The Relationship Between Lexical Frequency Profiling Measures and Rater Judgements of Spoken and Written General English Language Proficiency on the CELPIP-General Test

Scott Roy Douglas

Independent confirmation that vocabulary in use unfolds across levels of performance as expected can contribute to a more complete understanding of validity in standardized English language tests. This study examined the relationship between Lexical Frequency Profiling (LFP) measures and rater judgements of test-takers' overall levels of performance in the Speaking and Writing modules of the CELPIP-General test. In particular, the potential of measures such as lexical stretch and number of frequency bands accessed was examined. Randomized quota sampling from previously rated test-taker responses resulted in 200 speaking samples and 200 writing samples being compiled to create corpora of 211,602 running words and 70,745 running words respectively. Pearson r was used to examine the relationships between the LFP measures and rater judgements of CELPIP levels. Results point to significant correlations, with increasing CELPIP levels of performance generally accompanied by test-takers' increasing ability to produce greater numbers of words, deploy a greater variety of words, rely less on high-frequency vocabulary, tap into mid-frequency vocabulary, and access a greater number of frequency bands. These results underline the contribution of independently obtained lexical measures toward a fuller understanding of concurrent validity in standardized English language proficiency testing.

La confirmation indépendante que le vocabulaire d'usage se répand sur plusieurs niveaux de performance tel que prévu peut contribuer à une meilleure interprétation de la validité des tests standardisés de langue anglaise. Cette étude a examiné le rapport entre les mesures de profilage de la fréquence lexicale et les évaluations de la performance globale des élèves aux modules de parole et de rédaction du Programme canadien d'évaluation du niveau de compétence linguistique en anglais (CELPIP). Plus précisément, on a examiné le potentiel des mesures telles l'étendue lexicale et le nombre de bandes de fréquences atteintes. L'échantillonnage par quota aléatoire de réponses d'élèves déjà évaluées a entrainé la formation de 200 échantillons de parole et 200 échantillons de rédaction représentant deux corpora, un de 211 602 mots liés et l'autre de 70 745 mots liés. On a employé le coefficient de corrélation de Pearson pour examiner les rapports entre les mesures de la fréquence lexicale et les évaluations en fonction des niveaux du CELPIP. Les résultats dévoilent des corrélations significatives entre, d'une part, les meilleures performances au CELPIP et, d'autre part, une capacité à produire une quantité

et une variété plus importantes de mots; à moins recourir aux mots les plus fréquents; à puiser dans du vocabulaire à fréquence moyenne; et à accéder à un plus grand nombre de bandes de fréquence. Ces résultats soulignent la contribution des mesures lexicales obtenues indépendamment à la compréhension de la validité concourante des évaluations standardisées des compétences linguistiques en anglais.

Canada is a major immigrant-receiving nation. For the ten-year period from 2003 to 2012, approximately 2.5 million new immigrants came to Canada. Of those individuals, almost 1.5 million were economic class immigrants (Citizenship and Immigration Canada, 2013a). In addition to new immigrants, in the same ten-year period Canada received on average 159,202 temporary foreign workers per year, with 491,547 temporary foreign workers still present in Canada in 2012 (Citizenship and Immigration Canada, 2013b). For many potential new economic class immigrants and temporary foreign workers, there is a requirement for proof of language skills in order to apply for permanent resident or temporary foreign worker status.

The stakes are high for applicants taking the standardized tests that are the accepted measures of English language proficiency. If scores are too low, prospective immigrants and foreign workers, who are required to show evidence of English language proficiency, risk having their applications rejected. Thus, in order to ensure a fair process, accepted measures of English language proficiency have to be both reliable and valid. An important part of overall English language proficiency is the role that vocabulary plays as an underlying variable to language ability. In general, the ability to deploy and understand a precise and varied range of vocabulary is related to improved language capabilities (Roessingh, 2006). Examining the vocabulary elicited by an English language proficiency test can provide important information related to the validity of that test. Lexical evidence related to validity can be gathered by independently calculating Lexical Frequency Profiling (LFP) measures of breadth of vocabulary output (Laufer & Nation, 1995) in written and spoken test-taker responses and the strength of the relationships those lexical measures have with assessment ratings of the test-taker responses.

