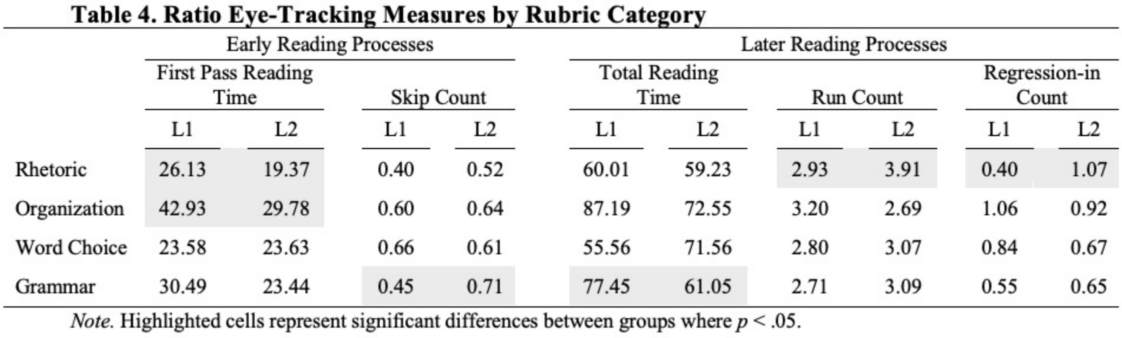

Table 4. Ratio Eye-Tracking Measures by Rubric Category

	Early Reading Processes				Later Reading Processes					
	First Pass Reading Time		Skip Count		Total Reading Time		Run Count		Regression-in Count	
	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2
Rhetoric	26.13	19.37	0.40	0.52	60.01	59.23	2.93	3.91	0.40	1.07
Organization	42.93	29.78	0.60	0.64	87.19	72.55	3.20	2.69	1.06	0.92
Word Choice	23.58	23.63	0.66	0.61	55.56	71.56	2.80	3.07	0.84	0.67
Grammar	30.49	23.44	0.45	0.71	77.45	61.05	2.71	3.09	0.55	0.65

Note. Highlighted cells represent significant differences between groups where $p < .05$.

Rhetoric. Our statistical analysis showed a significant difference for L1 and L2 texts in three rhetorical measures. The first pass reading time for rhetoric was higher for the L1 text (M = 26.13, SD = 18.49) compared to L2 (M = 19.37, SD = 11.09), $t(82) = 2.40$, $p = .018$, $d = .44$ while a Mann-Whitney test indicated that the run count and regression-in measures were both significantly higher for the L2 text as seen in table 5.

Table 5. Rhetorical Measurements

Label		N	Mdn	U	*sig.*	*d*
Run Count	L1	55	3	1628.5	.037	0.36
	ESL	75	3			
Regression Count	L1	55	0	1553.0	.011	0.44
	ESL	74	1			

Based on this, and conceptualized comprehensively in Table 4, raters appeared to spend more time in early reading processes of the L1 text but that they returned to L2 features more in later reading, a combination that seems to indicate initial decoding and effortful reading of L1 text and skimming or scanning of L2 text that gave way to increased confusion and comprehension difficulty of the L2 text.

First pass reading time is considered an early measure of reading, which is thought to relate to word recognition or lexical access (Winke & Lim, 2015). Thus, the raters seemed to spend more time decoding language associated with the rhetorical features of the L1 text either because they were more complex or required more effortful decoding. Run counts and regressions-in measure confusion and comprehension break-downs, and the measures seem to indicate that raters looked and re-looked at rhetorical features in the L2 text likely because these features were more cognitively demanding. Together, these data paint a picture of teachers scanning the rhetorical features of the L2 text quickly and then re-reading those rhetorical elements more deliberately later on. This may be partly due to expectancy bias in that the L2 writing instructors in our study may have had preconceived ideas of the rhetorical structure of an L2 text (i.e., clear argument in thesis, predictable use of evidence) and therefore scanned through quickly at first but then checked back to confirm their intuitions or further processed unclear rhetorical features in re-readings.

The L2 rhetorical structure was relatively predictable in its 5-paragraph form and had features which appeared writer-based (Flower, 1981) in that they reflected the sometimes contradictory internal thoughts of the writer. In contrast, the L1 rhetorical features illustrated some audience awareness by referencing other types of writers (e.g., "Some people can grasp") and utilizing more introductory devices, such as "for example". These differences in rhetorical style and proficiency may have triggered rater expectations that account for the differences in raters' reading behaviors.

