8-14-2019 2:00 PM

Does Lexical Frequency affect rater judgement of essays? An experimental design using quantitative and qualitative data

Mohammad Muneer Ul-Huda
The University of Western Ontario

Supervisor
Faez, Farahnaz
The University of Western Ontario
Boers, Frank
The University of Western Ontario

Graduate Program in Education
A thesis submitted in partial fulfillment of the requirements for the degree in Master of Education

ABSTRACT

Many correlational studies show a positive relation between written assessments of language and use of more diverse vocabulary (Lexical Diversity) and more infrequent words (Lexical Frequency). However, there have been no experimental studies that have isolated the effects of Lexical Frequency from Lexical Diversity. In the present study, 14 raters judged two versions of the same essay that differed only in Lexical Frequency. A Paired T-test showed no difference in mean scores between essays ($t(13) = .396$, $p = .70$) when the Lexical Frequency of 23.5% of Content Words were changed in a 347 word essay. Comments explaining scores given to essays showed that features other than vocabulary had a far greater influence on rater judgement. It is possible that the Lexical Frequency manipulations were not great enough to affect rater judgement, whether subliminal or conscious. Implications of these results for standardized language proficiency tests and future research in vocabulary are discussed.

KEYWORDS

Lexical Frequency, Lexical Diversity, Rater Judgement, Lexical Features, Lexical Indices

Standardized Language Tests, TOEFL

SUMMARY FOR LAY AUDIENCE

Studies have shown a positive relation between essay scores and the use of a greater variety of words (referred to as Lexical Diversity) and use of words that are less common (referred to as Lexical Frequency). These two variables, Lexical Diversity and Lexical Frequency, are often measured together. Because of this, we don't know how Lexical Frequency alone (use of less common words) affects a rater's perception of a writing. We conducted an experimental design where we produced two essays that were identical, except with regard to their Lexical Frequency: one essay used more common words (High Frequency Essay) and the other used less common words (Low Frequency Essay). Statistical analysis showed that raters (n=14) gave both essays the same score on average. After looking at rater comments explaining why they scored essays the way they did, we see that features of the essay other than vocabulary affected rater judgement far more. This could be one reason why we saw no difference in essay scores. Another reason could be that the change in Lexical Frequency that we created was simply not big enough to affect rater judgement. Implications of these results are discussed.

ACKNOWLEDGMENTS

I would like to thank Dr. Farahnaz Faez for accepting me as her student and allowing me to pursue a project of my own desire. The circumstances under which we came together were unexpected; in hindsight, I would have it no other way. Her guidance, support, and confidence in my abilities have made this an instructive and rewarding journey.

I would also like to thank Dr. Frank Boers, my instructor and co-supervisor. Dr. Boers' passion for teaching, grammar, and quality research has been an inspiration. His critical eye for academic hogwash has kept me on my toes, for which I am grateful.

I am also thankful to my grad mate and fellow teacher, Tomlin Gagen. He has been a soundboard when I needed to bounce ideas, a voice of cool reason when I was fed up with my own, and a chimney when I needed to vent. Thank you, my friend; it has been a pleasure being on this journey with you.

Finally, I would like to thank my mother, Nahid. She supported me on this journey in her own way, as only a mother could.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Thesis organization

This thesis is written in an integrated article format and organized into three parts. Chapter one provides a quick introduction to the literature, highlights unanswered questions, and explains what I hope to achieve with the present study. Chapter 2 is the integrated article, which provides a more detailed literature review of the written assessment and vocabulary field, explains the methodology used in the present study, gives an analysis of the results and what conclusions can be drawn. Chapter 3 concludes and summarizes the most relevant details of the study.

1.2 Thesis introduction

Vocabulary is judged as an independent construct in all second language assessments, written or spoken. It is an essential part of standardized language proficiency tests (SLPTs), like IELTS and TOEFL. Whether the test is being scored holistically or analytically, vocabulary in some form or manner is mentioned in the scoring criteria.

There is a large body of research in linguistics that has been dedicated to determining what vocabulary features are most salient to raters. Early research in vocabulary focussed on Lexical Diversity and Lexical Frequency (Arnaud, 1984; Linnarud, 1986; Grobe, 1981; Laufer, 1994; Laufer & Nation, 1995). More recent research has shifted focus to other features, such as N-Grams (multi-word units), collocations (words that often occur together), hypernymy (specific vs less specific words), polysemy (words with multiple senses/meanings), and Range (words used in fewer or greater contexts) and other psycholinguistic word features, such as word

imageability (how easily a word evokes the senses) and word concreteness (tangible vs abstract words) (Kyle & Crossley, 2015; Kyle & Crossley, 2016; Kyle et al., 2017; Kim et al., 2018).

Many of these vocabulary features interact with each other. For example, if you increase the number of words in an essay indefinitely, after a point you can only increase Lexical Diversity by including less frequent words (Lexical Frequency; Malvern et al., 2004). Similarly, words that are found in fewer contexts (Range) are usually less frequent in the language than words that are found in greater contexts (Kim et al., 2018). Because of this interaction effect it can be difficult to isolate the influence of individual vocabulary features on rater judgements using correlational studies.

This thesis proposes an experimental design to determine how measured and objective changes in specific vocabulary features affect rater judgement of essays. Specifically, it looks at how changes in Lexical Frequency affect rater judgement of essays, as determined by raters' scores of and comments on essays.

1.3 Ethics approval

This study required human participation. Participants were recruited after approval from the Western University Non-Medical Research Ethics Board (see Appendix F for approval letter). All procedures, including recruitment and storage of participant data, were compliant with guidelines set and approved by the ethics board.

CHAPTER 2: ARTICLE

2.1 Introduction

Vocabulary is judged as an independent construct in all second language written assessments, written or spoken, whether scored holistically or analytically. It is an essential part of standardized language proficiency tests (SLPTs), like IELTS and TOEFL.

There is a large body of research in linguistics that has been dedicated to determining what vocabulary features are most salient to raters. Early research in vocabulary focussed on Lexical Diversity and Lexical Frequency (Arnaud, 1984; Linnarud, 1986; Grobe, 1981; Laufer, 1994; Laufer & Nation, 1995). More recent research has shifted focus to other features, such as N-Grams (multi-word units), collocations (words that often occur together), hypernymy (specific vs less specific words), polysemy (words with multiple senses/meanings), and Range (words used in fewer or greater contexts) and other psycholinguistic word features, such as word imageability (how easily a word evokes the senses) and word concreteness (tangible vs abstract words) (Kyle & Crossley, 2015; Kyle & Crossley, 2016; Kyle et al., 2017; Kim et al., 2018).

Many of these vocabulary features interact with each other. For example, if you increase the words in an essay indefinitely, after a point you can only increase Lexical Diversity by accessing less frequent words (Lexical Frequency; Malvern et al., 2004). Similarly, words that appear in fewer contexts (Range) are usually less frequent than words that appear in greater contexts (Kim et al., 2018). Because of this interaction effect it can be difficult to isolate the influence of individual vocabulary features on rater judgements using correlational studies.

To the best of my knowledge, there have only been two experimental studies that have looked at the influence of Lexical Frequency and Lexical Diversity on rater judgement using experimental designs (Fritz & Ruegg, 2013; Vögelin et al., 2019). This study builds on and addresses limitations from these previous studies to determine how measured and objective changes in Lexical Frequency affect rater judgement of essays, as determined by rater scores and comments on essays.

The results of this study will help inform the effectiveness of holistic rubrics used in SLPTs, like the TOEFL iBT. It will also help validate the large body of correlational studies already done in the field. Many of the correlational studies done in the field of vocabulary and written assessment indicate that Lexical Frequency is a fairly salient feature when it comes to affecting rater judgement of essays. I hope to isolate the effect of Lexical Frequency on rater judgement to confirm a cause-effect relationship. Lastly, I hope to propose guidelines for methodology and how results should be reported in future studies in lexical feature manipulation in vocabulary and written assessment research.

2.2 Literature Review

2.2.1 Lexical Features and Terminology

Lexical Frequency (LF) is how frequently a word appears in a language according to some reference corpus. A corpus is a collection of texts, written and/or spoken, that is supposed to be representative of a discourse. For example, the Corpus of Contemporary American English (COCA) is supposed to be representative of contemporary American English, containing 560 million words from fiction, non-fiction, written and spoken text (Davies, 2008). In its simplest

form, the LF of a text is calculated by assigning a frequency value to each word in the text, and then summing those values and dividing it by the total number of words in the text.

Lexical Diversity (LD) can be measured by a simple Type-Token Ratio (Johnson, 1939, 1944), where Tokens refers to the total number of running words (including repetitions of the same words) and Types refers to the total number of unique words. A Type-Token Ratio is subject to text length, where the longer the piece of writing, the more likely it is that the Type-Token ratio will fall, as common words (such as articles and prepositions) will be repeated more often (Jarvis, 2013). To date, there have been many proposed measurements of LD that have tried to control for the text-length dependent problem. For this research, we use the Measure of Textual Lexical Diversity (MTLD; McCarthy, 2005), which has shown resilience against varying text-length (McCarthy & Jarvis, 2010; Jarvis, 2013; Koizumi, 2012) and been used in more recent studies looking at Lexical Diversity (Gonzalez, 2017a, 2017b; McNamara et al., 2010; Vögelin et al., 2019). MTLD calculates LD by processing the Type-Token Ratio of a text in chunks (called TTR factors) in a forward sequence (left to right) and a backward sequence (right to left), giving an average value of the two (see McCarthy & Jarvis, 2010, for a more detailed explanation).

There is a large and growing body of research on vocabulary that looks at LF and LD. However, not all researchers use these same terms. Where possible in this paper we will identify the different terms used by other researchers. Many previous researchers have used the term Lexical Sophistication for Lexical Frequency (Laufer, 1994; Laufer & Nation, 1995; Fritz and Ruegg, 2013; Saito et al., 2016; Vögelin et al., 2019). However, Kristopher Kyle and colleagues (Kyle & Crossley, 2015; Kyle & Crossley, 2016; Kyle et al., 2017; Kim et al., 2018) have proposed defining Lexical Sophistication as a multi-dimensional construct that ought to be measured

using multiple lexical features (indices), of which Lexical Frequency is only one. For this reason, we do not use the term “Lexical Sophistication.””

2.2.2 The relationship between LD, LF, and language proficiency

The literature is in agreement that as language proficiency increases so does use of diverse and infrequent words by first language (L1) and second language (L2) users, in writing and speaking (Arnaud, 1984; Linnarud, 1986; Grobe, 1981; Laufer, 1994; Laufer & Nation, 1995; Tidball and Treffers-Daller, 2008; Crossley & McNamara, 2009; Crossley et al., 2011a; Crossley et al., 2013; Gonzalez, 2017b).

For example, Arnaud (1984) found that French learners of English produced less diverse and more frequent vocabulary than American English native-speakers. The author used the term “Lexical Richness” as a multi-dimensional construct that captured Lexical Diversity *and* Lexical Frequency (referred to as “rare words” in paper). The study found a small but significant correlation between an independent productive vocabulary test and measures of Lexical Richness, implying that as student proficiency increased (as determined by the vocabulary test) so did their use of more diverse and infrequent words.

A similar study by Linnarud (1986) compared Swedish learners of English to native speakers of English. The author found that native speakers of English produced more Lexical Diversity and less frequent words in their writing compared to their Swedish counterparts.

Grobe (1981) looked at which features, including spelling, text length, syntactic maturity, and LD, best predicted scores given to essays written by 5th, 8th, and 11th grade students (assuming most were L1 users of English). The author found that Lexical Diversity (referred to as

"TYPES" in the paper) was one of the best predictors of holistic scores received on essays, with higher scores being given to essays with more diverse vocabulary. The 11th graders produced more Lexical Diversity than the 5th and 8th graders, implying that as language proficiency increases (assumed by grade level of student), so does knowledge and use of more diverse vocabulary.

Laufer (1994) compared essays by L2 students over a year at three intervals. With time, students started using more infrequent vocabulary (referred to as "Lexical Richness" in the article), though there was no change in Lexical Diversity (referred to as "Lexical Variation" in the article). It is unclear from the paper, but Laufer (1994) might have used a simple Type-Token Ratio (TTR) to measure Lexical Diversity. The topic and genre of writing was not controlled for either. Research has shown that text length (see Jarvis, 2013; Crossley et al., 2011a; Xie, 2015; Grobe, 1981; Ferris, 1994; Frase et al., 1998), and topic (O'Loughlin, 1995; Yu, 2010) can affect Lexical Diversity, which may explain why no change in Lexical Diversity was observed.

In a follow up study, Laufer and Nation (1995) compared the Lexical Frequency (referred to as "sophistication") of students' essays to their vocabulary knowledge, as determined by an independent vocabulary size test. Unlike the previous study, this study controlled for length and genre of essays. Students who performed better on the vocabulary size test used more infrequent vocabulary in their writing, suggesting that more proficient language users can use more advanced and rare vocabulary.

After the Laufer (1994) and Laufer and Nation (1995) studies, more researchers started using Lexical Frequency Profiles to measure student use of "sophisticated" vocabulary. Tidball and

Treffers-Daller (2008) wanted to know whether rater judgement *or* lexical frequency lists based on corpora were better able to distinguish between different proficiency levels of L2 learners. They hypothesized that raters, experienced French tutors, would be better at distinguishing between three different groups of French speakers: first year university L2 French students, final year university L2 French students, and L1 French speakers. The results showed that, with increased proficiency, L2 French speakers used less frequent words, with L1 French speakers using the highest percentage of advanced and rare vocabulary. The results also showed that human judgement of basic and advanced vocabulary, rather than frequency-based vocabulary lists, were better able to distinguish between groups of different proficiency levels. These result supports Jarvis' (2013) argument of having human raters validate computational-based results of what is deemed sophisticated vocabulary (of which Lexical Frequency is only *one* measure). However, other studies show that, when word frequencies are relatively close together, even highly educated individuals in the field of linguistics (or related) have a hard time distinguishing between which words are of higher or lower frequency (Schmitt & Dunham, 1999; Alderson, 2007; McCrostie, 2007).

Advancements in Computational Linguistics and the availability of large corpora of essays has allowed researchers to use large datasets and regression analysis models to determine the relationship between LD, LF, and language proficiency. Results continue to show that as language proficiency increases, writers use less frequent words in their essays, whether this is comparing SAT (Scholastic Aptitude Test) essay scores between writers at the high school and college level (Crossley et al, 2011a), or between L1 and L2 writers of varying proficiency (Crossley & McNamara, 2009; Crossley et al., 2013; Gonzalez, 2017b). In addition, Gonzalez

(2017b) also found that L1 writers had greater LD in their writing, with words in the mid-frequency bands (words found in the 3000 to 9000 most frequent word families) making the greatest contribution to LD.

