Assessment of L2 Student Writing:
Does Teacher Disciplinary Background Matter?

Grant Eckstein, Rachel Casper, Jacob Chan and Logan Blackwell

Brigham Young University, UT | USA

Abstract: This preliminary study examines the rating behavior of five composition and five ESL writing teachers while evaluating a text from a university-level non-native (L2) English speaking student. Using an eye tracker, we measured raters' dwell times and reading behaviors across four areas of interest—rhetoric, organization, vocabulary, and grammar. Results indicate that raters with differing disciplinary backgrounds read the text differently. L2 writing teachers tended to spend more time on and re-read the rhetorical, lexical, and grammatical features of the text while skipping over more of the grammar errors, while composition teachers read the text more deliberately. The findings suggest L2 writing teachers were more prone to skim and scan for information on which to base a grade while composition teachers delayed rating decisions until after reviewing the entire text, which is corroborated in previous research. These findings can expand our understanding of how disciplinary background can influence rating processes, which can inform rater training procedures, especially in disciplinary writing contexts where L2 writing is judged by individuals with and without expertise in composition or second language writing. Moreover, it demonstrates the utility of eye-tracking methods to examine the cognitive processes associated with reading and scoring student writing.

Keywords: Eye-tracking; Composition; L2; Writing and Assessment; Cognitive process

journal of WRITING RESEARCH

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. http://dx.doi.org/10.17239/jowr-2018.10.01.01

Contact: Grant Eckstein, Brigham Young University, 4071 JFSB, Provo, UT 84602 | USA – grant_eckstein@byu.edu

Scholars in both composition studies and applied linguistics acknowledge that their two disciplines have evolved from distinct epistemologies. Silva and Leki (2004) trace these differences to Grecian times, explaining that modern composition studies took theories from the Greek concept of rhetoric which originally focused on types of speeches, audience appeals, and steps in the rhetorical process. They further argue that applied linguistics is inspired by the early Greek investigation of language, which focused almost exclusively on words, their meanings, and grammar.

While rhetoric and linguistics were originally focused on the analysis and construction of oral text, these disciplines now also include written textual analysis and construction (Leki, 2000; Silva & Leki, 2004). In fact, the modern teaching of composition, specifically in the U.S., dates back only to the late 1800s[1] when, at Harvard College, the administration insisted that students be instructed on the rhetorical merits of literary classics and contemporary works in order to reproduce the style and substance in their own communication (Brereton, 1996). On the other hand, the professionalized teaching of writing to second language learners by applied linguists, again in a U.S. context, began only as recently as the 1940s (Matsuda, 1999; Matsuda, 2006a) when the end of World War II led to an increase in immigrants studying in U.S. institutions (Ferris, 2009). It was not until about the 1980s that second language (L2) writing evolved into a legitimate academic field (Matsuda, 2006a).

Although these disciplines support one another and ostensibly should work hand-in-hand, a disciplinary divide between composition for native and non-native writers developed, according to Matsuda (1999), as compositionists and applied linguistics mutually argued for the establishment of independent, specialized programs for second-language writers wherein teachers trained in linguistic analysis could apply their knowledge to students' language problems. At the time, this was regarded as a win-win situation, but in retrospect, this divide created an environment where compositionists and second language writing professionals rarely communicate (Ferris, 2009). Not only do they have different professional organizations, publications, and political orientations (Silva & Leki, 2004), they have different citation styles and even different disciplinary vocabularies (Costine & Hyon, 2011). Thus the general sub-fields of composition (teaching native English writers) and L2 writing (teaching non-native English writers) provide a useable framework for distinguishing two rather different epistemological paradigms of writing instruction and evaluation, particularly in a North American writing setting.

1. Disciplinary Differences

While both compositionists and L2 writing specialists have similar end goals[2] (e.g., teaching students to become effective English academic writers) and even use many similar practices (Eckstein, McCollum & Chariton, 2011), a number of factors contribute to their distinct approaches to writing instruction and assessment. One factor is that compositionists generally expect to teach students whose native language is English and

whose familiarity with a hegemonic, contemporary Western culture can be assumed (Atkinson & Ramanathan, 1995), even if such a common cultural framework is a myth (Matsuda, 2006b). L2 writing specialists, on the other hand, tend to work mostly with non-native or multilingual English writers who sometimes need fundamental language instruction in addition to writing support (Ferris, Eckstein & DeHonde, 2017; Eckstein & Ferris, 2018) and who cannot be assumed to share a hegemonic western cultural understanding.

In a particularly salient investigation of composition and L2 writing practices and paradigms in a large U.S. university, Atkinson and Ramanathan (1995) contrasted a University Composition Program with an English Language Program for matriculated, developmental, non-native English writers. In their ethnographic comparison, the researchers identified numerous cultural differences between the two programs, which they attributed fundamentally to the epistemology associated with each (rhetoric vs. applied linguistics). The composition program was characterized by an assumption that students shared Western cultural knowledge (i.e., values of critical thinking, originality, creativity, logic, and rationality), that writing should be learned developmentally by "constantly striv[ing] for greater 'depth' in…thought and writing" (p. 559), and that formulaic writing (i.e, the 5-paragraph essay) should be rejected. Meanwhile, the English language program was characterized by an assumption that students did not share Western cultural knowledge, that writing should be learned through strategic instruction of immediately useful study and writing skills, and that deductive/formulaic writing was appropriate for language learners, even if it did appear stifled. In other words, the composition approach saw writers themselves as the primary source of writing enlightenment (an inductive approach) whereas the L2 system saw teachers as the provider of writing development (a deductive approach). Other researchers have more recently contrasted composition and L2 writing paradigms. For instance, Costino and Hyon (2011) highlighted differing disciplinary vocabularies by explaining that compositionists value (and L2 writing specialists eschew) such terms as *critical pedagogy*, *power*, and *ideology* while L2 specialists value the terms *skill* and *practice* to the chagrin of compositionists. These terms serve to reinforce a composition view of writing-as-unified-thought compared to an L2 view of writing-as-discrete-skill.