Organization. Only one organization measurement was statistically significant, that of first pass reading time associated with decoding. Raters spent significantly more time when first fixating on L1 organizational features (M = 42.92, SD = 33.79) compared to L2 features with a medium effect size (M = 29.78, SD = 19.95), $t(46) = 2.02$, $p = .049$, $d = .47$.

The organization data tell a slightly different story than the rhetoric data in that raters appeared to process organizational features of the L1 text more in early reading. These results indicate that raters initially spent more time accessing the organizational features of the L1 text than the L2 text, which mirrors behavior in the rhetorical data. But unlike the rhetorical results, raters did not appear to re-read organizational features more in one text than another. We speculate that the longer first pass reading times were related to the cognitive difficulty of cohesive devices used in the L1 essay. The L1 essay used a more narrative structure with more complex, multi-word transitions which led from one point to another while the L2 writer, dividing his text into five paragraphs, utilized several explicit, single-word transitions between ideas that were easier to scan.

Grammar. The grammar skip count measure indicated that raters were more likely to skip L2 grammar features than L1 with a large effect size, X^2 (1, N = 200) = 12.76, $p < .001$, ϕ =.25. A Chi-Square analysis was run because skip rate is a nominal count indicating whether an AOI was skipped (1) or not (0). Complementing this, total reading time showed raters spending more time reading L1 grammatical features (M = 77.44, SD = 67.09) compared to L2 features with a small effect size (M = 61.05, SD = 39.66), $t(190) = 2.16$, $p = .032$, $d = .29$. In other words, the raters actually skipped more L2 grammar errors and then were less likely to return to them than L1 errors.

It seems almost paradoxical that TESOL professionals, who are familiar with L2 grammar errors, would overlook these features initially and instead dwell more on L1 grammar. One reason may be that raters compensated for the language acquisition process by overlooking L2 errors as has been suggested in the literature (Sakyi, 2000; Song & Caruso, 1996; Sweedler-

Brown, 1993). Additionally, raters' familiarity with these errors might also explain why they read past them in the L2 text. Eckstein and Ferris (2018) demonstrated significant differences in the kinds of errors L1 and L2 writers make at the first-year composition level, which seems to typify the essays in this study. The L2 essay, for instance, included missing words and verb errors. The L1 text, however, included punctuation and run-on sentence errors, features which are likely less often marked and therefore perhaps more salient or distracting to TESOL-trained teachers. Thus, raters in this study likely noticed typical L2 errors early in the essay and subsequently gave less heed to them while reading. This process of seeing representative grammar errors and then overlooking others seems to represent the way in which the raters performed multilingual equity (Hall, 2014) and compensated for L2 language development in their scoring.

In sum, a number of differences emerged when ESL-trained writing teachers read and assessed L1 and L2 student texts in this study. These raters read the rhetorical, organizational, and grammatical features of the L1 text more deliberately while skimming through and then returning to rhetorical features of the L2 text and skipping over many L2 grammatical structures. It is worth noting that no significant differences emerged in the word choice features of each essay. We anticipated some differences given that L2 writers often struggle with vocabulary (Crossley & McNamara, 2009). We suspect therefore that word choices were not a major factor contributing to rater score differences.

Interviews

To answer the second research question, we interviewed raters immediately following the rating task to capture their recollections and self-reported rating processes and used content analysis techniques (Bengtsson, 2016; Krippendorff, 2004) to organize their responses. All of the raters reported approaching L1 and L2 writing differently during regular essay rating outside of this study. When asked what differences they attend to, all reported varying their grammar expectations as the major modification. Josh and Jenna (all names are pseudonyms) both reported specifically being more lenient on grammar errors made by L2 writers and harsher on errors made by L1 writers.

When asked if they approached the two essays in this study differently, three raters responded that they imagined both essays to be written by non-native speakers. However, only two raters (LeDean and Jenna) reported remaining consistent in this approach after they noticed differences in the writers' organization and grammar.

We asked raters what they were looking for in terms of rhetoric, organization, word choice, and grammar. Raters had the most to say about grammar, some even listing the kinds of errors they recalled seeing in the two texts (some of which were not present in the text at all!) and listing off other errors they associated with L1 and L2 writing. This is notable given the grammar skip rate in this study and other research showing that mechanics were the least-attended to feature in essay rating (Winke & Lim, 2015). It is possible that our TESOL-trained raters were so sensitized to L2 grammar errors that such errors no longer required additional processing resources; instead, the raters overlooked these errors in L2 texts and made rating decisions based on their *perceptions* of grammar errors.