However, Gonzalez (2017b) presents reservations on simply using computational-based frequency indices to classify 'sophisticated' vocabulary. In one example, the author shows how the words *diction* and *English* in two separate excerpts are classified as low frequency (words that occur less frequently than the 9000 most frequent word families). Qualitative human judgement might perceive *diction* as more sophisticated than *English* (see Jarvis, 2013, and Tidball and Treffers-Daller, 2008). Using other criteria, such as Word Familiarity (Coltheart, 1981) or word specificity (Hypernymy; Fellbaum, 2010) might yield a different conclusion of which word is more sophisticated. This highlights the drawbacks of solely using LF as a measure of sophisticated vocabulary, and of using large corpora for data analyses, where it's not possible for researchers to qualitatively comb through the data due to the large volume. Based on these limitations, Gonzalez (2017b) recommends future research on LD and LF gather qualitative data from raters and use experimental manipulation of the two metrics to parse out their effects.

To the best of my knowledge, there are only two recent studies that show no increase in use of lower frequency words by L2 users over time (Crossley et al., 2010; Kim et al., 2018). Both these studies use spoken rather than written data and both suffer from the same limitations.

Crossley et al. (2010) conducted a year long study with 6 L2 speakers to see how their use of polysemic words and LF change with time. The participants were beginner learners in an Intensive English Program (IEP) at an American university. The L2 speakers were interviewed

through the duration of the year on conversational topics that allowed for spontaneous and natural speech. The interviews were transcribed before being analyzed. The results showed that, while the L2 speakers' mastery of polysemic words increased, such that they started using the same words in a greater variety of meanings, their Lexical Frequency did not increase (i.e., they continued using high frequency words, albeit in more diverse ways). An assumption in the study design is that a year would be enough time to see an increase in language proficiency, such that it could potentially affect Polysemy and LF measures in speech. However, there are several limitations with the data used. Firstly, 6 students make for a small sample size, even though the researchers had a total of 99 transcripts at the end for their statistical analysis. Secondly, it is difficult to compare the data collected (conversational interviews with spontaneous and natural speech) with data in other studies mentioned. For example, many of the other studies use texts that were produced under classroom or test conditions (Laufer, 1994; Laufer & Nation, 1995; Crossley et al., 2011a; Gonzalez, 2017b), where participants would be expected to perform with a mindset to 'do well' or impress. The limitations of the second study (Kim et al., 2018) that showed no change in LF in L2 users over time are the same as those in the first, because I believe the exact same dataset was used, going by the description (6 students, 99 transcripts, conversational interviews with spontaneous and natural speech). There is no reason to expect different results with the same dataset.

Summary of the relationship between LD, LF, and language proficiency

In summary, the literature consistently shows that an increase in language proficiency, in L1 *and* L2 users, is characterized by greater Lexical Diversity and a greater use of infrequent words (Lexical Frequency).

2.2.3 The relationship between LD, LF, and written assessment scores

This section of the literature review explores what we know of how Lexical Diversity and Lexical Frequency relate to written assessment scores.

2.2.3.1 Studies that show a positive relationship between LD, LF, and written assessment scores

The greater portion of written assessment research shows a positive relation between LD, LF and written assessment scores (Grobe, 1981; Engber, 1995; Yu, 2009; McNamara et al., 2010; Crossley et al., 2011a; 2011b; Crossley & McNamara, 2012; Gonzalez, 2017a, 2017b; Kim et al., 2018; Vögelin et al., 2019).

For example, when comparing the essays of 5th, 8th, and 11th graders, Grobe (1981) showed that, after controlling for length of essay and spelling errors, LD explained the greatest variation in scores, with higher scores being awarded to essays with greater LD.

In another study, Engber (1995) compared the essays of L2 English students of varying proficiency. Errors (in form and meaning; Nation, 2001) and LD (referred to as "Lexical Variation" in paper) correlated strongest with scores, with more errors resulting in lower scores and more LD resulting in higher scores. However, the strongest correlation was found in LD *without* any error, suggesting that using a greater variety of words is not as important as using a greater variety of words *correctly*.

In a study that looked at the written and spoken assessments from a standardized language proficiency test (the MELAB), Yu (2009) found that LD (measured using 'D'; Malvern & Richards, 2002) made a small but significant contribution to the written (11%) and spoken (23%) assessments, such that higher scores were awarded to texts with more LD.

McNamara et al. (2010) looked at the correlation between essay scores and several lexical features, including Lexical Diversity and Lexical Frequency. The study did not control for genre or essay length, both factors that have previously shown to influence written assessment scores (see Jarvis, 2013; Crossley et al., 2011a; Xie, 2015; Grobe, 1981; Ferris, 1994; Frase et al., 1998, O'Loughlin, 1995; Yu, 2010). However, an Analysis of Variance (ANOVA) showed no difference between high and low scoring essays based on genre or length (i.e., no significant influence of length or genre on essay scores). The three indices that best predicted essay scores were Syntactic Complexity (measured by number of words before the main verb in a sentence), Lexical Diversity (MTLD; McCarthy, 2005), and Lexical Frequency (referred to as "Word Frequency"), with higher scoring essays exhibiting more Lexical Diversity and less frequent words. Of relevance to the present study is that McNamara and colleagues tested for collinearity between Lexical Diversity and Lexical Frequency, which was significant (r = -.51, p < .001). Although this correlation was not strong enough for the cut off criterion of r > = 0.70, it does illustrate that the effects of LD and LF are hard to separate from one another.

Crossley and colleagues have conducted several studies that looked at large corpora of essays, by L1 and L2 writers, and have repeatedly shown that both high LD and use of low frequency words correspond with higher essay scores (Crossley et al., 2011a; Crossley & MacNamara, 2012).

In a variation of the previous studies, Crossley et al. (2011b) looked at another corpus comprised of essays by L2 writers of various proficiencies and L1 writers, but these essays were scored based on a rubric created to assess Lexical Proficiency (*not* overall writing proficiency). Lexical proficiency should be treated as only one aspect of writing proficiency (Diederich at al.,

1961). The final regression model showed that the three variables that best explained lexical scores were Lexical Diversity (measured using 'D'; Malvern, & Richards, 2002), Word Hypernymy (use of general vs specific words), and Lexical Frequency. Together, these variables explained about 44% of the variance in the scores.

Gonzalez (2017a) compared the academic writing of L2 and L1 writers at university. Genre and length of essay was not controlled, but statistical analysis found no significant difference in scores based on these two features. Results showed that raters awarded higher scores to essays that had more Lexical Diversity (measured using MTLD; McCarthy, 2005) and ones that used less frequent words. Of relevance to the present study is that Gonzalez (2017a) found a moderately strong correlation between LD and LF ($r = -0.44$, $p < .001$). Further analysis showed that, when going from low scoring essays to average scoring essays, the biggest difference in the writing profiles was an increase in Lexical Frequency (use of more low frequency words). However, at the top end of the spectrum Lexical Diversity made a greater contribution to essay scores than Lexical Frequency. This suggests that Lexical Frequency and written assessment scores may not have a linear relationship, such that, after a certain point, using more infrequent words could have a negligible or perhaps even a negative effect on scores.

Finally, Vögelin et al (2019) manipulated the LD and LF of four essays (two High Quality, two Low Quality) on the same topic. Each essay had two versions: High Lexical Richness and Low Lexical Richness. Raters were given four essays each, but never two versions of the same essay. The results showed that essays with High Lexical Richness received higher scores than essays with Low Lexical Richness. However, this was also true for High Quality essays over Low Quality essays, whether or not the High Quality essay had High or Low Lexical Richness. Because of this

it is difficult to determine what essay features (lexical or otherwise) were responsible for the high/low scores.

2.2.3.2 Studies that show no relation or a negative relation between LD, LF, and written assessment scores

Far fewer studies in written assessment research show no or a negative correlation between LD, LF, and written assessment scores.

Unlike the previous studies mentioned, Ruegg et al. (2011) used an analytic rubric for their study to determine whether the grammar and lexical components in the rubric measure similar things. The authors used 140 essays written by L2 students of English. For vocabulary features, they looked at errors in word class, word choice, spelling errors, Lexical Diversity (simple Type-Token ratio) and Lexical Frequency. The grammar component of the rubric and Lexical Errors were significantly correlated with lexical scores, such that essays with more errors had lower lexical scores. Lexical scores were positively correlated with LD and negatively with LF (use of infrequent words resulted in lower scores), but neither of these were significant. The negative correlation between LF and lexical scores can be explained by the Lexical Errors. As Engber (1995) pointed out, it's more important to raters that words be used correctly (form, meaning, and use) rather than simply being used. Those writers who attempted to use more advanced vocabulary might have used them incorrectly, which would explain the negative correlation between LF and lexical scores, and the significant correlation between Lexical Errors and lexical scores. The authors recommended experimental designs in the field of written assessment research, particularly on the manipulation of vocabulary in writing.

Following through with their recommendation, Fritz and Ruegg (2013) designed an experiment to see how three lexical metrics, Lexical Diversity (referred to as "Range" in the paper), Lexical Accuracy (errors in form, meaning, and use), and Lexical Frequency (referred to as "Sophistication" in paper) correlated with the lexis component of an analytical rubric. The authors manipulated the content words of a single essay to represent three different levels (high, medium, low) for the three different lexical metrics, resulting in a total of 27 different combinations of essays. 39 essays were assigned to 27 raters, of which each rater received 3 of the experimental essays. The raters were asked to score essays based on an analytic rubric, of which "lexis" was one scale. This acted as the dependent variable in the statistical analysis. Significant results were only found for Lexical Accuracy, with higher scores being given to papers with fewer lexical errors. There were no significant results for changes in Lexical Diversity and Frequency, but there was a trend of increased scores going from Low to Mid LD and LF, but a drop in scores from Mid to High for both features.

To the best of my knowledge, this is the first experimental design of its kind in written assessment research that attempted to manipulate Lexical Errors, LF, and LD. Almost inevitably, given the complexity of the research problem, there are issues of validity and design in the study that may render some of the conclusions invalid.

Appendix D of the research paper shows essays with the highest Lexical qualities (High in LD, LF, and Low in Lexical errors). Below is an example sentence from this essay:

"I have comprehended that meats need our mortal torso because meats fortify our hemoglobin and support our anatomy, so only vegetable is not advantageous for our health."

It should be evident to any proficient speaker of English that the choice of wording is awkward, nonsensical at times, and it is difficult to imagine why an L2 learner would make these word choices if indeed they knew their true meaning.

Such extreme manipulations of the essays could have rendered the experiment ecologically invalid, and the raters might have picked up on the manipulations, as reflected by the scores given. Indeed, the authors mention that, "a total of three similar essays was possibly too many and may have aroused some suspicion" (Fritz & Ruegg, 2013, pg. 179).

Xie (2015) looked at two different strategies used by L2 writers: a Risk-Taking Approach or a Defensive Approach. The Risk-Taking Approach included students using more sophisticated sounding vocabulary (low frequency words) and complex sentences. Of the different lexical features measured, only essay length and errors significantly correlated with essay scores, with longer essays and essays with fewer errors awarded higher scores. This is consistent with previous research (Engber, 1995; Crossley et al. 2011a; Crossley & McNamara, 2012; Santos, 1988). Lexical sophistication (measured by Lexical Frequency) correlated with lower scores rather than higher scores. A closer look at the data might explain the results: the students who employed Risk-Taking strategies also committed the greatest number of errors, which included lexical and grammatical errors. Therefore, what the results actually show is that students who use infrequent words *incorrectly* get a lower score. This conclusion is consistent with previous research (Engber, 1995) which has shown that error-free lexical diversity explained rater scores to a greater extent than lexical diversity alone.

2.2.3.3 Other lexical features and their relation to written assessment scores

Thus far we've only touched on the lexical features LD and LF and their relation to written assessment scores. However, more recent research using large corpus-driven data and computational linguistics have put other lexical feature front and center in written assessment research (Kyle & Crossley, 2015; Kyle & Crossley, 2016; Kyle et al., 2017; Kim et al., 2018).

These studies primarily involve use of the lexical analysis software, TAALES: the Tool for the Automatic Analysis of Lexical Sophistication (NOTE: "Lexical Sophistication" here is used as an umbrella term to cover various lexical features, *not* just Lexical frequency). Unlike previous research that has assumed a relationship between LF, LD and written assessment scores and student language proficiencies, Kyle and colleagues take all possible lexical features (referred to as "indices") in the vocabulary research field and examined which indices best predict Lexical Scores (Kyle & Crossley, 2015; Kyle et al., 2017; Kim et al., 2018), Written Assessment Scores (Kyle & Crossley, 2016; Kim et al., 2018), and Spoken Assessment Scores (Kyle & Crossley, 2015; Kim et al., 2018) using multiple regression analysis.

The original version of TAALES (Kyle & Crossley, 2015) consisted of 130 lexical indices, while TAALES 2.0 (Kyle & Crossley, 2017) added an additional 300 indices. Results from these studies have shown that the lexical features that best explain lexical and/or written assessment scores are N-Gram frequencies, Word Range, Word Familiarity, Hypernymy, and Polysemy (Kyle & Crossley, 2015; Kyle & Crossley, 2016; Kyle et al., 2017; Kim et al., 2018. NOTE: this is a partial list summarizing results from all four studies. See individual studies for full list). While Lexical Frequency indices also explain variation in scores, their contribution was minimal compared to these other indices.

Kim et al. (2018) suggest that, rather than trying to measure the concept of Lexical Sophistication using single lexical features, we should group similar and related lexical features into “dimensions” to help explain aspects of variation in writing scores. Of relevance to the present research is the suggestion that the construct of Lexical Sophistication might be too complex to measure using just a single measurement, namely Lexical Frequency, and that Lexical Frequency might strongly be related to other, multiple features, such as Hypernymy (because low frequency words tend to be comparatively specific) or Word Acquisition Properties (because low frequency words tend to be acquired later in life). Future research may need to start using a multidimensional construct to measure Lexical Sophistication and correctly describe what linguistic construct Lexical Frequency is actually measuring.

Note that Lexical Diversity was NOT measured in any of these studies and Kim et al. (2018) do not regard LD as a measure of Lexical Sophistication, but rather a separate lexical construct (but see Jarvis, 2013, for counter argument).

Summary of the relationship between LD, LF, and written assessment scores

In summary, the literature shows that using more diverse vocabulary and less frequent words correlates to higher scores on written assessments. The few studies that show no or a negative correlation between LD, LF and written assessment scores can be explained by research design limitations and sampling issues (example: low-proficiency L2 learners). More importantly, recent studies in computational linguistics show that while LF may correlate to higher written assessment scores, there are other lexical features with a stronger correlation, such as N-gram frequencies, word hypernymy, Range, and other psycholinguistic word features such as word

imageability (how easily a word evokes the senses) and word concreteness (tangible vs abstract words).