Not all compositionists and L2 writing specialists see the historical divide in their labor quite so starkly. Leki (2000, p. 99) noted that some writing instructors of L2 students readily aligned themselves with an applied linguistic identity, but others rebuffed and denounced applied linguistics as "pointlessly aspiring to be scientific." Ferris (2003) argued that some theorists have embraced an application of composition praxis to L2 writing instruction while others have been critical of it because of the unique characteristics of L2 learners. More recently, developments in genre pedagogies (Bawarshi & Reiff; 2010; Ramanathan & Kaplan, 2000) and translingualism (Canagarajah, 2016) sample from both composition studies and applied linguistics, suggesting a move toward reconciling these divided fields. Yet disciplinary differences in ideology still persist and function as filters to restrict full integration (Santos, 1992;

Silva & Leki, 2004). A translingual approach in composition studies, for instance, has been criticized by L2 writing specialists as simplifying the needs of L2 writers and failing to acknowledge theoretical and empirical developments in L2 writing (Atkinson, et. al, 2015; Matsuda, 2014). Thus, although some practitioners and theorists have ignored the composition/L2 writing divide and others have sampled from both ideologies, there nevertheless appear to be legitimate practical, cultural, and theoretical differences that separate the two fields.

2. Disciplinary Background and Writing Evaluation

Beyond the culture and instruction of writing classes, disciplinary backgrounds also appear to impact the way that writing teachers evaluate a piece of performance-based student writing. Composition theories support an evaluation process in which raters read an entire text, form a mental image in their minds of that text, and then compare that image to an established rubric (Edgington, 2005; Wolfe, 2005). Further evidence suggests that experienced raters withhold their judgements of a text until they have finished reading it; whereas, less-experienced raters make early judgements and revise these as they encounter evidence during the reading task (Wolfe, 2005). Cumming, Kantor, and Powers (2001), however, found that ESL/EFL raters approached decision making differently from composition teachers by utilizing a process of "step-by-step reporting and progressive decision making" (p. 39).

Despite the assumed assessment and instructional differences of writing teachers as described above, researchers have not observed consistent differences in actual essay evaluation based on teachers' disciplinary differences. Presumably, composition teachers should focus more on holistic rhetorical and content features while L2 writing instructors should focus more on organization and grammar. And indeed Song and Caruso (1996) seemed to find this when they directly compared the scores assigned by 32 English and 30 ESL professors on two native and two non-native texts. Their results showed that on holistic ratings, the English faculty rated all essays significantly higher than the ESL faculty, and that the English faculty appeared more focused on rhetorical features than language features. Furthermore, Cumming, Kantor, and Powers (2002) found that composition teachers tended to balance their evaluations on ideas, argumentation, and language use in student essays while ESL/EFL instructors were more attentive to students' language use. In contrast, Santos (1988) found that professors rated content lower than language when he investigated the ratings of 178 professors without specific L2 writing training on two non-native student texts. Similarly, Brown (1991) compared the ratings of 8 English and 8 ESL teachers on 112 student compositions divided equally by native and non-native writers. While his results showed no significant differences in professors' scores, he did find that disciplinary background affected the way in which professors arrived at their scores. Namely, English faculty focused more on cohesion, syntax, and mechanics while the ESL faculty attended more to organization and content. Combined, these findings seems to indicate

that disciplinary background may play a role in writing evaluation, but that clear-cut and consistent differences have yet to emerge through empirical investigation. Given the differences in disciplinary epistemologies which are assumed to lead to differing instructional and assessment practices, the lack of consistent findings suggests a need for additional, and perhaps more fine-grained, approaches to investigating disciplinary differences in writing assessment.

A major similarity of many prior studies is that researchers investigated teacher cognition by recording teacher ratings and self-reports. While these approaches certainly reflect the thinking of disciplinary raters, a more sophisticated approach utilizing eye-tracking technology has the potential to measure the reading behaviors associated with rater decisions to better identify areas in which teachers differentially interact with student writing. Eye-trackers are also thought to limit construct interference, particularly in contrast to think-aloud protocols (Godfroid & Spino, 2015). Although eye-tracking methodologies have long been applied to reading studies (see Rayner, 1998), they are still emerging as a methodological approach within writing research (see Polio & Friedman, 2016) and student essay evaluation (Winke & Lim, 2015). Eye-trackers are used to measure eye movements including fixations and saccades (tiny eye movements between fixations) that reflect attentional focus (Conklin & Pellicer-Sancez, 2016; Rayner, 1998, 2009). That attentional focus certainly reflects reading behavior but is also thought to indicate cognition to some extent since individuals tend to cognitively process those things to which they give attention (Rayner, 2009; Reichle, Pollatsek, & Rayner, 2006).