Raters also seemed consistent, though more vague, about expectations for word choice. Two raters (LeDean and Jenna) reported looking for academic words such as those from the

Academic Word List when evaluating word choice. The other raters described things such as readability, effective word choice, and precision as relevant word choice considerations.

Raters all reported looking for thesis statements and topic sentences when assessing organization. Other organizational considerations included paragraph structure, logical ordering of the essay, and coherence in general. Raters had the hardest time expressing their criteria for assessing rhetoric. Three reported the importance of responding to the prompt; otherwise, raters tended to include organizational considerations under rhetoric as well, such as clear thesis and topic sentences, well-developed paragraphs, and coherence between ideas. One rater, Jenna, mentioned audience awareness as a critical factor. Overall, raters seemed most comfortable assessing grammar and organization but were more vague about word choice and rhetoric (see Song & Caruso, 1996). We suspect that their professional training influenced these rating criteria (Cumming et al., 2001) since the TESOL profession stems from the study of linguistics, which is focused on language structures more than rhetoric (Silva & Leki, 2004).

We also noted during the study that raters consistently moved their eyes through both texts at least three times and for decreasing amounts of time for each pass. When asked about this behavior, raters explained that they used the first pass to become familiar with the text and subsequent passes to confirm or alter those perceptions, especially relative to the prompt. Raters were also genuinely surprised to learn that other raters similarly re-read the essay multiple times.

Eye Tracking and Difficulties with Writing Research

An important purpose of this research was to determine the feasibility and processes for utilizing eye-tracking methods in TESOL-based writing research. Although eye-tracking methods produced interesting results and participants were enthusiastic about the study, we encountered practical constraints that researchers are advised to consider when devising their own research of this kind.

Stimuli selection and coding was perhaps the most complicated venture in our study because we attempted to compare L1 and L2 writing. Thus, we were of necessity setting up an unequal comparison in terms of word and phrase length, word usage and complexity, style, organization, and so forth. Further we marked multi-word units for analysis rather than single words or sentences (see Carrol & Conklin, 2014 for difficulties measuring multi-word structures). It is difficult to match authentic texts sufficiently so that fixation durations on one text can be compared to durations on another. Conklin and Pellicer-Sánchez (2015) advocate for the creation of *a priori* AOIs which can be well controlled and matched for length. We attempted to match AOIs by using ratio data: dividing reading time by the number of letters in a particular word, for instance. This way, we could compare a time/letter ratio between the two texts instead of time/word, time/phrase, or time/sentence ratios. Yet we nevertheless recognize that the AOIs we chose are still open to criticism since we selected them based on expert decisions. Choosing AOIs empirically based on rater fixations (e.g., stimulus-generated AOIs or attention maps) for each rubric category as a preliminary step in a study design may be more objective (Holmqvist et. al., 2011). Otherwise, researchers have used *a posteriori* approaches by selecting each word as an AOI (Paulson, Alexander, & Armstrong, 2007) or even each sentence (Cop, Drieghe, & Kuyck, 2015), though this invites criticism if the AOIs in comparison texts are not well matched. Another option is to describe raters' reading/assessment processes of different texts and compare them abstractly or through more complex statistical operations, such as hierarchical linear modeling. In any event, we recommend that researchers comparing dissimilar or multiple texts select AOIs based on a solid empirical justification and match them across stimuli for length and possibly other features including lexical complexity and word frequency recognizing that single-word analysis may be best.

Our stimuli also proved difficult because of limited screen space. Most text-based eye-tracking experiments investigate word or sentence-level phenomena (Carrol & Conklin, 2014). Displaying continuous text on an eye-tracking screen is more complicated by virtue of text size. Text must be displayed with enough surrounding white space to make it clear where participants are looking, and thus, only short pieces of writing can usually be displayed at a time. Other researchers have compensated for this by showing multiple screens of text one after another (Cop, Drieghe, & Duyck, 2015), but in essay assessment where raters would flip through the screens of text repeatedly and may wish to go back to view an earlier part of an essay, this would substantially complicate data collection.