2.2.4 Experimental designs in vocabulary and written assessment research

This section talks about the benefits of using experimental designs in vocabulary and written assessment research and critically examines previous experimental studies in the field.

2.2.4.1 What experimental designs can tell us

All studies mentioned thus far, with the exception of two (Ruegg & Fritz, 2013; Vögelin et al., 2019), have been correlational studies. Correlational studies collect measurements on two or more variables to find a relationship between them. There are many advantages to correlational studies, particularly in education, where data is often produced under 'natural' conditions and the results can be said to be ecologically valid. Developments in technology and computational linguistics and the availability of large digital corpora have allowed for correlational studies using large sample sizes. However, it is difficult to draw definitive conclusions on cause and effect relations between variables using only correlational studies. There are many variables that are not controlled for and some whose effects we cannot measure. This is particularly true when looking at vocabulary in written assessment research that uses holistic rubrics.

As Read (2000) has noted:

> One problem with holistic rubrics is that many test takers may have varying strengths and weaknesses in different aspects of writing and speaking. For example, a test taker can have very good ideas that are structured in a coherent and unified way, but may not

> have the vocabulary to express the ideas. Or the learner may have great grammar and vocabulary, but may not be familiar with common phrasal expressions in the Target Language. A holistic score is then "a compromise between competing considerations." For vocabulary and assessment research, this means that we cannot confidently say to what degree the vocabulary alone in a piece of writing affected the score given (Ch 7, p. 214).

Experimental designs have the advantage of isolating the effects of specific lexical features (such as Lexical Frequency) and seeing how these features impact rater judgement.

2.2.4.2 A critique of past experimental studies

In the experimental design by Fritz and Ruegg (2013), 3 lexical variables were manipulated: errors (form, meaning, and use; Nation, 2001), Lexical Diversity (referred to as “Range” in paper), and Lexical Frequency (referred to as “Sophistication” in paper). The results showed that essay scores significantly increased with fewer errors, however there was no significant relationship for LD and LF. The following points highlight the limitations in the experimental design:

1. Below is an excerpt from the essay that represented the “Highest Lexical Quality,” meaning, Low Lexical Errors, High LF (more infrequent words) and High LD:

“I have comprehended that meats need our mortal torso because meats fortify our hemoglobin and support our anatomy, so only vegetable is not advantageous for our health. These reasons support my conviction which I deem we should eat also meats. I envisage that animals to eat by hominid have not been reared compassionately.”

The underlined words represent content words that were manipulated by the researchers. To an experienced ESL teacher, the above excerpt should appear strange, perhaps even nonsensical. It is difficult to believe that a writer who would know some of the low-frequency and specialized words used in the "Highest Lexical Quality" essay – such as hemoglobin, envisage, deem – would produce such unconventional phrasing and grammatical errors. This would violate the ecological validity of the study, as it is highly unlikely that an L2 student would naturally produce such writing. At the very least, this would alert the raters to the manipulation effect, which would undoubtedly affect their judgement of the paper. In fact, Fritz and Ruegg (2013) state that some raters may have become aware of the manipulation. For the experiment, the raters were told they were marking actual student essays for grades (as opposed to being told they were part of a research project). Fritz and Ruegg (2013) note that some of the manipulated papers received a mark of "0," suggesting that some raters suspected foul play by students.

2. Some of the manipulated words might not be accurate synonyms of the words they are supposed to replace. For example, "hemoglobin," which is supposed to replace "blood" (see Fritz & Ruegg, 2013, Appendix B), is a specialized/technical term often associated with use in medicine and biology and used in far fewer contexts (Range) than "blood." According to the WordNet database (Miller et al., 1990; Fellbaum, 2010), "Blood," can have multiple senses, including "family," "lineage," and "liquid part of the body" (taken from wordandphrase.info). The word "hemoglobin" does not fall under the synset (words grouped by the same sense) for any of these senses. The WordNet database was built primarily on human intuitions of words

and relations (psycholinguistics) (Miller et al., 1990), which means words that do not fall under the same synset are likely to evoke different impressions on raters.

3. Lastly, Fritz and Reugg (2013) report their manipulated essays in terms of Low, Medium, and High Lexical Qualities. There are no objective measurements reported for Low, Medium, and High, such as in reference to a corpus. One person's 'Medium' Lexical Frequency might be another person's 'Low,' and may change depending on numerous factors, such as context of writing (high school essays vs fourth year university essays).

In a more recent study, Vögelin et al. (2019) manipulated the LD and LF of four essays (two High Quality, two Low Quality) on the same topic, producing two versions of each essay: High and Low Lexical Richness. The authors provide objective measurements of Lexical Diversity (MTLD (McCarthy, 2005) and D (Malvern & Richards, 2002)) and Lexical Sophistication (Lexical Range values from multiple corpora using TAALES 2.0[1]). The results showed that High Lexical Richness (high LD and more infrequent words) had a significant influence on rater judgement of texts. However, the conclusions that can be drawn are limited by the methodology. The authors state they selected four essays with the "same frame," but the essays are not provided for comparison, and no explanation is given for what I meant by "same frame." Four essays of two different qualities (High and Low) written by four different students are likely to be different in many ways. The question posed to the students ("Do you agree or disagree with the following statement? As humans are becoming more dependent on technology, they are gradually losing

[1] Vögelin et al. (2019) give measurements of Lexical Range to represent Lexical Sophistication, however they used the British National Corpus (BNC, 2007) to manipulate Lexical Frequency of words. I believe reporting measurements of Lexical Frequency would have been more appropriate, which is what we do in this study.

their independence.") is also fairly subjective, and it's difficult to imagine how any two essays would be the same, or even similar. Furthermore, The ANOVA results show that *all* High Quality essays (High + Low Lexical Richness) on average scored higher than *all* Low Quality essays. In addition, *all* essays with High Lexical Richness (High + Low Quality) on average scored higher than *all* essays with Low Lexical Richness. This begs the question: did raters score High Quality essays highly *despite* Low Lexical Richness in some of them? Or did raters score essays with High Lexical Richness highly, *despite* poor quality writing? Based on how the data is aggregated, it is impossible to determine this. Lastly, since LD and LF were manipulated *together* for the same essay, it is impossible to parse the effects of the individual lexical feature on rater judgement.

The experimental design proposed in this study builds on and tries to address some of the limitations in previous research. Specifically, we address the issue of producing ecologically valid ("realistic") student texts, parsing out the effects of LD from LF, and controlling how much LF is changed in an objective and replicable manner.

2.2.5 Literature review summary

The literature review is virtually consistent that as language proficiency increases, for L1 and L2 language users, their use of Lexical Diversity and use of more infrequent words (LF) increases in writing (Arnaud, 1984; Linnarud, 1986; Grobe, 1981; Laufer, 1994; Laufer & Nation, 1995; Tidball and Treffers-Daller, 2008; Crossley & McNamara, 2009; Crossley et al., 2011a; Crossley et al., 2013; Gonzalez, 2017b). Only two studies (drawing from the same relatively small database) were found that appear to contradict this, but these were shown to have limitations

regarding the conditions under which data was collected (Crossley et al., 2010; Kim et al., 2018).

The literature is also virtually consistent on writing assessment scores increasing with increased Lexical Diversity and use of infrequent words (Grobe, 1981; Engber, 1995; Yu, 2009; McNamara et al., 2010; Crossley et al., 2011a; 2011b; Crossley & McNamara, 2012; Gonzalez, 2017a, 2017b; Kim, Crossley & Kyle, 2018; Vögelin et al., 2019). There are a few studies that report no or a negative correlation between LD or LF and written assessments (Ruegg et al., 2011; Fitz & Ruegg, 2013; Xie, 2015), but their findings could be accounted for by research design issues, such as having extremely low-level L2 speakers. However, more recent research using large corpora of writing samples and computational linguistics shows that, while there may be a positive correlation between LF and written assessment scores, such that more use of infrequent words results in higher scores, this effect is insignificant compared to other factors, such as use of collocations (N-grams), words from specific contexts (Word Range),words that are more specific and narrow in meaning (Hypernymy) and use of words with diverse senses (Polysemy) (Kyle & Crossley, 2015; Kyle & Crossley, 2016; Kyle et al., 2017; Kim et al., 2018).

To date there have only been two studies that use experimental designs to show how changes in vocabulary features (LD and LF) affect rater judgement (Fritz & Ruegg, 2013; Vögelin et al. (2019). These studies have design limitations that we hope to improve on in this study.

When measuring the effects of lexical change on holistic rubric scores, it can be difficult to ascertain how much of the score awarded to a essay is due to its lexical quality, rather than

other qualities of the essay, such as strength of argument, cohesion, unity, syntactical prowess, etc. (Read, 2000).

Based on the above information, we propose an experimental design to see whether changes in Lexical Frequency affect written assessment scores when Lexical Diversity is held constant. This will allow us to isolate the effect of one lexical feature while holding the other constant and validate results from previous correlational studies.

Specifically, our research questions are:

1. When holding Lexical Diversity constant as well as other factors that might influence written assessment scores, such as topic, text length, main arguments, sentence structures, number of errors, does a change in Lexical Frequency significantly affect written assessment scores?

2. Based on comments provided by raters, what can we determine about why raters gave a higher, lower, or the same score to essays that differ only at the level of Lexical Frequency? What can we determine about the influence of Vocabulary vs Non-Vocabulary features on rater judgment?

Based on previous research, we hypothesize that, depending on how many words we can change while keeping the essence of the essay the same, there will be a significant ($p < .05$) but small change in written assessment scores, such that the essay with the lower frequency words will receive a higher score. We are unsure whether this result will necessarily reflect itself in rater comments of essays. Previous studies have shown that even highly educated people in the field of linguistics (or similar) are poor judges of Lexical Frequency, especially when word frequencies are closer together (Schmitt & Dunham, 1999; Alderson, 2007; McCrostie, 2007).

However, this does not mean that a change in LF will *not* affect scores, since a change in LF could have a subliminal rather than conscious effect on rater judgement.

The results of this study will help inform rater training and rubric designs for Standardized Language Proficient Tests (SLPTs), like the TOEFL. From a research perspective, we hope to propose guidelines on methodology and reporting on results to allow comparison between other studies of a similar nature. The results can also be used to validate the large body of correlational studies that have been done in the field of Lexical Frequency and Lexical Diversity.

2.3 Methodology

2.3.1 Study design and procedure

We created a within-subjects experimental design with two versions of the same essay (henceforth the treatment essays) modified at two different frequency levels: a High Frequency Essay (HF Essay) and a Low Frequency Essay (LF Essay). Statistical analysis of rater scores and qualitative analysis of rater comments were used to answer the research questions.

The study took place online and was presented to raters using the Qualtrics Online Survey Platform (www.qualtrics.com). Raters were presented with a total of 9 essays on the same topic, including the two treatment essays. To control for exposure effect (raters having seen the 'same' essay already), the following measures were taken:

1. To minimize primacy and recency effects (Ranjith, 2012) on raters, we avoided placing the treatment essays at the beginning and end.

2. At the same time, we presented the treatment essays with as many 'distractor' essays in between (see figure 2).

3. The essays were presented sequentially, with raters being informed that they could not go back once they scored and commented on an essay. This ensured that, even if raters suspected seeing a similar essay, they could not confirm or determine in what ways the essay was similar.

4. Raters were informed they were partaking in a study to determine how essay scores correlated to certain essay indices as calculated by a computer. This was done to ensure that raters did not feel compelled to seek out or penalize essays that they thought were similar / the same (see Fritz and Ruegg (2013) critique in section 2.2.4.2).

5. Rater written comments were analyzed to look for Noticing comments (did raters notice a similarity in essays?) and to determine what, if any, impact the this had on their judgement of essays.

Raters randomly received either the HF Essay or LF Essay first to avoid any order effect.

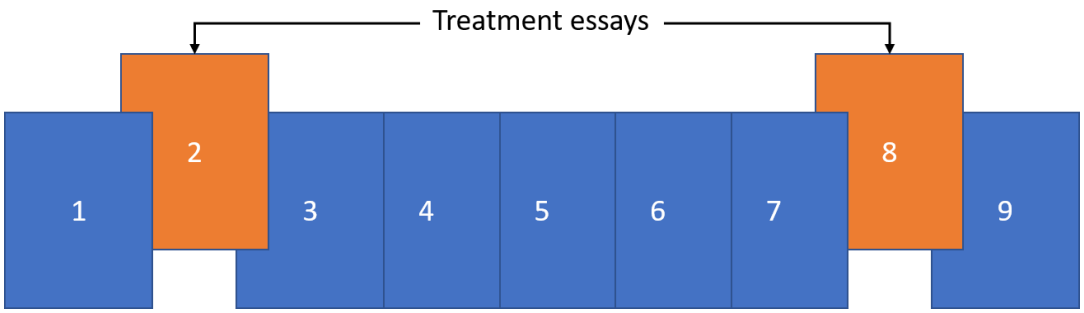

Figure 1: Order of essays presented to raters. Half the raters received the HF Essay first, the other half received the LF Essay first.

Before starting the study, raters were provided with a Training Package which consisted of the TOEFL rubric used to score the essays and five sample essays with varying scores and score explanations taken from the Official Guide to the TOEFL Test, third edition (ETS, 2009). Raters

were asked to score the 9 essays within 1 hour according to the rubric provided. Raters were also told "to provide an explanation for [their] score," in written comments after scoring each essay.

2.3.2 Essay selection

The essays used in the study came from the Independent Writing Task portion of a TOEFL iBT database, provided by Educational Testing Service. According to the Official Guide to the TOEFL Test, third edition, "The independent writing essay is scored on the overall quality of the writing: development, organization, and appropriate and precise use of grammar and vocabulary" (ETS, 2009). This is also reflected in the holistic rubric used to score the essay (Appendix A). 8 essays answering the same prompt were selected with a minimum and maximum score of 2.5 and 4.5 (maximum score one can achieve is 5) to be representative of the larger database.

Table 1: Essays selected from the TOEFL iBT database

Original TOEFL essay scores	Number of essays
2.5	1
3*	3
3.5	2
4	1
4.5	1

**one of these was selected to be the treatment essay*

The treatment essay chosen for modification was selected based on the following criteria:

1. To be a good representative of the rest of the database, an essay with a score of 3.0 (Mode score) was selected.

2. An essay between 300 to 350 words was selected. While there is no minimum or maximum word count criterion set for the Independent Writing Task, TOEFL recommends a minimum of 300 words for an "effective response" (ETS, 2009). NOTE: *all* 8 essays were within 50 words of each other.