A better understanding through eye-tracking of teachers' reading behaviors during the evaluation of student writing is important because it can be used to improve teacher evaluation and ultimately feedback practices. Furthermore, the results of an exploratory eye-tracking study of writing raters can justify further research into the approaches of writing teachers who assign, evaluate, comment on, and assess the writing of native and non-native English writers across the curriculum. More research of this nature may lead to a clearer understanding of how disciplinary background affects assessment practices and ultimately more equitable writing instruction and evaluation for all students. The present, exploratory study was thus informed by the following research questions:

1. How do composition and L2 writing teachers differ in their reading of L2 texts in terms of grammatical, lexical, organizational, and rhetorical features of the writing?
2. What features do composition and L2 writing teachers review more when reading and assessing L2 student texts?

3. Methods

3.1 Participants

Ten teachers (5 composition and 5 L2 writing specialists) participated in the study. Their average age was 27 years old, and all were either enrolled in or had recently graduated from a master's degree program in their respective fields (English and TESOL). Among the composition teachers, 3 were male and 2 female, and they had an average of three years of experience teaching mainly a required first year composition (FYC) course at a large western university in the U.S. As part of their master's degree program, they took a semester-long teacher training course concurrent with their first semester teaching in which they practiced reading, responding to, and evaluating student writing. They further reported receiving training on essay response and evaluation through a one-week orientation to teaching FYC prior to each year of teaching and weekly 50-minute in-service meetings during each semester.

Of the L2 writing teachers, 2 were male and 3 female and also averaged 27 years of age. They had between four to sixteen years of ESL teaching experience. They were all employed at an intensive English program associated with the same university as the composition teachers. The L2 writing teachers all had taken several graduate-level courses involving essay rating and assessment training. They also received extensive rater training and regular norming on both holistic and analytic essay rubrics at least three times per year.

3.2 Instruments

Essays

For this study, the stimuli presented to the participants was a single essay written by an international student during the first week of his college writing course. The writer was a native Mandarin-speaking student who had been studying English for approximately 13 years. He was in his sophomore year with an undeclared major and had taken one developmental writing class before the course. The class in which he wrote the essay used in this study was a ten-week first-year composition course held at a four-year university in the western U.S. During a 60 minute time period, the student was asked to respond to the prompt "Explain your background and process as a writer" after having been primed on the topic as part of a homework assignment. He completed the essay in class at a computer lab where he uploaded his draft to the course management system without an opportunity for subsequent revision. The essay was 339 words long and written in a 5-paragraph format, though this was not an assignment requirement. This particular essay was chosen because it included aspects of typical L2 writing including characteristic grammar errors, formulaic organization, and simplistic/repetitive vocabulary. Moreover, it had previously been reliably rated by two expert raters (Ferris,

Eckstein & DeHonde, 2017), and its relatively short size made it amenable to our eye-tracking study since we were constrained by display screen space.

Rubric

A "semi-structured" holistic rubric was used for the rating for this study (see Appendix A). The rubric was divided into four areas: rhetoric, organization, word choice, and grammar. Because research has shown that raters vary in their scoring only minimally when using an analytic rubric, a holistic rubric was adopted to better examine the differences between composition and L2 writing teachers. While the raters did not provide individual scores for each of the four categories, they were asked to think over the characteristics in each category of the rubric, determine how well they corresponded to the essays, and generate a mental score. They then orally provided a single, final score (4-12) at the completion of the assessment task. The rubric was designed by the researchers to resemble similar rubrics that were familiar to both groups of teachers. The researchers further piloted the rubric on several essay samples, adjusted the wording for clarity, and then digitized it for integration with the eye-tracker.

Procedure

The essay was first coded into areas of interest (AOIs) that could be programmed into the eye-tracking software. AOIs are predefined, semantically meaningful sections from which eye-tracking software takes measurements (Conklin & Pellicer-Sancez, 2016). We coded the essay into four types of AOIs: (a) rhetorical phrases that related to or supported the thesis of the essay; (a) organizational words and phrases that connected or transitioned among ideas; (c) lexical phrases in which wording was awkward but not strictly a grammatical error; (d) grammatical errors which included ungrammatical words or phrases as well as spelling or punctuation mistakes. Each member of the research team coded the essay individually, and then a group consensus was reached after discussing the codes. The essay and its corresponding AOIs were then uploaded into the eye tracking software (see Appendix B for essay and AOI codes).

The rubric was displayed on the eye-tracking computer screen for the participants to look over for any duration of time prior to and immediately following each essay. The participants did not receive any training on the rubric, but they did receive the rubric electronically prior to arriving at the lab.

When the participants arrived in the lab, they were asked demographic questions relating to their education and language experience and were then seated in front of the eye tracker. The machine used in this study was an SR Research EyeLink 1000 Plus (spatial resolution of 0.01°) sampling at 1000 Hz, requiring head stabilization throughout the experiment. A computer screen with a display resolution of 1600 x 900 (approximately 3.5 characters subtended 1° of visual angle) displayed the essays and rubric and was positioned 63 centimeters from the participants. The research assistant then calibrated and validated the eye tracking camera for each participant individually.

At this point, the participants could advance the program at their own rate to allow them adequate time to read, review, and rate the essay.

The participants were first presented with an instruction screen informing them of how the study would proceed. Verbal directions from the research assistant were also supplied if further explanation was required. Participants then completed a practice trial starting with the rubric, then a practice text, and then the rubric again in order to become accustomed to the task. They then proceeded with the L2 essay trial, after which the participants were asked to give a rating to the essay according to the rubric; a research assistant recorded the given score. Participants were not informed of the nationality, background, or linguistic status of the essay writer. Following the rating trial, the participants were asked about their experiences teaching writing and assessing compositions in general and about approaching the trial essay and utilizing the trial rubric in particular.