Vocabulary and Concurrent Validity

The concept of validity is connected to how well a test measures what it is meant to measure, and a determination of validity can contribute to an appropriate and meaningful understanding of test results (Bachman & Palmer, 1996; Gay, Mills, & Airasian, 2012). A key aspect of a test's validity is that of concurrent validity, which is based on the relationship between the results of the test under investigation and another valid measure (Gay et al., 2012). Bachman and Palmer (1996) maintain that high-stakes tests, such as

the CELPIP-General test, require a wide range of evidence in order to support the validity of test score interpretations and decisions based on those interpretations. Concurrent validity explorations such as those undertaken in the present study can contribute to providing needed evidence to support interpretations based on standardized test scores. A proposed aspect of concurrent validity for English language proficiency testing is that connected to vocabulary and the relationship between independent measures of lexical performance elicited by a test instrument and the overall test scores of the instrument under investigation. Vocabulary in use is an important part of the standardized assessment of English language proficiency. Generally, it can be expected that more highly rated speaking and writing samples demonstrate greater control and deployment of the English language vocabulary appropriate for the task. O'Loughlin (2013) maintains that a standardized test that employs and elicits vocabulary representative of the vocabulary which test-takers can be expected to use and understand in real-world contexts can be understood as having lexical validity. For the purposes of this research study, the aspect of concurrent validity under investigation is the extent to which measures of vocabulary breadth of knowledge correlate, as determined by LFP, with the Canadian English Language Proficiency Index Program (CELPIP) General Test levels of performance.

Vocabulary as an Underlying Variable

Vocabulary has been identified as an underlying variable of English language proficiency, with more sophisticated lexical output and understanding being associated with overall improved additional language competencies (Roessingh, 2006). For example, in ratings of speaking performance, measures of lexical richness significantly and positively correlate with general English language proficiency (Yu, 2009). In addition to speaking performance, the skilled employment of vocabulary knowledge leads to improved generation, development, and presentation of ideas, particularly in written text (Engber, 1995; Grabe, 1984; McNamara, Crossley, & McCarthy, 2010; Raimes, 1983, 1985). Generally, the ability to deploy an increasing range of vocabulary accompanies improved writing skills (Smith, 2003). Without this ability to deploy an appropriate range of vocabulary, the conveyance of precise meaning can become lost (Spack, 1984). As a result, the amount of vocabulary available for use can be directly associated with quality of a written text (Brynildssen, 2000). Robust vocabulary usage appears to have a positive impact on readers (Laufer, 1994), with higher ratings of writing quality given to writers with more available vocabulary to use (Nation, 2001). It has also been shown that writing samples with low ratings are typically accompanied by simple vocabulary (Cobb, 2003; Hinkel, 2003), but highly rated writing samples correlate with measures of increasing lexical richness (Laufer & Nation, 1995). Roessingh (2008) also identified that general evaluations of writing quality

are affected by low vocabulary ratings. Roessingh (2008) analyzed the results of the Alberta English 30 Diploma examination, an examination worth 50% of students' final course mark for Grade 12 English Language Arts 30-1. When considering the subscores for the written response components of the examination, it was found that lower vocabulary subscores were associated with lower subscores for other measures, while higher vocabulary subscores were associated with higher subscores for other measures. The conclusion was that measures of lexical ability were associated with the overall ability to make and communicate meaning in the Alberta English 30 Diploma examination, with the inference that vocabulary is an underlying variable of English language proficiency.

Lexical Output in Standardized English Language Testing

If vocabulary is an underlying variable of English language proficiency, evidence of the relationship between vocabulary output and overall outcomes in standardized English language testing should be apparent. For example, Douglas (2010) found that there were moderate to strong correlations between independent measures of lexical breadth of knowledge and overall final assessments on a large-scale Canadian test of university entrance-level writing competence. Banerjee, Franceschina, and Smith (2007) also investigated the relationship between vocabulary richness and judgements of writing performance, specifically in the International English Language Testing System (IELTS) Academic Writing module. One measure of vocabulary richness considered was that of lexical output. Results for the lexical output analysis showed that the mean total number of words (tokens) and the mean total number of different words (types) increased with each IELTS band level. Test-takers with lower IELTS scores produced fewer words in general as well as fewer unique instances of words. Further analysis found moderate positive correlations between tokens and IELTS band levels and between types and IELTS band levels, suggesting a relationship between total lexical output and judgements of IELTS scores. Along with lexical output, there also appeared to be a relationship between lexical sophistication and judgements of IELTS band levels. For Banerjee et al. (2007), lexical sophistication was determined by the percentage of low-frequency words in a text as measured by LFP. Results determined that the percentage of low-frequency words in a text increased with increasing IELTS band levels and that the percentage of high-frequency words in a text decreased with decreasing IELTS band levels. However, there did appear to be a point at which the trend levelled off and other aspects of language proficiency became more important in determining the IELTS band score.