Eye-tracking experiment set-up and data interpretation can also be a challenge for teachers who may hope to quickly run an eye-tracking experiment. While the experiment builder program of most eye-tracking machines has a user-friendly interface, it is not especially intuitive at first, and researchers may benefit from some basic programming expertise in order to turn their experimental design into a functioning eye-tracking experiment. When analyzing data, it can also be challenging to know what dependent variables to select since eye-trackers can collect scores of AOI, interest period, pupil-size, and gaze path measures. Thus, researchers should consult methods books such as *Eye tracking: A Comprehensive Guide to Methods and Measures* (Holmqvist et al, 2011), *Eye Tracking: A Guide for Applied Linguistics Research* (Conklin et al, 2018), or *Conducting Eye Tracking Research in Second Language Acquisition and Bilingualism* (Godfroid, forthcoming) when beginning eye-tracking research.

We found that other study elements were less problematic: participants were eager to learn about their reading processes and how they might rate students differently and were also very willing to share their personal beliefs about assessment. They were curious about the eye-tracking apparatus and delighted to see and discuss their eye-movement data.

Conclusions and Limitations

Our study revealed interesting differences in the way a small number of TESOL-trained writing teachers read and assessed texts written in English by L1 and L2 students. The raters tended to read L1 texts more deliberately while initially skimming through L2 texts and even skipping grammar errors in L2 texts altogether. The self-reflection portion of our study found that the raters most agreed on features that made up the organization and grammar categories but were less agreed on features of word choice and rhetoric. This likely stemmed from their professional training and experience and may have effected their scoring behaviors, especially since raters collectively reported approaching L1 and L2 texts differently by at least offering leniency on L2 grammar errors.

Taken together, these findings suggest that within our limited pool, raters indeed adopted different rating approaches for L1 and L2 writing. Raters read the L1 text more carefully and linearly while repeatedly skimming and reviewing the L2 text. Also, in order to be lenient on grammar for L2 writers, raters appeared to observe the presence of some typical grammar errors found in the L2 text and purposefully ignore others. The differential rating approaches could be related simply to the word choice and sentence structure used in each text: more sophisticated or less common structures in L1 writing could explain slower processing, more rereading, and less skipping. On the other hand, these findings could have implications for rater judgement. For instance, raters may be more likely to overlook grammar errors and read less linearly if they observe certain textual characteristics such as typical L2 grammar errors or stylized rhetorical or organizational features within the text. More research with larger participant and essay pools is needed, however, before generalizations of any kind can be made.

We acknowledge that there are a number of limitations of this research which constrict its generalizability just to our local context. One obvious limitation is that our participant pool and sample texts were extremely limited and thus we recommend caution in generalizing findings based on this study alone; more representative L1 and L2 texts are needed in order to make broader claims about raters' interactions with a variety of L1 and L2 writing styles and types. Another limitation is that the rubric and essays never appeared together, which made it impossible to connect eye-movement data between the essay and the criteria by which it was judged, though presenting both documents together would necessarily change the type of variables that could be analyzed. In future investigations, we hope to expand our data collection by examining how composition-trained raters read and asses these same or additional L1 and L2 texts. Although the insights of TESOL-trained raters are meaningful, it is much more likely for composition-trained teachers to rate L1 and L2 writing together. Thus insights from this group will perhaps be more relevant to L1/L2 writing assessment.

The major contribution of this study lies in its novel use of eye-tracking to investigate whether TESOL teachers approach L1 and L2 writing differently and with different criteria in mind (i.e., Haswell, 2007; Lindsey and Crusan, 2011; Rubin & Williams-James, 1997). By utilizing eye-tracking methods, we can gain a greater awareness of the reading behaviors that may affect rater scores. We believe that additional and robust research of this nature that draws on observable eye-tracking measures could inform teacher and rater training and potentially lead to better writing instruction and more equitable rating procedures for both L1 and L2 writers.

About the Authors

Grant Eckstein is a professor of Linguistics at Brigham Young University. His research interests include response to student writing and second language writing development and pedagogy.

Wesley Schramm has a master's degree in TESOL from Brigham Young University where he investigated raters' approaches to assessing grammar through eye-tracking methods. His interests include teaching and assessing L2 writing.

Madeline Noxon has a bachelor's degree in Linguistics from Brigham Young University.

Jenna Snyder holds a bachelor's degree in English with a minor in TESOL. She teaches ESL classes at Brigham Young University's English Language Center and plans to continue her studies in TESOL and expand her research of writing.

References

Ai, H., & Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (eds.), *Automatic treatment and analysis of learner corpus data* (pp. 249-264). Amsterdam/Philadelphia: John Benjamins.