3. Of the three essays with a score of 3.0, the one with the highest percentage of Content Words was selected. For simplicity, we only counted Nouns, Verbs, Adjectives, and Adverbs as Content Words. A higher percentage of Content Words would give us more options for modification in producing our High Frequency and Low Frequency versions of the essay.

The final essay chosen for modification had an original score of 3.0, was 348 words in length, and had a Content Word density of 51.44% (179/348 words).

2.3.3 Essay modification

The website wordandphrase.info was used to find words of low and high frequency. Word And Phrase provides individual frequency values for words according to the Corpus of Contemporary American English (COCA; Davies, 2008). It also provides definitions according to WordNet (Fellbaum, 1998) and synonyms for these words ranked by frequency and ordered by Word Sense (synsets), which are the different senses of a word. These tools were used to ensure that appropriately high and low frequency words were selected and that synonyms were based on an objective criterion (Word Sense) rather than the subjective intuition of the researcher. The HF essay was first created/modified from the original essay. Next, the LF Essay

was modified from the HF Essay to maximize the difference in Lexical Frequency. In the final version of the treatment essays, a total of 23.5% (42/179) of Content Words were changed from the HF Essay to the LF Essay (see Appendix B1 and B2 for final version of essays).

23.5% may seem like a small number, especially when considering the objective of this study is to determine whether changes in LF affect rater judgement. However, the number of words we could change was limited by the following conditions:

1. Content Words used in prompts were not changed in either the LF Essay or HF Essay[2]. Research has shown using words from the prompt can influence rater judgement of essays (Plakans & Gebril, 2013; Gebril & Plakans, 2016).

2. The original, unmodified essay had many of the same high frequency words that were also highly polysemous (for example, "good" was used a total of 9 times). Polysemous words have many different (but related) senses. There are many low frequency words that can replace a high frequency word like "good," (e.g., "Excellent," "Decent," "Remarkable") but there are *not* many high frequency words that can do the same. We were limited by how many highly frequent and polysemous words we could replace in the HF Essay to maintain Lexical Diversity in *both* essays.

3. We did not change word phrases ("financially stable," "first of all," "point of view") and common collocations, as these may have unintended and uncontrolled effects on rater judgement.

[2] With the exception of one word that was changed in *both* essays to maximize the difference in LF

The above conditions were necessary for us to conclude that whatever effects we saw in rater judgement came from changes in Lexical Frequency, and to minimize the effects of other, unaccounted variable as much as possible.

All grammatical errors (e.g., wrong word form) were matched between both essays. Spelling errors had to be corrected in order to accurately calculate LD and LF values. However, after calculating LD and LF, we decided not to reinsert spelling mistakes. Spelling errors in a high frequency words (e.g., “basis”) may not leave the same impression on a rater as a spelling error in a low frequency word (e.g., “foundation”). Since there weren’t many egregious spelling mistakes in the original essay to begin with, we thought this an acceptable modification to keep as many variables as possible constant between both essays.

2.3.4 Ecological validation of essays

After a first round of changes, both the LF Essay and HF Essay were presented to three other researchers (two faculty members and a graduate student) in the Applied Linguistics department with experience in teaching and scoring L2 essays. The researchers were asked to make sure that both essays:

1. Conveyed the same ‘essence’ and meaning as each other, despite the frequency changes

2. Were passable as written by L2 writers of mid-level proficiency

After a first round of comments, changes were made to the essays and presented to the three researchers again. During the second round of inspection, all three researchers agreed that

both essay carried the same 'essence' and meaning, and were passable as writing by L2 writers of mid-level proficiency.

2.3.5 Measuring LD and LF

The online tool Coh-Metrix 3.0 (Graesser et al., 2004) and the software TAALES 2.2 (Kyle et al., 2017) were used to calculate Lexical Diversity and Lexical Frequency values, respectively. For Lexical Diversity, we took the Type-Token Ratio (TTR), MTLD (McCarthy, 2005) and VOCD (Malvern et al., 2004) for both treatment essays. For Lexical Frequency, we took the frequency values of Content Words (CW) from three different corpora for both essays: the SUBTLEXus (Brysbaert & New, 2009), the British National Corpus (BNC, 2007), and the Corpus of Contemporary American English (COCA; Davies, 2008). The frequency values from the COCA corpus are most relevant to this study, as this was the corpus that was used to change the frequency of words in both essays. In addition, we present the Range, Polysemy, Hypernymy, and Academic World List (AWL; Coxhead, 2000) values of both essays for comparison and discussion (Table 2).

Table 2: Indices of Lexical Features for the HF and LF Essay, including Lexical Diversity and Lexical Frequency

Description	High Frequency Essay	Low Frequency Essay
Word Count*	347	347
Lexical diversity, type-token ratio, all words	0.42	0.42
Lexical diversity, MTLD, all words	65.34	65.34
Lexical diversity, VOCD, all words	69.17	69.37
SUBTLEXus_Freq_CW	80712.61	72369.34
BNC_Written_Freq_CW	1.05	0.99
COCA_Academic_Frequency_CW	964.85	921.80
COCA_fiction_Frequency_CW	912.20	844.58

COCA_magazine_Frequency_CW	1033.60	971.03
COCA_news_Frequency_CW	1005.91	948.28
COCA_spoken_Frequency_CW	1435.78	1333.05
COCA_All_Frequency_CW (avg)~	1070.47	1003.75
COCA_Academic_Range_CW	0.44	0.38
COCA_fiction_Range_CW	0.42	0.36
COCA_magazine_Range_CW	0.35	0.3
COCA_news_Range_CW	0.36	0.31
COCA_spoken_Range_CW	0.41	0.35
All_AWL_Normed^	0.07	0.11
Polysemy Content Words	12.42	10.20
Hypernymy Nouns and Verbs (Sense Mean, Path Mean)	4.38	4.43

* There is one less word than the original essay
~ TAALES 2.2 does not calculate the word frequency based on the entirety of the COCA database. This was calculated by summing the frequency values from the sub-corpora and dividing it by the number of sub-corpora (5)
^ Fraction of words in text from the Academic Word List (Coxhead, 2000)

We also calculated the lexical frequency profiles of both essays using the online tool Compleat Lexical Tutor (www.lextutor.ca; Cobb, 2019) (Table 3). Unlike the lexical frequency values from TAALES 2.2 (Kyle et al., 2017), lexical frequency profiles tell us what percentage of a text falls under which Frequency Bands, which are bands of the 1000 most frequent words, 2000 most frequent words, etc., according to a reference corpus[3].

Table 3: Lexical Frequency Profiles of the HF and LF Essay

	HF Essay		
	Types (%)	**Tokens (%)**	**Cumulative Token (in %)**
K-1 Words:	118 (80.27)	303 (87.3)	87.3
K-2 Words:	18 (12.24)	29 (8.4)	95.7**^**
K-3 Words:	6 (4.08)	10 (2.9)	98.6*****

[3] Lexical frequency profiles calculated based on BNC (BNC, 2007) and COCA (Davies, 2008)

K-4 Words:			
K-5 Words:			
K-6 Words:	1 (0.68)	2 (0.6)	99.2

	LF Essay		
	Types (%)	**Tokens (%)**	**Cumulative Token (in %)**
K-1 Words:	105 (70.95)	281 (81.0)	81
K-2 Words:	23 (15.54)	39 (11.2)	92.2
K-3 Words:	10 (6.76)	16 (4.6)	96.8**^**
K-4 Words:	3 (2.03)	3 (0.9)	97.7*****
K-5 Words:			
K-6 Words:	3 (2.03)	5 (1.4)	99.1

^ 95% coverage of text
***** 98% coverage of text
Types = number of unique words
Tokens = total number of words

2.3.6 Raters

Raters were selected using snowball sampling methods and by contacting people within the researchers' network. The following criteria were set for rater participation in the research:

Two years of experience in:

1. Scoring essays at the high school level or higher using a holistic rubric (assigning a single overall score to an essay) in any subject; and/or
2. Scoring standardized language proficiency tests (SLPTs) for English, like TOEFL or IELTS (or any other); and/or
3. Any combination of the above two.

We did not limit raters to those with L2 experience because we wanted to increase the probability of hitting optimal participant numbers and because we believe that, if there is an effect of change in LF between the treatment essays, this would be manifest in any rater with relevant experience in scoring essays holistically, whether in an L2 or L1 context. Information on raters was collected, such as years of experience, type of students (L1 and/or L2), and subjects taught (see Table 9).

In the end, a total of 16 raters with a wide range of experience completed the study. On one end, one rater had 2 – 3 years of experience in holistically scoring essays by L2 writers at the high school level or higher, and on the other end one rater had over 10 years of experience solely with scoring essays for Standardized Language Proficiency Tests (SLPTs). Table 9 in the Results section shows rater information by years of experience and type of experience in scoring essays holistically.

2.3.7 Analysis

To answer the first research question, we used a Paired T-test to see if there was a difference in mean scores assigned to the two treatment essays. Paired T-test assumes that data is continuous, independent, normally distributed, and contains no outliers. There is no reason to assume the data (essay scores) wouldn't be normally distributed, but a Shapiro-Wilk Test on both treatment essays was done to test for normality. An Inter-Correlation Coefficient (ICC) was calculated using scores from the 7 non-treatment essays to determine inter-rater reliability.

To answer the second question, rater written comments were coded into Feedback Points, instances of written comments that identify a single reason for which a rater may have awarded or penalized the essay/essayist. Feedback Points were employed by Hyland and Hyland (2001) to code rater comments on student written assignments and has subsequently been used in similar research (Vögelin et al., 2018). Feedback Points were divided into praise (+) and criticism (-) (Hyland and Hyland, 2001), and coded into Vocabulary and Non-Vocabulary categories. Li and Lorenzo-Dus (2014) used a similar coding scheme when investigating the effects of vocabulary on rater judgement of oral proficiency. Our coding scheme was modified to fit our research question and includes lexical features identified in more recent research (Kyle & Crossley, 2015; Kyle & Crossley, 2016; Kyle et al., 2017; Kim et al., 2018) that are said to be predictive of essay scores, such as Polysemy, Hypernymy, and Range (see Appendix C for definitions of coding categories).

The final Feedback Point coding scheme (Table 3) was based on the following considerations:

1. The TOEFL Independent Writing Task rubric that raters used to score the essays (Appendix A)

2. Previous studies that coded rater comments on student writing, either based on a rubric (Vögelin et al., 2018; Li & Lorenzo-Dus, 2014) or by inductively analysing rater comments (Hyland & Hyland, 2001; Barkaoui, 2007).

3. Recent research on the prominence of lexical features in essay scores (Kyle & Crossley, 2015; Kyle & Crossley, 2016; Kyle et al., 2017; Kim et al., 2018).

Two researchers independently identified and coded the Feedback Points. Discrepancies were resolved through discussion, and a 100% consensus was reached in the end.

Table 4: Scheme used to code rater comments. For full definitions see Appendix C

Non-Vocabulary	**Vocabulary**
Organization (Unity, Coherence, Progression, Paragraphs)	Lexical Diversity
Ideas (Use of examples to support position)	Lexical Frequency
Addresses Topic	Collocations and Idiomatic Expressions
Errors (including spelling, grammar, word form)*	Polysemy
Syntactic variety (Variety of sentence structures used)	Hypernymy
Global / Holistic impressions	Word Range (special/specific terms, including academic words)
Other	Word choice
	Vague

*Does NOT include Word Choice errors

2.4 Results

RQ1: When holding Lexical Diversity constant as well as other factors that might influence written assessment scores, such as topic, text length, main arguments, sentence structures, number of errors, does a change in Lexical Frequency significantly affect written assessment scores?

Sixteen raters completed the study. Scores from two raters (**Rater 2** and **Rater 14**) were removed from the statistical analysis because of outlier values (see next section for explanation), but their comments were kept to help answer RQ2. Table 4 shows descriptive statistics for the scores given to the HF Essay and LF Essay by the remaining 14 raters.

Table 5: Descriptive statistics for the HF and LF Essay (N = 14)

Descriptive Statistics	*HF Essay*	*LF Essay*
Mean	3.0	3.1
Median	3.0	3.0
Mode	3.0	3.0
Standard Deviation	0.7	0.7
Minimum Score	2.0	2.0
Maximum Score	4.0	4.5
Total Count	14	14

A Shapiro-Wilk test for normality showed that scores for both essays were normally distributed ($p > .05$; see Table 6). The relatively low value (0.09) for the HF Essay is likely due to the low sample size (14) and would probably be much higher with more data points.

Table 6: Results from Shapiro-Wilk test for normality for the LF and HF Essay

	Statistic	df	Sig.	Kurtosis	Skewness
LF ESSAY	0.926	14	0.267	0.4	0.2
HF ESSAY	0.893	14	0.090	-0.6	0.1

A Paired T-test showed no significant difference in mean scores between the High Frequency Essay and the Low Frequency Essay; $t(13) = .396$, $p = .70$.

The results indicate that changing the Lexical Frequency of 23.5% of Content Words from 1070.47 (High Frequency) to 1003.75 (Low Frequency), based on average COCA Content Word frequencies in a 347 word essay, does not significantly affect written assessment scores.

Figure 2: Box Plot showing distribution of essay scores for the HF and LF Essay.

Table 7 shows ICC estimates for the 7 non-treatment essays based on scores by all 16 raters (including the two that were removed from the Paired T-test). ICC estimates and their 95% confidence intervals were calculated using SPSS based on a mean rating ($k = 16$), absolute

agreement, two-way random effects model. Here we report the ICC based on the mean of *k* = 16 since many standardized language proficiency tests, including TOEFL, are scored by more than one rater and agreement or average between the ratings is sought to determine the final score (ETS, 2009). An ICC with a confidence interval of 0.796 – 0.983 is considered good to excellent reliability (Koo & Li, 2016)[4].

Table 7: Intraclass Correlation (ICC) used to measure Inter-rater reliability between 16 raters.

	Intraclass Correlation (ICC)	95% Confidence Interval		F Test with True Value 0		
		Lower Bound	Upper Bound	Value	df1	df2
Average Measures	.918	.796	.983	17.316	6	90

RQ 2: Based on comments provided by raters, what can we determine about Why raters gave a higher, lower, or the same score to both essays? What can we determine about the influence of Vocabulary vs Non-Vocabulary features on rater judgement?

A total of 133 Feedback Points were generated from the 16 raters for both treatment essays. 86% of the Feedback Points were coded into the Non-Vocabulary category and the remaining 14 % into the Vocabulary category. Number of Feedback Points were evenly distributed across both essays (see Table 7). For a full breakdown of Feedback Points distribution see Appendix D.