Eye Tracking Measures

The following measures were used in interpreting the eye-movement data we collected from our participants. All time durations were measured in milliseconds:

Early reading measures

- *First run dwell time:* the total time of all fixations during just the first pass of an AOI
- *Skip count:* an indication of whether a fixation occurred in an AOI during first-pass reading

Late reading measure

- *Total dwell time*: the total time of all fixations across all passes of an AOI
- *Run count:* the number of times a participant's gaze entered and left an AOI irrespective of which direction the gaze originated
- *Fixation count:* the number of discrete fixations across all runs within an AOI
- *Regression-in count:* number of times that an AOI was entered into from a later part of the sentence (e.g., looking back at a previous AOI)

The literature on eye-tracking measures distinguishes late from early reading measures (Inhoff, 1984; Staub & Rayner, 2007) in which, according to Conklin & Pellicer-Sanchez (2016), "early measures tap into automatic processes and the initial stages of processing" (p. 455) which include lexical access and text decoding. Later reading measures represent "strategic processing and include revisits and reanalysis that result from processing difficulty" (Conklin & Pellicer-Sanchez, 2016, p. 445) and therefore include processes associated with comprehension, integration, and evaluation. Thus we included measures associated both with early and late reading processes to determine how teachers automatically decode or process text and then subsequently comprehend, integrate, and evaluate it.

Data Analysis

We collected holistic essay ratings from all participants and compared raw scores. We then analyzed the eye-tracking data by arranging all AOI data into their corresponding rubric categories and matching these across teaching background, resulting in eight sets of data. We then used non-parametric Wilcoxon Signed-Rank tests to analyze the matched data with participants' dwell times (reported in milliseconds) and run counts as dependent variables in each test. Non-parametric (and associated median scores) were used in reporting the results because of the relatively small sample sizes and non-normal distribution of the count data. Given the number of independent analyses, we used a Bonferroni corrected alpha of .01 (.05/5 analyses) to reject the null hypothesis. Finally, the skip count, which resulted in nominal data (1=skipped, 0=not skipped) was analyzed using a Chi-Square test.

4. Results

4.1 Holistic Scores

The holistic scores showed some descriptive variation between the two groups of teachers as seen in Table 1. The maximum value per teacher was 12 points, yet the mean score among composition teachers was 6, and the mean for L2 writing teachers was 7.6, indicating that L2 writing teachers were generally more lenient raters, though this did not bear out statistically ($Z = 1.23$, $p = .216$, $r = 39$).

Table 1: Holistic Scores by Compositionists (L1) and L2 Writing Specialists

Rater	L1	L2
1	5	10
2	8	6
3	5	7
4	6	7
5	6	8
Mean	6.0	7.6
Median	6	7

4.2 Rubric Categories

Beyond holistic scores, we saw numerous differences in the reading behavior of teachers across discipline in all four rubric categories. The results of these analyses as presented below and arranged by rubric category.

Rhetoric

As can be seen in Table 2, which shows the results of individual Wilcoxon Signed-Rank tests for each dependent variable, the composition and L2 writing teachers read the

same rhetorical features in quantifiably different ways. For instance, the L2 teachers had more or longer interactions with AOIs than the composition teachers with medium to large effect sizes[3]. The obvious difference from this trend being first run dwell time, which measures early reading processes such as text decoding, and showed that composition teachers had longer median times than L2 teachers.

Table 2: Disciplinary differences in reading behavior for Rhetorical Structures

	Teacher	n	median	*Z*	sig.	*r*
First Run Dwell Time	Comp	74	637.00	3.33	< .001*	0.39
	L2	74	374.50			
Total Dwell Time	Comp	75	992.00	3.65	< .001*	0.42
	L2	75	1187.00			
Fixation Count	Comp	75	5.00	3.40	< .001*	0.39
	L2	75	6.00			
Run Count	Comp	75	2.00	5.29	.001*	0.61
	L2	75	3.00			
Regression-in Count	Comp	74	0.00	3.91	< .001*	0.46
	L2	74	1.00			

* $p < .01$

We also investigated initial skip count, which measures whether a participant skipped a particular AOI during initial reading. Since participants may have gone back and re-read areas skipped initially, skip count reflects the degree to which readers skim a text during initial reading. Descriptive measures showed an obvious difference in that composition teachers skipped 28% of rhetorical AOIs on initial reading while L2 writing teachers skipped 52% of them. A Chi-Square analysis showed that the L2 writing skip rate was significantly higher than the composition skip rate X^2 (1, N = 150) = 9, $p < .003$, $\varphi = .24$, a small to medium effect size.

Organization

The organization results are less pronounced than those for rhetoric. As displayed in Table 3, L2 teachers had longer total dwell times and higher run counts while composition teachers had longer first-run dwell times. Regression-in counts, which signal re-reading, were not significant in the organization category. Initial skip count did not differ significantly between teacher groups either.