Adler-Kassner, L., & Wardle, E. (2015). *Naming what we know: Threshold concepts of writing studies*. University Press of Colorado.

Anson, C., & Schwegler, R. (2012). Tracking the mind's eye: A new technology for researching twenty-first-century writing and reading processes. *College Composition and Communication, 64*(1), 151-171.

Ball, A. (1997). Expanding the dialogue on culture as a critical component when assessing writing. *Assessing Writing 4*, 169–202.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54–74.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing, 28*(1), 51–75. http://doi.org/10.1177/0265532210376379

Bengtsson, M. (2016). How to plan and perform a qualitative study using content analysis. *NursingPlus Open, 2*, 8-14.

Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes whilst completing reading tasks: A mixed-method eye-tracking and stimulated recall study*. ARAGs Research Reports Online. AR, 2015/001. London: British Council.

Carrol, G., & Conklin, K. (2014). Eye-tracking multi-word units: some methodological questions. *Journal of Eye Movement Research, 7*(5), 1–11. http://doi.org/10.16910/jemr.7.5.5

Cop, U., Drieghe, D., & Duyck, W. (2015). Eye movement patterns in natural reading: A comparison of monolingual and bilingual reading of a novel. *PloS ONE, 10*(8), e0134008. doi:10.1371/journal.pone.0134008.

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

Conklin, K., Pellicer-Sánchez, A., & Carrol, G. (2018). *Eye-tracking: A guide for applied linguistics research*. NY: Cambridge University Press. doi: 10.1017/9781108233279

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. http://doi.org/10.1016/j.jslw.2009.02.002

Crossley, S. A., Kyle, K., Allen, L. K., Guo, L., & McNamara, D. S. (2014). Linguistic microfeatures to predict L2 writing proficiency: A case study in Automated Writing Evaluation. *The Journal of Writing Assessment, 7*(1), 1–16.

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

Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework*. (TOEFL Monograph Series N 22). Princeton, NJ: Educational Testing Service.

DeRemer, M. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, 5*, 7–29.

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

Eckstein, G., & Ferris, D. R. (2018). Comparing L1 and L2 texts and writers in first-year composition. *TESOL Quarterly, 52*(1), 137–162. http://doi.org/10.1002/tesq.376

Edgington, A. (2005). "What are you thinking" Understanding teacher reading and response through a protocol analysis study. *Journal of Writing Assessment, 2*(2), 125–145.

Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions*. TOEFL Research Report RR-03-17. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-03-17.pdf

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning & Verbal Behavior, 20*, 641-655.

Flower, L. S. (1981). Revising writer-based prose. *Journal of Basic Writing, 3*(3), 62-74.

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.

Godfroid, A. (2018, March). One tool, many applications: Robust eye-tracking research across SLA disciplines. Paper presented at the meeting of the American Association of Applied Linguistics, Chicago, IL.

Godfroid, A. (Forthcoming). Conducting eye tracking research in second language acquisition and bilingualism. NY: Routledge.

Godfroid, A. & Spino, L. (2015). Reconceptualizing reactivity of think-alouds and eye tracking: Absence of evidence is not evidence of absence. *Language Learning, 65*(4), 896-928.

Godfroid, A., & Winke, P. (2015). Investigating implicit and explicit processing using L2 learners' eye-movement data. In P. Rebuschat (Ed.), *Implicit and explicit learning* (pp. 325–348). Philadelphia, PA: John Benjamins.

Hall, J. (2014). Language background and the college writing course. *Journal of Writing Assessment, 7*(1). Retrieved from http://journalofwritingassessment.org/article.php?article=77

Haswell, R.H. (2007). *Researching teacher evaluation of second language writing via prototype theory*. Retrieved from http://www.writing.ucsb.edu/wrconf08/Pdf_ Articles/ Haswell-Article.pdf

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

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van De Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford: Oxford University Press.

Huang, J. (2009). Factors affecting the assessment of ESL students' writing. *International Journal of Applied Educational Studies, 5*(1), 1-17.

Huot, B. (1993). The influence of holistic scoring procedures on reading and rating student writing. In M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 206-236). Cresskill, NJ: Hampton Press Inc.

Huot, B. (2002). *(Re)Articulating writing assessment for teaching and learning*. Logan: Utah State University Press.

Huot, B., O'Neill, P., & Moore, C. (2010). A usable past for writing assessment. *College English, 72*(5), 495–517.