[4] Koo and Li (2016) recommend 30 samples for 3 raters. We had 7 samples (essays) for 16 raters.

Table 8: Distribution of Feedback Points across both essays and across the Non-Vocabulary and Vocabulary categories.

	LF Essay	HF Essay	Total
Non-Vocabulary Feedback Points	54	61	115 (86%)
Vocabulary Feedback Points	8	10	18 (14%)
Total	62	71	133 (100%)

Table 9 summarizes all raters by scores they assigned to the treatment essays, the number of Feedback Points they produced, years of experience, and type of experience.

Table 9: Summary of raters, scores assigned to essays, comments, and experience in scoring essays holistically.

Raters	High Frequency Essay Score	Low Frequency Essay Score	Score Change	High Frequency Essay Feedback Points	Low Frequency Essay Feedback Points	Years of experience	Experience Type	Notes on Experience
Rater 1	3	3.5	Increase	3	4	3 - 5	SL, SLPT	-
Rater 2	5	4.5	Decrease	0	0	5 - 10	FL, SL, OTHER	Linguistics (university) English for Academic Purposes
Rater 3	4	3	Decrease	6	4	5 - 10	OTHER	Business subjects.
Rater 4	3	3.5	Increase	8	9	5 - 10	SL	-
Rater 5	3.5	3	Decrease	1	0	10+	FL, SL	-
Rater 6	2.5	3	Increase	3	3	5 - 10	SL	-
Rater 7	2	2	Same	8	5	2 - 3	SL	-
Rater 8	3	3.5	Increase	5	1	3 -5	SL	-
Rater 9	4	4.5	Increase	7	6	2 -3	SLPT, OTHER	Elem. ed.
Rater 10	3	4	Increase	4	4	3 - 5	SL, SLPT	-
Rater 11	2.5	2	Decrease	1	8	10+	SL, OTHER	English Language, French Language, Linguistics (both English and French)
Rater 12	2	3	Increase	4	2	5 - 10	SLPT, OTHER	English Literature, History, IELTS Exams
Rater 13	4	3	Decrease	5	3	10+	FL, SL, SLPT, OTHER	Engineering, Business.
Rater 14	4	0	Decrease	5	1	2 - 3	SL	-
Rater 15	3	2.5	Decrease	6	7	10+	SLPT	-
Rater 16	3	3	Same	5	5	10+	SL	-

LEGEND
SL: Scoring essays by Second Language (L2) writers at the high school level or higher
FL: Scoring essays by First Language (L1) writers at the high school level or higher
SLPT: Scoring essays for Standardize Language Proficiency Tests (IELTS, TOEFL, etc.)

OTHER: Scoring essays by First OR Second Language writers in a subject OTHER than English Language
NOTE: Scores from Rater 2 and Rater 14 were removed from the Paired T-Test, but their written comments were kept for analysis

Seven of the sixteen raters explicitly stated they noticed a similarity in essays. Despite 'noticing', we believe all raters (with the exception of **Rater 14** and **Rater 2**) proceeded to score the essays "as if seeing it for the first time" (**Rater 11**), as can be inferred from their comments and scores (see below).

Seven raters scored the LF Essay higher than the HF Essay, another seven scored the LF Essay lower than the HF Essay, and two raters gave the same score to both essays. Each of these outcomes is discussed below with respect to rater praise (+) and criticism (-) for each essay.

2.4.1 Raters that scored the LF Essay higher than the HF Essay

Of the seven raters that scored the LF Essay higher than the HF essay, none of them explicitly stated they noticed the treatment essays (i.e., raters did not comment on whether they saw a similar/same essay previously). Five of the seven raters increased their score by 0.5, and two others by 1. Four of the seven raters explicitly mentioned 'Vocabulary' in their comments. Each of these outcomes is discussed with respect to rater praise (+) and criticism (-) for each essay, and with what can be inferred about the influence of Vocabulary vs Non-Vocabulary features on rater judgement.

Rater 1 provides similar criticism and praise for both essays, except for an additional criticism for the LF Essay: "The first and second reason both focus on financial stability or security." This comment refers to how the writer deviates from the question asked (choosing to study subjects of interest vs subjects that will help with a career), i.e. not addressing the topic. Despite the additional criticism, the LF Essay is awarded a higher score by 0.5. **Rater 1** makes no comments on vocabulary for either essay.

Rater 4 praises the LF Essay for using "good vocabulary" and "strong collocations," but criticizes the HF Essay for having limited "academic vocabulary." **Rater 4** mentions more errors in the LF Essay, including errors in sentence fragments, Run-On-Sentence, and Word Forms, however this does not stop them from awarding the LF Essay a higher score by 0.5

Rater 6 praises both essays for use of "transition words" and criticizes both essays for "word choice." In addition, they criticize the LF Essay for improper "preposition and article" use, but this is not mentioned for the HF Essay. **Rater 6** makes no other comments on vocabulary for either essay.

Rater 8 takes issue with how neither essay "Addresses Topic." This is elaborated more substantially in his/her comment of the LF Essay:

> "Doesn't effectively address the topic, which is whether a person should study SUBJECTS they are interested in, or subjects to prepare for a job or career. The writer focuses on why it is important to have/get a good job."

I infer that **Rater 8** emphasized this point in the LF Essay because they felt it was necessary to justify their score after seeing the same/similar essay a second time, even though they provided no Noticing comment. The rater still awarded the LF Essay an increased score of 0.5, though the tone/wording of the comment suggests that the LF Essay should be further penalized (even though both essays have addressed the topic in the exact same manner). **Rater 8** makes no comments on vocabulary for either essay.

Rater 9 praised both essays for addressing the topic and organization. The HF Essay was praised for its ideas and "good explanations," but this is not mentioned for the LF Essay. The LF Essay

was further criticized for “errors in word usage,” though this was not mentioned for the HF Essay. The HF and LF Essay were scored 4 and 4.5 respectively. Of the 16 raters, **Rater 9** has one of the lowest experiences in marking essays holistically (2-3 years) and no experience in teaching Second Language students (though they do have experience marking SLPTs). This may explain the relatively high scores awarded to both essays. Other than “errors in word usage” for the LF Essay, **Rater 9** makes no other comments on vocabulary for either essay.

Rater 10 praised both essays for the “range of vocabulary,” however criticized the LF Essay for “collocation errors.” They also praised the LF Essay for “unique and expanded ideas” but made no such mention for the HF Essay.

Rater 12 criticized the HF Essay for multiple errors, including errors in word choice, word forms, and grammar. In fact, **Rater 12** only had critical comments for the HF essay, while the LF Essay is praised for addressing the topic and “using somewhat developed explanations, exemplifications and/or details” and not criticized for any of the issues that were apparent in the LF Essay. Other than criticizing the LF essay for errors in word choice, **Rater 9** makes no other comments on vocabulary for either essay.

From the above analysis we can see that most raters were inconsistent in their praise and criticism for both essays, and even though some raters had more critical Feedback Points for the LF Essay, they still ended up awarding the LF Essay a higher score (**Raters 1**, **4**, **6**, **9**). The inconsistency and distribution of comments makes it impossible to infer any one reason for raters awarding a higher score to the LF Essay.

Of the 7 raters that scored the LF Essay higher than the HF Essay, only 1 (**Rater 4**) correctly criticized the HF Essay for limited "academic vocabulary" (see Table 2 for comparison of academic vocabulary in the treatment essays). **Rater 4** also praised the LF Essay for "good vocabulary," however since this was not elaborated, we cannot be sure if they are referring to the Low Frequency words that were present. The limited reference to Vocabulary features in general would indicate that raters were not greatly influenced by the changes in lexical frequency, *as far as we can infer from their written comments*.

2.4.2 Raters that scored the LF Essay lower than the HF Essay

Of the seven raters that scored the LF Essay lower than the HF essay, 6 of them explicitly stated they noticed the similarity in essays (**Rater 11**: "This is a repeat of a prior piece"). Four of the seven raters decreased their score by 0.5, two decreased their score by 1, and one rater (**Rater 14**) gave the LF Essay a 0 for "plagiarism." Each of these outcomes is discussed with respect to rater praise (+) and criticism (-) for each essay, and what can be inferred about the influence of Vocabulary vs Non-Vocabulary features on rater judgement.

Rater 2 provided 0 comments for both essays, however they gave a score of 4.5 and 5 to the LF and HF Essay, respectively. I speculate that **Rater 2** encountered the HF Essay after the LF Essay (there is no way to be certain with the way the survey data was generated), and scored the HF Essay a 5 – the only 5 awarded to all 9 essays by any rater – because he/she was not sure what to do with it, or suspected an error in the survey. All of **Rater 2**'s other essay scores and comments seem to be within the expected range. A score of 5 is not only an outlier in the HF and LF Essay data, but for the entire dataset, thus **Rater 2**'s scores were removed from the statistical analysis.

Rater 3 provided praised the HF Essay for structure, grammar, and general organization of the essay ("Does tie the two sides of the argument together […] flows very well") and vague/holistic impressions such as "English. . .[is] very solid." A mix of praise and criticism was offered for the LF Essay, but there was no overlap in the praise **Rater 3** gave the HF Essay. **Rater 3** made no comment on vocabulary for either essay.

Rater 5 provided no comments for the LF Essay and only provides a noticing comment for the HF Essay: "(Wasn't this the first sample?)"

Rater 11 criticizes the LF Essay for its organization, ideas, lack of phrasing diversity, and errors. They only make a Noticing comment for the HF Essay:

> "This is a repeat of a prior piece. Since I am unable to return and see the rating I gave that one, I'll proceed as if seeing it for the first time."

Despite stating this is a "repeat of a prior piece," after proceeding "as if seeing it for the first time," **Rater 11** awards the HF Essay an extra score of 0.5.

Rater 13 praises the HF Essay for its clear thesis and progression of ideas and examples, and criticizes it for its "structure" and "grammar." None of the same criticisms or praises are mentioned for the LF Essay. **Rater 13** notices the similarity in essays and comments on the HF Essay: "This sample was / is the same as was previously listed in the nine essays, OR it is very similar! I can't for sure remember what grade that I assigned, but I think it should be a 3.5 or a 4 because […]" The LF Essay was in fact awarded a 3.0, and the HF Essay was awarded 4.0. The rater goes on to praise the HF Essay for its organization and criticize it for its grammar, both points that are missing for the LF Essay. I infer that, since **Rater 13** noticed the similarity in the

previous essay, it forced them to pay more attention to the HF Essay. Despite this, **Rater 13** makes no mention of vocabulary for either essay.

Rater 14 penalized the LF Essay for "plagiarism" and gave it a 0. The HF Essay was given a score of 4 and praised for organization and syntactic variety, however was criticized for "word-choice errors" and grammar errors. **Rater 14**'s scores were removed from the statistical analysis.

Rater 15 provides the same comments, almost verbatim, for both essays, praising them for general impressions and addressing the topic, and criticizing them for word choice and errors in grammar. An additional comment is provided for the LF Essay: "but more information could be included to ensure message is clear." **Rater 15** noticed the essays and begins his/her comments for the HF Essay with the following before explaining his/her score:

"This response is the same as the first one in this series of nine responses. I give this response a 3 in its first appearance here and I give this response a 3 in its second appearance here of the 9 responses being reviewed." **Rater 15** actually awarded the LF Essay a score of 0.5 less than the HF Essay.

From the above analysis we can see that, even though 6 of the raters appeared to have noticed that two essays in the dataset had very strong resemblance, not one of them mentioned 'Vocabulary,' or anything related, in their comments for *either* essay. This is true for even the most experienced raters, with **Rater 15** having 10+ years of experience in scoring SLPTs, like the ones presented in this study, and **Rater 13** with 10+ years of scoring essays holistically by L1 and L2 writers at the high school level or higher, scoring SLPTs, and essays for business and engineering courses.

There is no apparent trend in the Feedback Points that sheds light on why these raters scored one essay higher or lower than the other. The limited reference to Vocabulary features in general would indicate that raters were not greatly influenced by the change in lexical frequency, *as far as we can infer from their written comments*.

2.4.3 Raters that gave the same score to both essays

Of the two raters that assigned the same score to both essays, only one of them explicitly stated they noticed the resemblance between the two treatment essays. Each of these outcomes is discussed with respect to rater praise (+) and criticism (-) for each essay, and what can be inferred about the influence of Vocabulary vs Non-Vocabulary features on rater judgement.

Rater 7 assigned both essays a score of 2. Both the LF and HF Essays were criticized for not addressing the topic and for errors in grammar, though the rater further elaborated his/her stance by mentioning the lack of lexical diversity and 'sophistication' in the HF Essay: "Their vocabulary choices are not very varied: mostly everyday words used in ordinary conversation."

Rater 16 assigned both essays a score of 3. Both essays were criticized for not addressing the topic and lack of organization ("Conclusion does not summarize major ideas"). The HF Essay is praised for "sentence structures and transitions" and for academic vocabulary, both points that are missing for the LF Essay.

From the above analysis we can see that **Rater 7** is the *only* rater of 16 that correctly mentions the lack of lexical 'sophistication' in the HF Essay, criticizing the vocabulary as "mostly everyday

words used in ordinary conversation," and how the writing is in fact "like conversation written down" (see Appendix B1 and B2 for final treatment essays).

2.5 Discussion

Changing the Lexical Frequency of 23.5% of Content Words from 1070.47 to 1003.75, based on average COCA Content Word frequencies in a 347 word essay, did not significantly affect written assessment scores.

With regards to rater comments, two things become apparent: 1) Non-vocabulary features have far more influence on raters' conscious judgement than vocabulary features, and 2) halo effects and rater inconsistences (Knoch et al., 2007) are a source of possible variance in scores.

With regards to why there was no difference in mean scores, it's possible that the changes in Lexical Frequency were far too small to influence rater judgement, whether subliminally or consciously. However, this begs the question: *should* the changes in Lexical Frequency have influenced rater judgment?

These points are discussed below.

2.5.1 The influence of non-vocabulary vs vocabulary features and rater effects

From analysing rater's Feedback Points, it becomes apparent that Non-Vocabulary features (86% of all Feedback Points) are far more influential to rater judgement than Vocabulary features, as far as we can infer from the written comments. This could be a quirk of holistic rubrics, as opposed to analytic rubrics, where raters' attention is not explicitly drawn to specific essay features (like vocabulary).