Table 3: Disciplinary Differences in Reading Behavior for Organization

	Teacher	n	median	*Z*	sig.	*r*
First Run Dwell Time	Comp	51	274.00	0.87	.381	0.13
	L2	52	224.00			
Total Dwell Time	Comp	55	487.00	2.95	.003*	0.40
	L2	55	600.00			
Fixation Count	Comp	55	2.00	2.61	.009*	0.35
	L2	55	3.00			
Run Count	Comp	55	1.00	3.70	< .001*	0.50
	L2	55	2.00			
Regression-in Count	Comp	51	0.00	1.45	.142	0.21
	L2	52	0.00			

* $p < .01$

Word Choice

L2 writing teachers only had significantly higher fixation and run counts on word choice measures as seen in Table 4. Effect sizes similarly spanned medium to large. As with organization measures, first run dwell time did not significantly distinguish teacher groups, though L2 writing teachers had a higher initial skip count than composition raters X^2 (1, N = 109) = 4.858, $p < .028$, $\varphi = .21$, a small effect size.

Table 4: Disciplinary Differences in Reading Behavior for Word Choice

	Teacher	n	median	*Z*	sig.	*r*
First Run Dwell Time	Comp	54	364.50	1.65	.099	0.23
	L2	54	301.00			
Total Dwell Time	Comp	55	587.00	2.42	.016	0.33
	L2	54	769.50			
Fixation Count	Comp	55	3.00	2.63	.008*	0.36
	L2	54	4.00			
Run Count	Comp	55	2.00	4.13	< .001*	0.56
	L2	54	2.50			
Regression-in Count	Comp	54	0.00	2.10	.036	0.29
	L2	54	0.00			

* $p < .01$

Grammar

The grammar results showed significant run count differences for teacher type as seen in Table 5. As with rhetoric, composition teachers had significantly longer first run dwell times, the only measure, across categories, where composition teachers had higher median scores than L2 writing teachers. Results also showed that L2 writing teachers initially skipped a surprising 70% of all grammar AOIs (60 out of 85) whereas composition teachers initially skipped less than half (40 out of 85), a result that showed significant difference X^2 (1, N = 170) = 9.714, $p < .002$, $\varphi = .23$.

Table 5: Disciplinary Differences in Reading Behavior for Grammar

	Teacher	n	median	Z	sig.	*r*
First Run Dwell Time	Comp	83	373.00	4.63	< .001*	0.51
	L2	83	209.00			
Total Dwell Time	Comp	85	741.00	1.01	.315	0.11
	L2	85	631.00			
Fixation Count	Comp	85	3.00	0.09	.932	0.01
	L2	85	3.00			
Run Count	Comp	85	2.00	3.55	< .001*	0.39
	L2	85	3.00			
Regression-in Count	Comp	83	0.00	0.41	.682	0.05
	L2	83	0.00	6.04		

* $p < .01$

In addition to these results, we also calculated total trial time (in milliseconds) per rubric category. Results in Table 6 show a consistent pattern of higher dwell times among the L2 writing teachers except in the category of grammar, where composition teachers had longer times.

Table 6: Disciplinary Differences in Average Reading Time in Milliseconds

	Composition		L2 Teachers	
	Average dwell times	% of AOI dwell	Average dwell times	% of AOI dwell
Rhetoric	16306	37%	22289	41%
Organization	5521	13%	7928	14%
Word Choice	7541	17%	11448	21%
Grammar	14700	33%	13203	24%
Total AOI Dwell Time	44068		54869	

Taken together, results from our analyses showed quantifiable differences in the way that the composition and L2 teachers read the same student text. The L2 teachers had higher total dwell times and run counts in all rubric categories with the exception of grammar; whereas, composition teachers had higher first run dwell times (though only rhetoric and grammar were significant), and initially skipped significantly fewer AOIs in all categories but organization.

4.3 Interviews

When asked about their general approaches for reading and rating essays from L1 and L2 writers, L2 teachers reported that they were more lenient on grammar for L2 writers. Most composition teachers similarly agreed. We then asked teachers to explain how they approached the rating task just completed, particularly in terms of features that informed their perceptions and scores on the rhetorical, organizational, lexical, and grammatical components of the L2 essay.

In the area of rhetoric, L2 writing teachers struggled to describe how they rated it. They mentioned a variety of items they looked for including audience awareness, cohesion, sense of the text, and relevance to the prompt, irrespective of whether these items appeared on the rubric. Further, some teachers listed criteria that dealt with organization as criteria for rhetoric, such as Josh[4], who stated that he looked for thesis statements and topic sentences when responding about rhetoric. In the area of word choice, several L2 teachers specifically said they looked for academic word choice, while others also mentioned adequate word choice and appropriate word choice as discriminating factors. Overwhelmingly, L2 writing teachers reported being most focused on grammar. Jenna said that she "tends to be heavy on grammar." Other raters seemed to concur, being specific about types of grammatical errors they noticed in the text (some of which were not actually present).

Composition raters reported being most focused on rhetoric, explaining that they focused on the argument and evidences to support the argument. When looking at the areas of organization in a text, Jean stated that she based her judgements on the transitions throughout the text. Others focused on the localized ideas within the paragraph. Composition teachers were vague about word choice, mostly looking for fluid, sensible, or "sparkly" (Jean) wording. In the area of grammar, Spencer and Jean, two of the composition teachers, reported that they looked for patterns of error instead of individual errors and generally only addressed specific grammar issues when they impeded the flow of the rhetorical argument.

5. Discussion

5.1 Holistic Scores

Our holistic results, which showed no significant difference in rater leniency (beyond a descriptive leniency among L2 teachers), corroborate Brown (1991), who found no significant difference between English and ESL raters when examining 112 essays, but stand in contrast to a conflicting report by Song and Caruso (1996) who found English faculty to be more lenient than ESL faculty when rating two ESL essays. Regardless of whether the scores actually differed may be irrelevant, however, since, as Brown (1991) admits, different faculty may "arrive at those scores from somewhat different perspectives" (p. 601). The exploratory results elsewhere in this study suggest that L2 writing teachers' rating behavior supports other research showing that teachers react differently to L2 writing based on their teaching backgrounds (Cumming, Kantor, & Powers, 2001, 2002). What triggers such discrimination is the focus of our further data analysis.