Institute of International Education. (2016). *Open doors 2016 fast facts*. Retrieved from www.iie.org/-/media/Files/Corporate/Open-Doors/Fast-Facts/Fast-Facts-2016.ashx?la=en&hash=9E918FD139768E1631E06A3C280D8A9F2F22BBE1

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*, 130-144.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275–304.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Thousand Oaks, California: Sage Publications.

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing, 31*(3), 329–348. http://doi.org/10.1177/0265532214526174

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

Lindsey, P., & Crusan, D. (2011). How faculty attitudes and expectations toward student nationality affect writing assessment. *Cross the Disciplines, 8*(4), 1-41. Retrieved from http://wac.colostate.edu/atd/ell/lindsey-crusan.cfm

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. New York: Peter Lang.

McDaniel, B.A. (1985). *Ratings vs. equity in the evaluation of writing*. Paper presented at the 36th Annual Conference on College Composition and Communication, Minneapolis, MN.

Nation, P. (2009). Reading faster. *International Journal of English Studies, 9*(2), 131-144.

NCTE. (2009). Assessment must be fair and equitable. *Standards for the assessment of reading and writing*. Retrieved from http://www.ncte.org/standards/assessmentstandards

Perkins, K. (1980). Using objective methods of attained writing proficiency to discriminate among holistic evaluations. *TESOL Quarterly, 14*, 61-69.

Pula, J., & Huot, B. (1993). A model of background influences on holistic raters. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237–265). Cresskill, NJ: Hampton Press.

Paulson, E. J., 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.

Rafoth, B. A., and Rubin, D. L. (1984). The impact of content and mechanics on judgements of writing quality. *Written Communication, 1*, 446-458.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422.

Rayner, K. (2009). Eye movements in reading: Models and data. *Journal of Eye Movement Research, 2*(5), 1–10.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading*. Hillsdale, NJ: Erlbaum.

Rayner, K., Sereno, S., Morris, R., Schmauder, R., & Clifton, C. (1989). Eye movements and on-line language comprehension processes. *Language and Cognitive Processes, 4*(3-4), SI21-SI49.

Reichle, E. D., Pollatsek, A., & Rayner, K. (2006). E-Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research, 7*(1), 4–22.

Roberts, F., & Cimasko, T. (2008). Evaluating ESL: Making sense of university professors' responses to second language writing. *Journal of Second Language Writing, 17*(3), 125–143. http://doi.org/10.1016/j.jslw.2007.10.002

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. http://doi.org/10.1016/S1060-3743(97)90031-X

Sakyi, A. (2000). Validation of holistic scoring for ESL writing assessment: A study of how raters evaluate ESL compositions on a holistic scale. In A. J. Kunnan (ed.), *Fairness and validation in language assessment* (pp. 130–153). Cambridge: Cambridge University Press.

Silva, T., & Leki, I. (2004). Family matters: The influence of applied linguistics and composition studies on second language writing studies–past, present, and future. *The Modern Language Journal, 88*(1), 1–13.

Singer, N. R., & LeMahieu, P. (2011). The effect of scoring order on the independence of holistic and analytic scores. *The Journal of Writing Assessment, 4*, 1-13.

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.

Sparks, J. (1988). Using objective measures of attained writing proficiency to discriminate among holistic evaluations. *TESOL Journal, 9*, 35-49.

Stratman, J. F., & Hamp-Lyons, L. (1994). Reactivity in concurrent think-aloud protocols: Issues for research. In P. Smagorinsky (ed.), *Speaking about writing: Reflections on research methodology* (pp. 89–111). Thousand Oaks, CA: Sage.

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

Van Gompel, R. P. (Ed.). (2007). *Eye movements: A window on mind and brain*. Elsevier.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (ed.), *Assessing second language writing in academic contexts* (pp. 111–125). Norwood, NJ: Ablex.

Wolcott, W., & Legg, S. M. (1998). *An overview of writing assessment: Theory, research and practice*. Urbana, IL: National Council of Teachers of English.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing, 25*, 37–53. http://doi.org/10.1016/j.asw.2015.05.002

Wolfe, E. W. (2005). Uncovering rater's cognitive processing and focus using think-aloud protocols. *Journal of Writing Assessment, 2*(1), 37–56.

Wolfe, E. W., & Ranney, M. (1996). Expertise in essay scoring. In D.C. Edelson & E.A. Domeshek (Eds.), Proceedings of ICLS 96 (pp. 545-550). Charlottesville, VA: Association for the Advancement of Computing in Education.

Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing, 27*, 37-53.

Notes

[1] We use the terms "L2 writing" and "L2 texts" in this essay to refer to texts written in English by a non-native English speaker. "L1 writing" and "L1 text" refers to texts written by native-English speakers. [back]

[2] For a copy of the full essay prompt and original scoring rubric, contact the first author. [back]

[3] In total, there are five matched L1/L2 comparisons for each rubric category. One reviewer recommended adjusting the experiment-wise alpha level to account for the possibility of type I errors, but given that each of the variables included independent data sets, several statisticians who consulted on this project deemed this procedure overly restrictive. [back]

Appendix A. Scoring Rubric

Scoring Criteria	**Possible**
Rhetoric Consider the following • Clarity of overall message and purpose • Sophistication of support and elaboration • Sense of audience awareness • Control of voice	**1 2 3**
Organization Consider the following • Cohesiveness of the whole text • Effectiveness of paragraph focus • Logical sequencing of ideas • Efficacy of transitions	**1 2 3**
Word Choice • Correctness of word choice • Sophistication of word choice • Variety of vocabulary	**1 2 3**
Grammar • Structure and coherence of sentences • Accuracy of grammar: • Verb tenses and agreement • Word forms, word order • Prepositions, articles • Mechanics: Punctuation, capitalization, spelling	**1 2 3**
Total Score	**___ / 12**

Appendix B. Student Essays and Interest Area Codes

Student 1 (L1 Text)

Prompt: Explain your background and process as a writer.

I am a good writer in the sense of thinking critically in my essay. When I start off my essays it's a habit for me to write down onto paper first and then later on transfer that information onto the computer. The reason I do this is because I can concentrate better without distraction and I am actually thinking about my essay and no other things on the web. So by doing this first I am able to brainstorm more ideas. Sometime when I start off my essay writing on my computer, I have the intention or interest to search the web. When I do this I can no longer concentrate on my essay. As a good writer I also need time. Everyone pace is different when writing essays. Some people can grasp the essay topic and ideas right after the instructor gives it out and some students (like me) take a longer time to process the information down and then start writing. I am not a good writer in the sense of making my paper the best but what I contribute to my essays.

It is always hard for me to narrow my ideas into a shorter sentence so for that I am a detail writer. I like giving and going into depth in my writing. When there is an idea that I really want to express but cannot find a word for it I like to describe what it is. For example, if I wanted to talk about my family and finding a specific vocabulary word to describe them would be impossible. In my opinion one word cannot describe my family. If I have the choice to pick multiple of words to describe them it would be a long one. But by describing my family in a lot of words can help the reader to have an idea of how I am describing my family. In my writing I try to make the reader understand what I am writing about and that can work sometime and other time it may not go as planned.

Rhetoric 12
Organization 7
Word Choice 10
Grammar 23

Student 2 (L2 Text)

Prompt: Explain your background and process as a writer.

When I just came to college, the only thing I knew about writing is fallowing the "SAT writing formula," which was giving a position with two examples. I repeated the same sentences many times in my essay because I do not how to write it in other ways. I did not have enough powerful examples for my view because I did not know much about English lectures. Sometimes, I could only get one example for my discussion. Like my reread essay, "Money and Happiness," I only gave one example, which was a piece of Chinese history. Therefore, writing is really difficult for me.

I think I have to rewrite more times than others do to get my essay better. However, I do not know how to rewrite an essay. It is because I can make my words more vivid in Chinese, but cannot make them in English. I cannot give a more powerful example and discussion with my poor words by rewriting. I think this is the reason I am not a good writer and I do not really like my writing.

However, I like writing. Writing is telling a story to your readers. I have a lot of interesting and meaningful story, and I enjoy sharing them with people, and people will stand on your position in the story. I will not alone in my story any more. I like writing and sharing.

I hope this class could teach us how to rewrite our essays with wonderful words, sentences, and paragraph. It is because only few people can write an excellent writing in one time. Rewriting is one of the best ways to get excellent writing. Moreover, I hope I can get better in my grammar skill. It is the basic skill to be a good writer.

All in all, I am not a good writer. Writing is really difficult for me. However, I like writing very much. I like sharing my stories with more people by writing. I hope this class could get me better in writing.

Rhetoric 15
Organization 11
Word Choice 12
Grammar 17