Previous research has shown that, when raters are asked to judge and comment on essay scores, vocabulary has little to no effect on rater judgement, *as far as we can infer from rater self-reported explanations and decision-making behaviours*. For example, Cumming, Kantor, and Powers (2002) looked at the decision-making behaviours of raters and what essay features they most attended to when asked to judge essays *without* reference to a rubric. Through rater Think-Aloud Protocols, the authors showed that rhetorical features (organization, cohesion, ideas) were attended to substantially more than language, with "Consider Lexis" comprising only 2.6% of all comments by raters. In a similar study, Barkaoui (2007) looked at the decision-making behaviours of raters judging essays, based on a holistic *and* analytic rubric, using the same coding scheme developed by Cumming et al. (2002). Of the 30 Think-Aloud Protocols averaging 27 comments each, "Consider Lexis" only comprised of 3 – 4% of all comments. Furthermore, written comments by the same raters (separate from Think-Aloud Protocols) explaining their scores showed 0 comments on lexis.

Even when asked to explicitly judge texts based on vocabulary, Non-Vocabulary features seem to be far more prominent to raters. When Li and Lorezno-Dus (2014) asked raters to judge an oral text specifically for its quality of vocabulary, the majority of rater comments still ended up being coded in the Non-Vocabulary category (57%: pronunciation, fluency, grammar, etc.). In this case, since this is an oral text (as opposed to written), a comparison might be unjustified. As the authors suggest, when it comes to oral texts specifically, it might be difficult for raters to distinguish certain text features from one another. For example, if raters cannot understand what is being said (pronunciation), it might be impossible for them to give a proper judgment of Vocabulary.

However, a similar phenomenon could be occurring with written texts in the form of halo effects (Thorndike, 1920). In assessment scoring, halo effects refer to when the judgement of one feature of an assessment (say, Vocabulary in writing) affects the judgement of another feature in the same assessment (say, Grammar) (Knoch et al., 2007). For example, Vögelin et al. (2018) manipulated Lexical Frequency, Lexical Diversity and Spelling mistakes to see how it affected teacher comments on essays. Results showed that texts that had high Lexical Diversity and more infrequent words also received more positive comments on grammar, and texts with more spelling mistakes received more negative comments on Vocabulary, Grammar, and other aspects of writing. We cannot infer any specific halo effects of Lexical Frequency on other essay features in the present study, due to limited data. However, it seems that, if the presence of High or Low Frequency words gives raters an overall positive OR negative impression of the essay, this is transferred into their judgement of the essay as a whole. This is shown in **Rater 3**'s comments of how the HF essay "[…] is so much better," (though they don't state *how* it's better) and how they proceed to give only praise for the HF Essay, while the LF Essay only received criticism. **Rater 12**, who only had criticisms to offer the HF Essay (including "inappropriate choice of words"), only had praise to offer the LF Essay, including how it addressed the topic and had "developed […] exemplification and/or details," points that are identical in both essays.

This could also simply be rater inconsistency (intra-rater reliability), which, along with halo effects, is another possible source of rater bias and error (Myford & Wolfe, 2003; Knoch et al., 2007). Some variation is to be expected between raters, but comments showed inconsistency by the same raters who end up criticizing or praising *different* aspects of both essays that were

in fact the same. For example, **Rater 4** criticizes errors in the LF Essay ("sentence fragments," "word forms," "Run On Sentences"), but fails to mention these same errors in the HF Essay. **Rater 10** praises both the LF and HF essay for "range of vocabulary," however they only praise the LF essay for "unique and expanded ideas."

Whether this is rater inconsistency or a halo effect, or raters simply being selective about what they choose to comment on, is difficult to ascertain.

2.5.2 Should raters have been influenced by the change in lexical frequency?

It is possible that the difference in Lexical Frequency between the two essays was simply not great enough to influence rater judgment, whether this judgement be subliminal or conscious.

When we look at the COCA Content Word Lexical Frequency counts for both essays (1070.47 and 1003.75), there is no reference guide to tell us how 'high' is high and how 'low' is low. There is no guide to show us how *much* lower the average lexical frequency of a text must be before a rater can be expected to award it a higher score.

However, as a surrogate, researchers in corpus-linguistics have created Frequency Bands (bands of the 1000 most frequent words, 2000 most frequent words, etc.) based on how many word families a reader needs to know to sufficiently comprehend a text. For example, Schmitt and Schmitt (2014) suggest that learners ought to know the first 3000 most frequent words in English to sufficiently understand graded reading material (covering 98% of vocabulary in the text). Based on these numbers, Schmitt and Schmitt (2014) consider the first 3000 word families to be High Frequency, and the 3000 to 9000 most frequent word families to be Mid Frequency. If we look at the lexical frequency profiles of both treatment essays (Table 3), we

see that the HF Essay has a 98% coverage of High Frequency words (up to 3000), with only 1 word being from the 6000 level. In contrast, the LF Essay only reaches 95% coverage at the 3000 level, with 6 words from the 4000 and 6000 level (Mid Frequency). This is by no means a large difference, especially when we consider that humans seem to be poor judges of ranking words by frequencies that are really close together.

When differences in lexical frequency is *explicitly* drawn attention to, some studies show that trained individuals (language teachers or people in the field of linguistics) can differentiate frequent from infrequent words when there is a large difference in frequencies (High vs Low Frequency words) (Tidball & Treffers-Daller, 2008; Schmitt & Dunham, 1999). However, when word frequencies are much closer, even highly educated individuals (in linguistics or related fields) seem to have subpar and inconsistent performance in ranking words by frequencies (Schmitt & Dunham, 1999; Alderson, 2007; McCrostie, 2007). From a psycholinguistic perspective of learning, storing, and recalling word frequencies, Ellis (2002) agrees that most humans are bad intuitive statisticians when it comes to accurately judging the frequency of words.

Based on the above information, perhaps it should not be a surprise that the change in Lexical Frequency had little to no effect on rater scores and comments, subliminally or otherwise. However, the more important question for language assessment is, *should* it have had an affect?

As mentioned in the beginning, Vocabulary is an imperative part of all language assessments, written or spoken, and is usually judged as an independent construct in all SLPTs, as can be

determined by the wording of the rubric (holistic or analytical) being used to score the assessment. Of the 16 raters, only two raters accurately criticized the High Lexical Frequency in the treatment essay, with **Rater 7** commenting on the highly common words used in the HF Essay ("mostly everyday words in ordinary conversation") and **Rater** 4 commenting on the "limited academic vocabulary" of the same essay. This is quite surprising given that 7 of the 16 raters (which did not include **Rater 7** and **4**) noticed a resemblance between the two treatment essays, but none of the 7 mentioned "vocabulary" in their comments.

This might be different if the raters were using an analytic rubric that drew their attention to the use of vocabulary. However, this too would not guarantee a change in rater judgement, because raters would then need to determine if the change in Lexical Frequency is *enough* to justify giving the writer a higher score. If a writer uses the words "foundation," "esteemed," and "expensive," rather than "basis," "respected," and "high," is that enough for them to earn a slightly higher score? How *many* words need to be of a sufficiently low frequency for raters to judge it appropriate to give a higher score? Does the context of the writing matter? When raters judge an essay for TOEFL vs a fourth year's honour thesis, does the 'standard' of Lexical Frequency (how low a frequency a word should be) shift in the mind of the rater? For that matter, is "frequency" even the correct construct by which raters should judge an essay? Many studies (McNamara et al., 2010; Crossley et al., 2011a; 2011b; Crossley & McNamara, 2012; Gonzalez, 2017a, 2017b; Kyle & Crossley, 2015; Kyle & Crossley, 2016; Kyle et al., 2017; Kim et al., 2018) have shown a correlation between low frequency words and written assessment scores, but how is word frequency interpreted in the mind of a rater? Could it be that other lexical features (like word Range, Polysemy, Hypernymy) that happen to correlate with Lexical

Frequency are better theoretical constructs that help explain rater judgements? As can be seen in Table 2, all these lexical features changed with Lexical Frequency. As Kim et al. (2018) suggest, the otherwise vague and abstract concept of Lexical Sophistication should perhaps be considered a multi-dimension theoretical construct, which consists of much more than just Lexical frequency.

2.6 Limitations

The present study and what can be inferred from its results are limited in the following ways:

1. We cannot conclude that just because certain features of an essay were missing from a rater's comments that those features did not influence the rater's judgement. At most we can say that raters commented on the essay features that they were *most* aware of affecting their judgement, not *all* the essay features that affected their judgement.

2. Though none of the 7 raters who noticed the similarity in the treatment essays pin-pointed what was similar (or different), we cannot conclusively say this did not affect their judgment. For example, it is possible that the raters tried to recall what they gave the original essay they came across (whether the HF or LF Essay), and then tried to match or justify the score to the second essay. This may have caused them to give a lower or higher score than they would have otherwise.

3. The sample size for the quantitative analysis (14 raters) is fairly small. Though the Paired T-test is robust against small sample sizes, a larger samples size would give more conclusive results.

2.7 Future research and recommendations

I recommend the following guidelines for future studies in lexical feature manipulation, for the sake of replicability and ecological validity:

1. Stating what percentage of words in a text have been changed.

2. Stating measured and objective changes in manipulated texts based on known lexical indices, such as those generated by TAALES (Kyle et al., 2017) or Coh-Metrix 3.0 (Graesser et al., 2004).

4. Using humans to judge the ecological validity of manipulated essays (are these essays passable as L2 writing?)

5. Presenting the manipulated essays as part of the appendices for critique from the larger research community.

These guidelines will allow researchers to reliably compare, contrast, and critique results in the field of written assessment research.

Lastly, we recommend a replication of the present study using a between-subjects rather than within-subjects design. We used a within-subjects design because of limited access to participants and financial resources. A good between-subjects design would require greater participants and would remove any possible exposure effects on raters from seeing two similar essays in a single dataset.

2.8 Conclusion

This study set out to determine whether a measured and objective change in Lexical Frequency in two otherwise similar essays affected rater judgements of the essays as determined by their

scores and comments. Unlike previous correlational or experimental studies, this study successfully kept the Lexical Diversity of both treatment essays constant to isolate the effects of Lexical Frequency. A Paired T-Test of 14 raters' holistic scores showed no difference in the mean score of both essays. Rater comments showed that Non-Vocabulary features had a far greater influence on their judgement than Vocabulary features. Raters were inconsistent in their praise and criticisms of both essays even when commenting on the same/similar feature. Changes in Lexical Frequency inevitably changed other lexical features, such as Range, Academic Word percentage, Polysemy and Hypernymy. These other lexical features may serve as better theoretical constructs to explain why, if at all, changes in Lexical Frequency affect rater judgement (although no effect for these other lexical features emerged here either). We recommend that future experimental designs in vocabulary and written assessment research follow a similar format of presenting results and state how many words are changed and by how much, to allow for study comparisons. We also recommend a replication study using a between-subjects design to remove any possible exposure effects on raters and to validate the results of the present study.

CHAPTER 3: CONCLUSION

The present study was done to determine whether changes in Lexical Frequency affect rater judgement of essays, as determined by scores assigned to essays and rater comments. I conducted an experimental study using a within-subjects design, where raters were presented with two essays (a Low Frequency Essay and a High Frequency Essay) to score and comment on. The essays were similar in every single way, including having the same Lexical Diversity, text length, main arguments, sentence structures, and number of errors. We assured that both essays conveyed the same meaning, despite changes in Lexical Frequency, by having three independent researchers with ESL experience judge the content of the essays.

Based on scores by fourteen raters, statistical analysis shows that changing the Lexical Frequency of 23.5% of Content Words from 1070.47 (High Frequency) to 1003.75 (Low Frequency), based on average COCA Content Word frequencies in a 347 word essay, does not significantly affect written assessment scores.

Based on comments by sixteen raters, we can see that Non-Vocabulary features have a far greater influence on rater judgement than Vocabulary features. Previous studies support this observation. Cumming, Kantor, and Powers (2002) looked at the decision-making behaviours of raters and what essay features they most attended to when asked to judge essays *without* reference to a rubric. Through rater Think-Aloud Protocols, the authors showed that rhetorical features (organization, cohesion, ideas) were attended to substantially more than language, with “Consider Lexis” comprising only 2.6% of all comments by raters. In a similar study, Barkaoui (2007) looked at the decision-making behaviours of raters judging essays, based on a holistic *and* analytic rubric, using the same coding scheme developed by Cumming et al. (2000).

Of the 30 Think-Aloud Protocols averaging 27 comments each, “Consider Lexis” only comprised of 3 – 4% of all comments. Furthermore, written comments by the same raters (separate from Think-Aloud Protocols) explaining their scores showed 0 comments on lexis. Even when raters are explicitly asked to judge texts based on vocabulary, Non-Vocabulary features seem to be far more prominent. Li and Lorenzo-Dus (2014) asked raters to judge an oral text specifically based on its vocabulary, yet most rater comments still pertained to Non-Vocabulary features, like pronunciation, fluency, and grammar. It is possible that for raters many of these supposedly independent constructs (pronunciation vs vocabulary) are hard to distinguish from each other, or have an interaction affect.

This is similar to what Thorndike (1920) labelled halo effect. In assessment scoring, halo effects refer to when the judgement of one feature of an assessment (say, vocabulary in writing) affects the judgement of another feature in the same assessment (say, grammar) (Knoch et al., 2007). Previous studies have implied halo effects of one essay feature on another. For example, Vögelin et al. (2018) manipulated Lexical Frequency, Lexical Diversity and spelling mistakes to see how it affected teacher comments on essays. Results showed that texts that had high Lexical Diversity and more infrequent words also received more positive comments on grammar, and texts with more spelling mistakes received more negative comments on vocabulary, grammar, and other aspects of writing.

Due to limited data we cannot infer any specific halo effects of Lexical Frequency on other essay features in the present study. However, it seems if the presence of High or Low Frequency words give raters an overall positive OR negative impression of the essay, this is transferred into their judgement of the essay as a whole. This is shown in **Rater 3**’s comments

of how the HF essay "[…] is so much better," (though they don't state *how* it's better) and how they proceed to give only praise for the HF Essay, while the LF Essay received only received criticism. **Rater 12**, who only had criticisms to offer the HF Essay (including "inappropriate choice of words"), only had praise to offer the LF Essay, including how it addressed the topic and had "developed […] exemplification and/or details," points that are identical in both essays.

This could also simply be rater inconsistency which, along with halo effects, are two possible sources of rater bias and error (Myford & Wolfe, 2003; Knoch et al., 2007).

Lastly, it is possible that the difference in Lexical Frequency between the two essays was simply not great enough to influence rater judgment, whether this judgement be subliminal or conscious. When we look at the COCA Content Word Lexical Frequency counts for both essays (1070.47 and 1003.75), there is no reference guide to tell us how 'high' is high and how 'low' is low. There is no reference to show how *much* lower the average lexical frequency of a text ought to be before a rater awards it a higher score.