5.2 Rubric Categories

The first research question sought to identify how teachers of differing disciplinary backgrounds read and process L2 texts in terms of four component features: rhetoric, organization, word choice, and grammar. Each category showed some significant differences, illustrating that L2 teachers generally have longer dwell times and more run counts than composition teachers.

Rhetoric

The rhetoric results were the most dramatic, showing significant differences across teacher backgrounds on each measurement. The L2 teachers had a longer total dwell time and higher reading counts while compositionists had a higher first run dwell time. This may be because the words used by the student writer were novel or unfamiliar to the composition teachers or that these teachers took slightly longer to cognitively access words within the rhetoric category. This second explanation may be more reasonable given that the rhetorical features in the text illustrated a conflicting argument in which the author presented himself as "not a good writer" but then also stated "I like writing very much." This seeming contradiction in argument, repeated throughout the essay, may have been more difficult for the compositionists to initially process than the L2 teachers who are conditioned disciplinarily to see writing in terms of a formula rather than an argument (Atkinson & Ramanathan, 1995) and thus may favor the presence of topic statements.

The fact that the L2 teachers otherwise had a longer total dwell time but also had higher run and regression-in counts, but that they also had higher initial skip counts, suggest that the L2 teachers used a different approach to reading the rhetorical features of the student text. This approach could be characterized by early skimming or scanning of the argumentative structure of the text followed by more careful re-reading

of the rhetorical features. This approach prioritizes later reading processes which are associated with cognitive processing of meaning, though the same findings could indicate that teachers were distracted or confused when reading. If we take the former interpretation, it is plausible that the L2 writing teachers created a mental outline of the rhetorical structure of the text and then re-read it to flesh out their understanding, an approach similar to that described by Cumming, Kantor, and Powers (2002) based on teachers' self-description of their reading and assessment behaviors.

Organization

The results from the organization data analysis indicate that the teachers approached the organizational features of the text in ways similar to those of rhetoric. That is, the L2 teachers tended to re-read the organizational features more than composition teachers. However, there was no evidence that either group of teachers initially skipped the organizational features or regressed into them any more than the other. In other words, both groups seemed to read the text's organization relatively linearly with the L2 teachers re-reading more often. Although these features were mostly simple, one-word transition phrases ("however," "moreover," "therefore"), their immediate functions within discourse are context-dependent and may require substantial cognitive resources to parse, prompting re-reading. Alternatively, however, is the explanation which accounts for the disciplinary difference: that because L2 teachers see these organizational features as necessary components of a formulaic essay design, they may have re-read them in order to assess their effect on the essay as a whole.

Word Choice

The data reveal a continuation of the trend seen in rhetoric and organization: that L2 teachers re-read the word choice features more than composition teachers. This is evidenced in higher fixation and run counts. Furthermore, the initial skip rate was higher for L2 teachers, suggesting yet again that these teachers skimmed the word choice features only to return to them in subsequent passes. The first run dwell times were not significantly different, indicating that neither teacher group spent more time decoding and accessing each lexical item. The items coded for word choice were not single vocabulary words, but rather were slightly unusual yet grammatically acceptable phrases. We had assumed that L2 teachers with some experience reading unusual English phrases would access them more quickly than composition teachers, but this turned out not to be the case, suggesting that both groups of teachers had equal difficulty (or ease) in initially accessing unusual wording, though L2 teachers did reread them more, suggesting greater processing difficulty or perhaps more judgement time. The overall interpretation of word choice data is that disciplinary background led composition teaches to read these items more linearly than L2 teachers when preparing to offer an assessment of writing.

Grammar

Much as discussed earlier, the grammar results show that the L2 writing teachers seemed to skim through these elements (initially skipping many) and then return to them in succeeding re-readings of the text. Composition teachers seemed to favor a more linear reading approach, though they spent more time accessing grammatical items during first runs, which we expected given that the items were legitimate language or mechanical errors. Since composition teachers are generally less accustomed to L2-type grammar errors, it was expected that they would need more time to access and process them.

Overall, the results from the eye-tracking component suggest that the teachers from different disciplinary backgrounds in this exploratory study tended to read L2 texts differently. In nearly all categories under investigation, the L2 writing teachers spent longer times reading components of the L2 text. L2 teachers also appeared to re-read features in each category more than the composition teachers. These observations seem to indicate different approaches to essay reading, namely that the composition teachers displayed a more linear approach while the L2 writing teachers were generally more recursive in their reading. This may also indicate a different assessment approach. As discussed earlier, Edgington (2005), Wolfe, (2005), and Cumming, Kantor, and Powers (2001) all used teachers' self-reports during and after assessment activities to demonstrate that disciplinary background influences rating behavior. Specifically, composition teachers are more likely to read an essay through in its entirety, form a mental image of it in their minds, and then consult a scoring guide before arriving at a final assessment whereas L2 teachers are more likely to make judgements during and throughout the reading process, adjusting their evaluation as they go. Winke and Lim (2015) have recently corroborated this process through an eye-tracking study of rubric rating among L2 teachers. The results from our eye-tracking research provide evidence of reading processes that support different disciplinary approaches. Composition teachers, after all, had lower dwell times and fewer re-reading counts across nearly all rubric categories while L2 teachers appeared to skim through many of the categories, reread more often, and overall spend more time on most reading measures for each rubric category and on the assessment task as a whole.