We acknowledge that the present study is limited in how far we can infer from rater comments on what influenced their scoring of the essays. We are also unsure of how noticing the treatment essays, in a series of essays, could have affected rater judgement, other than saying that raters were unable to articulate what exactly was different/similar between the two treatment essays.

We recommend that future experimental designs in vocabulary and written assessment research follow a similar format of presenting results and state how many words are changed and by how much, using objective indices of lexical measures, such as those generated by

TAALES (Kyle et al., 2017) and Coh-Metrix 3.0 (Graesser et al., 2004). Lastly, we recommend a replication study using a between-subjects design to remove any possible exposure effects on raters and to validate the results of the present study.

REFERENCES

Alderson, J. C. (2007). Judging the Frequency of English Words. *Applied Linguistics*, *28*(3), 383–409. https://doi.org/10.1093/applin/amm024

Arnaud, P. J. L. (1984). The lexical richness of L2 written productions and the validity of vocabulary tests. *Practice and Problems in Language Testing*, 14–28. Colchester, England: Department of Language and Linguistics.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86–107. https://doi.org/10.1016/j.asw.2007.07.001

BNC. (2007). *British National Corpus, version 3* (BNC XML ed.). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Cobb, T. Range for texts v.3 [computer program]. Accessed 15 June 2019 at https://www.lextutor.ca/cgi-bin/range/texts/index.pl

Coltheart, M. (1981). The MRC Psycholinguistic Database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497–505. https://doi.org/10.1080/14640748108400805

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System*, *41*(4), 965–981. https://doi.org/10.1016/j.system.2013.08.002

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

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115–135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011a). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561–580. https://doi.org/10.1177/0265532210378031

Crossley, S. A., Weston, J. L., McLain Sullivan, S. T., & McNamara, D. S. (2011b). The Development of Writing Proficiency as a Function of Grade Level: A Linguistic Analysis. *Written Communication*, *28*(3), 282–311. https://doi.org/10.1177/0741088311410188

Crossley, S., Salsbury, T., & McNamara, D. (2010). The Development of Polysemy and Frequency Use in English Second Language Speakers. *Language Learning*, *60*(3), 573–605. https://doi.org/10.1111/j.1467-9922.2010.00568.x

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision Making while Rating ESL/EFL Writing Tasks: A Descriptive Framework. *The Modern Language Journal*, *86*(1), 67–96. https://doi.org/10.1111/1540-4781.00137

Davies, M. (2008-) *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. Available online at https://corpus.byu.edu/coca/

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). Factors in Judgments of Writing Ability. *ETS Research Bulletin Series*, *1961*(2), i–93. https://doi.org/10.1002/j.2333-8504.1961.tb00286.x

Ellis, N. (2002). Reflections on Frequency Effects in Language Processing. *Studies in Second Language Acquisition*, 24(2), 297–339. https://doi.org/10.1017.S0272263102002140

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*(2), 139–155. https://doi.org/10.1016/1060-3743(95)90004-7

Fellbaum, C. (2010). WordNet. In R. Poli, M. Healy, & A. Kameas (Eds.), *Theory and Applications of Ontology: Computer Applications* (pp. 231–243). https://doi.org/10.1007/978-90-481-8847-5_10

Ferris, D. R. (1994). Lexical and Syntactic Features of ESL Writing by Students at Different Levels of L2 Proficiency. *TESOL Quarterly*, *28*(2), 414–420. https://doi.org/10.2307/3587446

Frase, L. T., Faletti, J., Ginther, A., & Grant, L. (1998). Computer Analysis of the Toefl Test of Written English. *ETS Research Report Series*, *1998*(2), i–26. https://doi.org/10.1002/j.2333-8504.1998.tb01791.x

Fritz, E., & Ruegg, R. (2013). Rater sensitivity to lexical accuracy, sophistication and range when assessing writing. *Assessing Writing*, *18*(2), 173–181. https://doi.org/10.1016/j.asw.2013.02.001

Gebril, A., & Plakans, L. (2016). Source-based tasks in academic writing assessment: Lexical diversity, textual borrowing and proficiency. *Journal of English for Academic Purposes*, *24*, 78–88. https://doi.org/10.1016/j.jeap.2016.10.001

González, M. C. (2017a). Profiling Lexical Diversity in College-level Writing. *Vocabulary Learning and Instruction*, *6*(1), 61–74. https://doi.org/10.7820/vli.v06.1.Gonzalez

González, M. C. (2017b). The Contribution of Lexical Diversity to College-Level Writing. *TESOL Journal*, *8*(4), 899–919. https://doi.org/10.1002/tesj.342

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3758/BF03195564

Grobe, C. (1981). Syntactic Maturity, Mechanics, and Vocabulary as Predictors of Quality Ratings. *Research in the Teaching of English*, *15*(1), 75–85.

Hyland, F., & Hyland, K. (2001). Sugaring the pill Praise and criticism in written feedback. *Journal of Second Language Writing, 10*, 185-212.

Jarvis, S. (2013). Capturing the Diversity in Lexical Diversity: Lexical Diversity. *Language Learning*, *63*, 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Johnson, W. (1939). *Language and Speech Hygiene: An Application of General Semantics. Outline of a Course*. Institute of General Semantics.

Johnson, W. (1944). I. A program of research. *Psychological Monographs*, *56*(2), 1–15. https://doi.org/10.1037/h0093508

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical Sophistication as a Multidimensional Phenomenon: Relations to Second Language Lexical Proficiency, Development, and

Writing Quality. *The Modern Language Journal*, *102*(1), 120–141. https://doi.org/10.1111/modl.12447

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*(1), 26–43. https://doi.org/10.1016/j.asw.2007.04.001

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction*, *01*(1), 60–69. https://doi.org/10.7820/vli.v01.1.koizumi

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12–24. https://doi.org/10.1016/j.jslw.2016.10.003

Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly*, *49*(4), 757–786. https://doi.org/10.1002/tesq.194

Kyle, K., Crossley, S., & Berger, C. (2017). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*. https://doi.org/10.3758/s13428-017-0924-4

Laufer, B., & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics*, *16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307

Laufer, B. (1994). The Lexical Profile of Second Language Writing: Does It Change Over Time? *RELC Journal*, *25*(2), 21–33. https://doi.org/10.1177/003368829402500202

Li, H., & Lorenzo-Dus, N. (2014). Investigating how vocabulary is assessed in a narrative task through raters' verbal protocols. *System*, *46*, 1–13. https://doi.org/10.1016/j.system.2014.06.006

Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. Malmö: Gleerup.

Malvern, D. D., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Houndmills, NH: Palgrave Macmillan.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, *19*(1), 85–104. https://doi.org/10.1191/0265532202lt221oa

McCarthy, P. M. (2005). *An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD)*. 200.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McCrostie, J. (2007). Investigating the Accuracy of Teachers' Word Frequency Intuitions. *RELC Journal*, *38*(1), 53–66. https://doi.org/10.1177/0033688206076158

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic Features of Writing Quality. *Written Communication*, *27*(1), 57–86. https://doi.org/10.1177/0741088309351547

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An On-line Lexical Database *. *International Journal of Lexicography*, *3*(4), 235–244. https://doi.org/10.1093/ijl/3.4.235

Myford, C. M., & Wolfe, E. W. (n.d.). *Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I*. 38.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge ; New York: Cambridge University Press.

O'Loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing*, *12*(2), 217–237. https://doi.org/10.1177/026553229501200205

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing*, *22*(3), 217–230. https://doi.org/10.1016/j.jslw.2013.02.003

Ranjith, N. (2012). Serial Position Curve. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 3050–3052). https://doi.org/10.1007/978-1-4419-1428-6_1816

Read, J. (2006). *Assessing vocabulary* (6. printing). Cambridge: Cambridge Univ. Press.

Ruegg, R., Fritz, E., & Holland, J. (2011). Rater Sensitivity to Qualities of Lexis in Writing. *TESOL Quarterly*, *45*(1), 63–80.

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical profiles of comprehensible second language speech. *Studies in Second Language Acquisition*, *38*(04), 677–701. https://doi.org/10.1017/S0272263115000297

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

Schmitt, N., & Dunham, B. (1999). Exploring native and non-native intuitions of word frequency. *Second Language Research*, *15*(4), 389–411. https://doi.org/10.1191/026765899669633186

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(04), 484–503. https://doi.org/10.1017/S0261444812000018

ETS. (2009). *The official guide to the TOEFL® test third edition*. New York, NY: McGraw-Hill

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*(1), 25–29. https://doi.org/10.1037/h0071663

Tidball, F., & Treffers-Daller, J. (2008). Analysing lexical richness in French learner language: What frequency lists and teacher judgements can tell us about basic and advanced words. *Journal of French Language Studies*, *18*(03), 299–313. https://doi.org/10.1017/S0959269508003463

Vögelin, C., Jansen, T., Keller, S. D., Machts, N., & Möller, J. (2019). The influence of lexical features on teacher judgements of ESL argumentative essays. *Assessing Writing*, *39*, 50–63. https://doi.org/10.1016/j.asw.2018.12.003

Vögelin, C., Jansen, T., Keller, S. D., & Möller, J. (2018). The impact of vocabulary and spelling on judgments of ESL essays: An analysis of teacher comments. *The Language Learning Journal*, 1–17. https://doi.org/10.1080/09571736.2018.1522662

Xie, Q. (2015). "I must impress the raters!" An investigation of Chinese test-takers' strategies to manage rater impressions. *Assessing Writing*, *25*, 22–37. https://doi.org/10.1016/j.asw.2015.05.001

Yu, G. (2010). Lexical Diversity in Writing and Speaking Task Performances. *Applied Linguistics*, *31*(2), 236–259. https://doi.org/10.1093/applin/amp024

Appendix A: TOEFL iBT rubric used to score the independent writing task.

SCORE	TASK DESCRIPTION
5	**An essay at this level largely accomplishes all of the following:** Effectively addresses the topic and task Is well organized and well developed, using clearly appropriate explanations, exemplifications and/or details Displays unity, progression and coherence Displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice and idiomaticity, though it may have minor lexical or grammatical errors
4	**An essay at this level largely accomplishes all of the following:** Addresses the topic and task well, though some points may not be fully elaborated Is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications and/or details Displays unity, progression and coherence, though it may contain occasional redundancy, digression, or unclear connections Displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form or use of idiomatic language that do not interfere with meaning
3	**An essay at this level is marked by one or more of the following:** Addresses the topic and task using somewhat developed explanations, exemplifications and/or details Displays unity, progression and coherence, though connection of ideas may be occasionally obscured May demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning May display accurate but limited range of syntactic structures and vocabulary
2	**An essay at this level may reveal one or more of the following weaknesses:** Limited development in response to the topic and task Inadequate organization or connection of ideas Inappropriate or insufficient exemplifications, explanations or details to support or illustrate generalizations in response to the task A noticeably inappropriate choice of words or word forms An accumulation of errors in sentence structure and/or usage
1	**An essay at this level is seriously flawed by one or more of the following weaknesses:** Serious disorganization or underdevelopment Little or no detail, or irrelevant specifics, or questionable responsiveness to the task Serious and frequent errors in sentence structure or usage
0	**An essay at this level** merely copies words from the topic, rejects the topic, or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.

Appendix B1: Final HF Essay

*Essay prompt removed on request by ETS

When it comes the question of which one is more important between interest and job or career when we have to choose, different people might have different answers. From my point of view, I completely agree with picking subjects to prepare for a job or career is more important because job or career is priority and basis of my happy life, good job will give me financially stable and keep me be respected or admired. The reasons why I think so go as follow:

First, the reason for my view is that job or career is priority and basis of my happy life. First of all, I live in Toronto, everything need to pay by myself. I need pay hydro bill, telephone bill, rent etc. So I have to find a job. Second, In order to find a good job, I have to raise myself to the next level, so I should study english and go to university in future. If I have a good job, I can make more money to pay the high tuition, books fee etc. I can focus on study, enjoy free time with my friends. All these activity need a basis of money.

Second, another reason for my view is that good job will give me financially stable. For example, if I have a good job, I do not need loan and can freely pick the program that I really want to pick to learn such MBA program although the tuition is so high. When I have enough saving, I can make a plan including trip to the whole world and enjoy my life.

Finally, the good job will provide me respect and approval because in order to get good job, I have to get into university to get great education, learning professional skills, and I will keep myself proper, nice, etc. All these will provide me be respect and approval by other people.

In conclusion, according to the reasons I have discussed above, it is a right idea for me to pick subjects to prepare a job or career at first.

Appendix B2: Final LF Essay

*Essay prompt removed on request by ETS

When it comes the dilemma of which one is more important between interest and job or career when we have to choose, different people might have different answers. From my point of view, I entirely agree with selecting subjects to prepare for a job or career is more important because job or career is priority and foundation of my happy life, decent job will offer me financially stable and keep me be esteemed or admired. The reasons why I think so go as follow:

First, the reason for my view is that job or career is priority and foundation of my happy life. First of all, I reside in Toronto, everything need to pay by myself. I need pay hydro bill, telephone bill, rent etc. So I have to secure a job. Second, In order to secure a decent job, I have to advance myself to the next level, so I should study english and go to university in future. If I have a decent job, I can earn more money to pay the expensive tuition, books fee etc. I can concentrate on study, enjoy leisure time with my friends. All these pursuit need a foundation of money.

Second, another reason for my view is that decent job will offer me financially stable. For example, if I have a decent job, I do not need loan and can freely select the program that I really desire to select to learn such MBA program although the tuition is so expensive. When I have adequate saving, I can make a plan including travel to the whole world and enjoy my life.

Finally, the decent job will grant me esteem and praise because in order to get decent job, I have to get into university to get distinguished education, learning professional skills, and I will keep myself cultured, nice, etc. All these will grant me be esteem and praise by other people.

In summary, according to the reasons I have discussed above, it is a sensible idea for me to select subjects to prepare a job or career at first.

Appendix C: Feedback Point Coding Categories and Definitions

Coding Scheme	Explanations
Non-Vocabulary category	
Organization (Unity, Coherence, Progression, Paragraphs)	Anything to do with structure of the essay, or structure of a paragraph, including unity of entire essay (following through with thesis), coherence of ideas, use of transition words
Ideas (Use of examples to support position)	Examples, strength of argument, reasoning
Addresses Topic	Does the essay answer the question? Does the writer complete the task assigned?
Errors (including spelling, grammar, word form)	Errors in spelling, punctuation, grammar, including sentence structures, word forms, all **except Word Choice**
Syntactic variety (Variety of sentence structures used)	Variety of sentences used (complex sentences, compound sentences)
Global / Holistic impressions OR Vague	Remarking on the overall impression of the essay; how the rater 'feels' about the essay, including the style, tone, register, proficiency level of writer, OR making a positive or negative comment on an aspect of the essay that is hard to pinpoint
Other	Remarking on something about the essay that is outside the scope a rater's consideration
Vocabulary Category	
Vocabulary - LD	Lexical diversity: variety or diversity of different words used
Vocabulary - LF	Lexical frequency: Use of uncommon (or common) words. Could also be referred to as 'sophisticated' or advanced words
Vocabulary - Collocations and Idiomatic Expressions	Use of collocations, idiomatic expressions, and other common phrases that lend the text a 'native-like' quality
Vocabulary - Polysemy	Words that can occur in many contexts or have multiple meanings are Polysemous. Example: Make a cake, Make the bed, Make arrangements, Make up with someone. Words that occur in limited contexts are less Polysemous. Example, you can say Bake a cake, but not Bake food.
Vocabulary - Hypernymy	Words that are more Hypernymic are more specific: Example, Greyhound is more specific than Dog is more specific then Animal

Vocabulary - Word Range (special/specific terms)	Words that are used in more limited contexts. For example, Transaction is a word with a limited Range that you'd find use more often in Business contexts than others. Includes Academic Words.
Vocabulary - Word choice	Commenting on writer’s choice of words
Vocabulary - Vague	Commenting on vocabulary but nothing specific

Appendix D: Full breakdown of feedback points and numbers

Coding	LF Essay	HF Essay	Total (+) and (-)	Total from each category	Total Non-vocab and Vocab
Organization (+)	11	16	27	34	115
Organization (-)	4	3	7		
Ideas (+)	3	3	6	12	
Ideas (-)	4	2	6		
Addresses Topic (+)	5	3	8	17	
Addresses Topic (-)	6	3	9		
Errors (+)	1	4	5	32	
Errors (-)	14	13	27		
Syntactic Variety (+)	0	1	1	2	
Syntactic Variety (-)	0	1	1		
Global / Holistic Impressions OR Vague (+)	3	5	8	11	
Global / Holistic Impressions OR Vague (-)	2	1	3		
Other	1	6	7	7	
Vocabulary - LD (+)	1	1	2	4	18
Vocabulary - LD (-)	1	1	2		
Vocabulary - LF (+)	0	0	0	1	
Vocabulary - LF (-)	0	1	1		
Vocabulary - Collocations / Idiomatic Expressions (+)	1	0	1	2	
Vocabulary - Collocations / Idiomatic Expressions (-)	1	0	1		
Vocabulary - Polysemy (+)	0	0	0	0	
Vocabulary - Polysemy (-)	0	0	0		
Vocabulary - Hypernemy (+)	0	0	0	0	
Vocabulary - Hypernemy (-)	0	0	0		
Vocabulary - Word Range (+)	0	1	1	2	
Vocabulary - Word Range (-)	0	1	1		
Vocabulary - Word Choice (+)	0	0	0	8	
Vocabulary - Word Choice (-)	3	5	8		
Vocabulary - Vague (+)	1	0	1	1	
Vocabulary - Vague (-)	0	0	0		
Total	62	71	133	133	**133**

Appendix E: All Content Words used and replaced in LF and HF Essay words, their ranks/frequencies, and Word Sense

Word in High Frequency Essay	Rank in COCA	Synonym used in Low Frequency Essay	Rank in COCA	Sense^	NOTES
Question in "when it comes to the question. . ."	197 (Question)	Dilemma	4151 (Dilemma)	Quandary; trouble or question resulting from complexity	"Dilemma" is not a direct synonym of "question," but after the validation process it was agreed that "Dilemma" is an appropriate Low Frequency synonym with the same sense as "question" in the context, "when it comes to the question of which one is more important. . ."
Completely	1170 (Completely)	Entirely	1868 (Entirely)	To the complete degree or to the full or entire context	
Picking	517 (Pick)	Selecting	1742 (Select)	Select carefully from a group	Form alliteration after change; "picking subjects" becomes "selecting subjects"
Good	110 (good)	Decent	3908 (Decent)	Having desirable or positive qualities especially those suitable for a thing specified	All instances of "good job" changed to "decent job" to maintain Lexical Diversity
Basis	1311 (basis)	Foundation	2270 (foundation)	A relations that provides the foundation for something, the fundamental assumptions underlying an explanation	
Give	98 (Give)	Offer	373 (offer)	Cause to have in the abstract of physical sense, transfer poession of something concrete or abstract to	

				somebody, convey or reveal information	
Respect	1559 (respect)	Esteem	11563 (Esteem)	An attitude of admiration or esteem, the condition of being honoured (esteemed or respected or well regarded)	In the first paragraph, the incorrect form of "esteem" is used (adjective: esteemed) to match the incorrect form of "respect" used in the same place for the High Frequency essay
Live	210 (live)	Reside	5342 (Reside)	Make one's home or live in	
Get	39 (get)	Secure	2505 (secure)	Obtain; come into possession of something concrete or abstract	
Raise	433 (raise)	Advance	2699 (advance)	Improve; raise the level or amount of something, raise from a lower to a higher position	
Make	45 (make)	Earn	1309 (earn)	earn; make or cause to be or become, cause to do, give rise to	
Focus	688 (Focus)	Concentrate	2373 (Concentrate)	Focus one's attention on something, cause to converge on or towards a central point	
free in phrase "free time"	473 (free)	Leisure	6353 (Leisure)	Time available for ease and relaxation, freedom to choose pastime or enjoyable activity	"Leisure" is not a direct synonym of "free," but after the validation process it was agreed that "Leisure" is an appropriate Low Frequency synonym with the same sense as "free" in the context, ". . .enjoy leisure time with my friend . . ."

Activity	537 (Activity)	Pursuit	3903 (Pursuit)	hobby; an activity that diverts or amuses or stimulates	
Want	83 (want)	Desire	3812 (desire)	Feel or have desire for, have a need of, wish or demand the presence of	
Provide	262 (Provide)	Offer	373 (offer)	Provide what is desired or needed, esp. support, food, or sustenance	
Pick	517 (Pick)	Select	1742 (Select)	Select carefully from a group	
High in "tuition is so high"	141 (High)	Expensive	1670 (Expensive)	Greater than normal in degree or intensity or amount	"Expensive" is not a directy synonym for "high" but after the validation process it was agreed that "Expensive" is an appropriate Low Frequency synonym with the same sense as "high" in the context, ". . .the tuition is so expensive."
Trip	963 (trip)	Travel	1082 (travel)	Change location, undertake a journey or trip, take a trip for pleasure	"Travel" is not a direct synonym of "trip," but after the validation process it was agreed that "Travel" is an appropriate Low Frequency synonym with the same sense as "trip" in the context, "I can make a plan including travel to the whole world. . ."
Great	160 (great - adj)	Distinguished	5632 (distinguished)	Set apart from other such things, standing above others in attainment or reputations	
Proper	2069 (proper)	Cultured	15984 (cultured)	Polite; marked by refinement in taste and manners	"Cultured" is not a direct synonym of "proper," but after the validation process it was agreed that "Cultured" is an appropriate Low Frequency synonym with the same sense as "proper" in the context, ". . .the decent job will grant me esteem

					and praise. . .and I will keep myself cultured, nice etc."
Enough	872 (Enough - determiner)	Adequate	3318 (Adequate)	Enough to meet a purpose	"Enough savings" changed to "adequate savings"
Get	39 (Get)	Grant	1917 (Grant)	Let have, allow to have, bestow esp. officially	"Grant" is not a direct synonym of "get," but after the validation process it was agreed that "grant" is an appropriate Low Frequency synonym with the same sense as "get" in the context, ". . .the decent job will grant me esteem and praise. . ."
Approval	2565 (Approval)	Praise	4976 (Praise)	Appreciation; a message expressing a favourable opinion	
Conclusion in "in conclusion. . ."	1672 (Conclusion)	Summary	4965 (Summary)	Position or opinion or judgement reached after consideration, a brief statement that presents the main points in a concise form	"Summary" is not a direct synonym of "conclusion," but after the validation process it was agreed that "Summary" is an appropriate Low Frequency synonym with the same sense as "Conclusion" in the context, "In summary, according to the reasons I have discussed above. . ."
Right	317 (Right)	Sensible	6208 (Sensible)	Being of striking appropriateness and preference, suitable and fitting	"Sensible" is not a direct synonym of "right," but after the validation process it was agreed that "Sensible" is an appropriate Low Frequency synonym with the same sense as "Right" in the context, "it is a sensible idea for me to select. . ."
What	*	That	*		Spelling mistake: "What" changed to "That," to minimize affect of spelling mistakes on rater perception of overall writing, since we're trying to test the effects of Low/High Frequency words
of	*	or	*		Spelling mistake: "Of" changed to "Or," to minimize affect of spelling mistakes on rater

					perception of overall writing, since we're trying to test the effects of Low/High Frequency words
NOTES:					

1. Changes made from High Frequency essay, NOT original essay.
2. Low Frequency essay has ONE more unique words than the High Frequency essay. This word is "earn," which replaced "make" in the High Frequency essay. "Make" appears twice in the High Frequency essay and once in the Low Frequency essay. The repeated occurrence is in the phrase, "make a plan." I could not find an appropriate synonym for "make" that was as collocationally appropriate with "plan," therefore this was not changed to maintained the Type-Token ratio in both essays. However, in an essay of 347 words, I believe one repeated word will have no affect on raters in terms of Lexical Diversity.
3. "Rank in COCA" refers to raw frequency of word in entire COCA database. Example: "Completely (1170)" means the word "Completely" (NOT its base word "complete") is the 1170th most common word in the database. The lower the number the more common the word.
4.^ Senses taken from WordNet Database through wordandphrase.info. Not all possible senses of words were presented, only those most pertinent to the context

Appendix F: Ethics approval letter by Western University Non-medical research ethics board

Western Research

Date: 8 November 2018 **To**: Dr. Farahnaz Faez **Project ID**: 112918

Study Title: How Do Changes in Lexical Frequency affect written assessment scores when Lexical Diversity is held constant?

Application Type: NMREB Initial Application

Review Type: Delegated

Full Board Reporting Date: 07/Dec/2018 **Date Approval Issued**: 08/Nov/2018 15:51

REB Approval Expiry Date: 08/Nov/2019

Dear Dr. Farahnaz Faez

The Western University Non-Medical Research Ethics Board (NMREB) has reviewed and approved the WREM application form for the above mentioned study, as of the date noted above. NMREB approval for this study remains valid until the expiry date noted above, conditional to timely submission and acceptance of NMREB Continuing Ethics Review.

This research study is to be conducted by the investigator noted above. All other required institutional approvals must also be obtained prior to the conduct of the study.

DOCUMENTS APPROVED:

Document Name	Document Type	Document Date	Document Version
1 Email Script for Recruitment	Recruitment Materials	11/Oct/2018	1
3 Letter of Information and Consent	Implied Consent/Assent	01/Nov/2018	2
5 Debriefing Document	Debriefing document	11/Oct/2018	1
Sample essay that participants will score	Other Data Collection Instruments	11/Oct/2018	1

Documents Acknowledged:

Document Name	Document Type	Document Date	Document Version
toefl_writing_rubrics	Supplementary Tables/Figures	06/Oct/2018	1

No deviations from, or changes to the protocol should be initiated without prior written approval from the NMREB, except when necessary to eliminate immediate hazard(s) to study participants or when the change(s) involves only administrative or logistical aspects of the trial.

The Western University NMREB operates in compliance with the Tri-Council Policy Statement Ethical Conduct for Research Involving Humans (TCPS2), the Ontario Personal Health Information Protection Act (PHIPA, 2004), and the applicable laws and regulations of Ontario. Members of the NMREB who are named as Investigators in research studies do not participate in discussions related to, nor vote on such studies when they are presented to the REB. The NMREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000941.

Please do not hesitate to contact us

if you have any questions.

Sincerely,

Katelyn Harris, Research Ethics Officer on behalf of Dr. Randal Graham, NMREB Chair

Note: This correspondence includes an electronic signature (validation and approval via an online system that is compliant with all regulations).

CV: MUNEER HUDA

EDUCATION

SEPT 2017 - PRESENT

MA IN EDUCATION (APPLIED LINGUISTICS), UNIVERSITY OF WESTERN ONTARIO

APRIL 2014

TESOL DIPLOMA (250 HOURS), COVENTRY HOUSE INTERNATIONAL

FEBRUARY 2010

B.SC HONOURS IN ENVIRONMENTAL SCIENCES, UNIVERSITY OF GUELPH

EXPERIENCE

JAN 2018 – APRIL 2019 (3 SEMESTERS)

TEACHING ASSISTANT (BIOLOGY), UNIVERSITY OF WESTERN ONTARIO

As a teaching assistant I was responsible for independently delivering tutorial material to first year biology students. I lectured and guided students in completing their assignments, marked and gave feedback on assignments. In my end-of-semester review, one of my students commented that I was The GOAT (The Greatest Of All Time).

JULY 2016 – APRIL 2019

EAP ENGLISH TEACHER, CULTUREWORKS

As an EAP (English for Academic Purposes) teacher I prepared international students to attend post-secondary school in Canada. My primary responsibilities were to plan lessons, lecture, assist students in and outside the class room, mark assignments, and provide feedback. As my secondary responsibilities I occasionally contributed to assignment, assessment, and curriculum design. During my time at CultureWorks, I taught Grammar, Writing, Reading & Listening, and an Independent Study course.

AUG 2015 – AUG 2016

FREELANCE EDITOR AND ONLINE ENGLISH TUTOR, ONLINE

As a freelance editor I worked on a variety of projects, including resumes, job descriptions, and copywriting. As an English tutor, I worked for several online organizations to tutor students from various parts of the world. I catered lessons to my students' needs, which varied

considerably, from beginner to advanced, from primary school students to working professionals.

SEPT 2014 – JUNE 2015

ENGLISH TEACHER, HONGYA PRIMARY SCHOOL, SHENZHEN, CHINA

As an international English teacher, I prepared lessons, taught, and managed classrooms of 50 students each. My lessons focused on communicative teaching and immersing students in a Western English sociolinguistic culture.

JAN 2014 – MARCH 2014

SCIENCE TEACHER, MAD SCIENCE

As a "mad scientist," I taught science and conducted experiments with primary school students in an after-school extracurricular program. I introduced basic concepts in Chemistry and Physics to students with the aim of getting them interested early in STEM subjects. I was responsible for teaching, demonstrating experiments, and having students conduct the experiments in a safe and fun environment.

SKILLS AND TRAITS

- Excellent speaker and presenter
- Great interpersonal and communication skills
- Experienced freelance editor and writer
- Former VP of Education at university Toastmasters club