5.3 Interviews

Immediately following the eye-tracking activity, we interviewed each participant to ask about their general approach to reading and rating essays and their approach to the rating task they just completed. L2 writing teachers elaborated on their grading style, stating that they familiarized themselves with the text first and later returned for a more thorough look. By contrast, composition raters tended to report making more judgements on the first pass. Spencer, a composition rater, reported that his approach in rating involved finding points of interest in the first run through, and then returning to compare those with other positive or negative points that he noticed.

In comparison to their interviews, the teachers' reading behaviors are somewhat different. Both populations said they would be lenient on L2 grammar, and the L2 writing teachers further said they would spend more time on grammar. As seen in Table 6, we know that this is not entirely the case since composition teachers spent slightly more time looking at grammar features than L2 teachers, and when taken as a proportion of all dwell times, L2 teachers only spent a quarter of the time on grammar features compared to the compositionists who spent a third of their dwell time there. L2 writing teachers also reported focusing on grammar and word choice in fairly equal ways, something that is partially born out in the eye-tracking data in that L2 writing teachers initially skipped more AOIs in these two categories than composition teachers and spent a little more than 20% of their dwell time on each of these categories. However, grammar was not the primary focus of L2 teachers, but instead they focused the greatest proportion of their time on rhetorical features, a category whose definition they struggled to articulate in the interviews.

Meanwhile, the compositionists' claims that they tended to focus more on rhetoric and weightier grammar errors that interfered with meaning appears to be reflected in their reading approach since they spent about one-third of their dwell times on each of these categories. Furthermore, the fixation data supports the composition teachers' claims of the importance of the first read through in their grading. Composition teachers spent less time on the essay and had significantly fewer runs through the AOIs, suggesting that they read through the essay more linearly while L2 writing teachers reread features of the essays multiple times.

These reports suggest that the composition teachers may have been slightly more self-aware raters while the L2 writing teachers were perhaps less aware of their treatment of L2 grammar errors either because they overlooked them after identifying typical errors at the outset or because their professional training allowed them to observe and account for grammar errors more quickly than composition teachers. In any event, the two rater groups appeared to diverge in their assessment approaches in ways predicted by their disciplinary training with compositionists examining texts more holistically and L2 writing teachers focused more on re-reading constituent components of the text.

6. Conclusions and Future Work

Although limited in scope and generalizability, this study supports observations that composition teachers, whose discipline is rooted in rhetoric and whose praxis tends to value critical thinking and originality, contrast with L2 writing teachers whose parent field is linguistics and who are more likely to embrace deductive essay organization and tolerate grammar errors in student writing (Atkinson & Ramanathan, 1995; Costino & Hyon, 2011; Silva & Leki, 2004). In particular, the findings from this exploratory study show that our writing teachers' educational backgrounds did impact their rating behaviors. Although the L2 writing teachers did not display significant differences in

their holistic scores compared to the composition teachers, they did tend to take more time reading the essay in total and generally did more processing of textual features associated with rhetoric, organization, and word choice. This is in contrast to the composition teachers who tended to read the rhetorical and grammatical features of the text more linearly and re-read less of the text overall, especially rhetorical and lexical features. They also initially skipped fewer AOIs within the rhetoric, word choice, and grammar categories. All this resulted in what appeared to be a more holistic approach to reading and rating the L2 text among the composition teachers. These findings support those of previous think-aloud protocol studies which similarly demonstrate more recursive decision making among L2 writing teachers and more linear rating among compositionists (Cumming, Kantor, & Powers, 2001, 2002).

The results of this exploratory study are not generalizable due to a small sample size and their exploratory nature, and this is something which must not be overlooked. Nevertheless, we found differences that cannot be measured through think aloud protocols alone. Because of this we believe that these results may have implications for writing teachers and administrators in both composition and L2 writing programs if they are born out in larger follow-up studies. These studies could include using a larger sample size in both number of participants and number of L2 essays. Based on our research, future studies on the grading of rhetoric would be most productive since this is the rubric category with the most striking differences between teacher groups. Furthermore, because of limitations in the study design, we were unable to place the student essay side-by-side with the rating rubric, and so future research could integrate these to determine whether and how teachers access a rubric during the rating task. Future researchers could also initiate a post-study reflective protocol to further test how teachers conceptualize their rating.

In addition to the findings, which are interesting in their own right, an important contribution of this research is a demonstration of the potential use of eye-tracking methodologies in the research of writing. This was a primary objective in our study design and one that composition researchers have encouraged (Anson, Horn, & Schwegler, 2009; Anson & Schwegler, 2012). Given the advancement of eye-tracking technology and its ability to measure reading behaviors that may relate to cognition, we hope others will continue to investigate writing and composition issues using eye-tracking applications.

Notes

1. Smit (2004) argues that composition studies as an academic field of research began only in the 1960s in the U.S. even though composition instruction itself began much earlier.
2. It should be noted that writing teachers do not universally agree on a single end goal. Smit (2004) explains that composition goals include empowering students toward personal freedom, socializing students into communities of practice, and fashioning participants of civic dialogue. This view is further complicated by national context in that writing teachers outside of a North American setting may differ substantially from these goals.
3. Effect size was measured using Rosenthal's (1994) r where $r = \frac{Z}{\sqrt{N}}$ with standard values of small < 0.3, medium = 0.3-0.5, and large > 0.5.
4. All names are pseudonyms; an internal ethics board approved all aspects of this study including the collection of interview data.

References

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.

Anson, C., Horn, S. R., & Schwegler, R. A. (2009). The promise of eye-tracking methodology for research in writing and reading. *Open Words Access and English Studies, 3*(1), 5-28.

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. http://doi.org/10.2307/3588074

Atkinson, D., Crusan, D., Matsuda, P. K., Ruecker, T., Simpson, S., & Tardy, C. (2015). Clarifying the relationship between L2 writing and translingual writing: An open letter to writing studies editors and organization leaders. *College English*, *77*(4), 383–387.

Bawarshi, A. S., & Reiff, M. J. (2010). *Genre: An introduction to history, theory, research, and pedagogy*. West Lafayette, IN: Parlor Press.

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly*, *25*(4), 587–603. http://doi.org/10.2307/3587078

Brereton, J. (Ed.). (1996). *The origins of composition studies in the American college, 1875–1925*. Pittsburgh, PA: University of Pittsburgh Press. https://doi.org/10.2307/j.ctt5hjsd2

Canagarajah, S. (2016). Translingual writing and teacher development in composition. *College English*, *78*(3), 265–274.

Conklin, K., & Pellicer-Sanchez, A. (2016). Using eye-tracking in applied linguistics and second language research. *Second Language Research, 32*(3), 453-467. https://doi.org/10.1177/0267658316637401

Costino, K. A., & Hyon, S. (2011). Sidestepping our "scare words": Genre as a possible bridge between L1 and L2 compositionists. *Journal of Second Language Writing*, *20*(1), 24–44. http://doi.org/10.1016/j.jslw.2010.12.001

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.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86,* 67–96. https://doi.org/10.1111/1540-4781.00137

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.

Eckstein, G., Chariton, J., & McCollum, R.M. (2011). Multi-draft composing: An iterative model for academic argument writing. *Journal of English for academic purposes, 10* (3), 162-172. https://doi.org/10.1016/j.jeap.2011.05.004

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

Ferris, D. R. (2003). *Response to student writing: Implications for second language students*. Mahwah, NJ: Lawrence Earlbaum Associates. https://doi.org/10.4324/9781410607201

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., Brow"n, J., Liu, H. S., & Stine, M. E. A. (2011). Responding to L2 students in college writing classes: Teacher perspectives. *TESOL Quarterly*, *45*(2), 207–234. http://doi.org/10.5054/tq.2011.247706

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.

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. ttps://doi.org/10.1111/lang.12136

Inhoff, A. (1984). Two stages of word processing during eye fixations in the reading of prose. *Journal of Verbal Learning and Verbal Behavior, 23*, 612-624. doi: 10.1016/S0022-5371(84)90382-7

Lerner, N. (2005). The teacher-student writing conference and the desire for intimacy. *College English*, *68*(2), 186. http://doi.org/10.2307/30044673

Leki, I. (2000). Writing, literacy, and applied linguistics. *Annual Review of Applied Linguistics*, *20*, 99–115. https://doi.org/10.1017/s0267190500200068

Matsuda, P. K. (1999). Composition studies and ESL writing : A disciplinary division of labor. *College Composition and Communication*, *50*(4), 699–721. https://doi.org/10.2307/358488

Matsuda, P. K. (2006a). Second-language writing in the twentieth century: A situated historical perspective. In P. K. Matsuda, M. Cox, J. Jordan, & C. Ortmeier-Hooper (Eds.), *Second-language writing in the composition classroom* (pp. 14–30). Boston, MA: Bedford/St. Martin's.

Matsuda, P. K. (2006b). The myth of linguistic homogeneity in U.S. college composition. *College English*, *68*(6), 637–651. https://doi.org/10.2307/25472180

Matsuda, P. K. (2014). The lure of translingual writing. *PMLA*, *129*(3), 478–483. https://doi.org/10.1632/pmla.2014.129.3.478

Moussu, L. (2013). Let's talk! ESL students' needs and writing centre philosophy. *TESOL Canada Journal*, *30*(2), 55–68. https://doi.org/10.18806/tesl.v30i2.1142

Polio, C., & Friedman, D. A. (2016). *Understanding, evaluating, and conducting second language writing research*. NY: Routledge. https://doi.org/10.4324/9781315747293

Ramanathan, V., & Kaplan, R. B. (2000). Genres, authors, discourse communities: Theory and application for (L1 and) L2 writing instructors. *Journal of Second Language Writing*, *9*(2), 171–191. http://doi.org/10.1016/S1060-3743(00)00021-7

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. https://doi.org/10.1037//0033-2909.124.3.372

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

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. https://doi.org/10.1016/j.cogsys.2005.07.002

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis*. (pp. 231-244). New York: Russell Sage Foundation.

Santos, T. (1992). Ideology in composition: L1 and ESL. *Journal of Second Language Writing. 1*.1–15. https://doi.org/10.1016/1060-3743(92)90017-j

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly*, *22*(1), 69–90. https://doi.org/10.2307/3587062

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. ttps://doi.org/10.1111/j.0026-7902.2004.00215.x

Smit, D. W. (2004). *The end of composition studies*. Carbondale, IL: Southern Illinois University Press.

Staub, A. & Rayner, K. (2007). Eye movements and online comprehension processes. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 327–342). Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198568971.001.0001

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.

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: L2 Student Essay and AOI Codes

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