Similar patterns of lexical output and a decreasing reliance on high-frequency vocabulary in output associated with test scores representing higher levels of English language proficiency also appear in large-scale standardized

assessments of spoken English language proficiency. Read and Nation (2006) looked at the percentage of low-frequency words in test-taker performance on the Speaking module of the IELTS test. Similar to Banerjee et al.'s (2007) study, the more highly rated that test-takers were, the more they were able to produce a greater total output of words. In addition, more highly rated test-takers used a wider range of low-frequency words, with the output of lower-rated test-takers being made up of an increased percentage of high-frequency word choices. To illustrate, Read and Nation (2006) found that 81.8% of the spoken output of test-takers rated as Band 4 on the IELTS scale was from the 2,000 most frequent words in English. However, only 68.4% of the spoken output of test-takers rated as Band 8 on the IELTS scale came from these same most frequent words in English.

Lexical Frequency Profiling

The Douglas (2010), Banerjee et al. (2007), and the Read and Nation (2006) studies all employed LFP as a tool for determining test-takers' reliance on high-frequency vocabulary and their ability to deploy low-frequency vocabulary. LFP provides computerized lexical frequency information about the words in a text by comparing the text to vocabulary frequency lists. This produces information about the percentage of words in a text that come from varying frequency levels. For example, in Laufer and Nation (1995), LFP was used to reveal the percentage coverage of the 1,000 most frequent word families, the 1,000–2,000 most frequent word families, a list of frequent academic word families, and word families not found on those three lists. Laufer and Nation (1995) found LFP to be a stable measure of lexical sophistication in that it was able to distinguish different proficiency levels as well as produce reliable results for writers producing two different texts. It also proved to be a valid measure of lexical competence in that LFP results correlated with other measures of lexical proficiency. As a result, Laufer and Nation (1995) determined that LFP is a valuable research and diagnostic instrument. While Meara and Bell (2001) and Meara (2005) have raised some questions as to the sensitivity and reliability of LFP, especially for estimating productive vocabulary size, Laufer (2005) has pointed out that LFP is a tool for examining productive vocabulary in use rather than a means of calculating the size of a participant's vocabulary. Furthermore, Laufer countered Meara's (2005) approach that used artificial data to criticize LFP by underlining that LFP is meant to be used with actual language in use. For Laufer, LFP is both valid and reliable for measuring the vocabulary that English as an additional language (EAL) users choose to deploy in their productive output. Bolstering Laufer's assertions, LFP has been employed in a number of studies of vocabulary in use in productive texts (for example, Cobb & Horst, 1999; Douglas, 2013; Laufer & Paribakht, 1998; Lee & Muncie, 2006; Morris, 2003; Morris & Cobb, 2004; Muncie, 2002; Read & Nation, 2006).

The CELPIP-General Test

Developed in Canada by Paragon Testing Enterprises, a subsidiary of the University of British Columbia, the CELPIP-General Test is a computer-delivered standardized test of general English language proficiency that focuses on varieties of English in use in Canadian contexts. The test consists of four modules: Listening (40 minutes), Reading (60 minutes), Speaking (20 minutes), and Writing (60 minutes). The two modules focused on in the current study are the Speaking and Writing modules. For the Speaking module, test takers attempt eight speaking tasks such as giving advice, talking about a personal experience, describing a scene, making predictions, comparing and persuading, dealing with a difficult situation, expressing opinions, and describing an unusual situation. Test-takers are presented with a prompt, and they deliver their answers through a microphone on the test-takers' computers. For the Writing module, test-takers are presented with two writing tasks such as writing an e-mail or responding to an opinion survey. After reading each writing prompt, test-takers type their responses to each task directly into their computers (Paragon Testing Enterprises, 2015).

There are 12 CELPIP levels of performance ranging from a minimal level of proficiency to advanced levels of proficiency in workplace and community contexts. In evaluating test-takers' responses to each speaking or writing task, raters focus on meaning and evaluate language across four dimensions. Each module of the CELPIP-General Test is assigned an overall CELPIP level. Reported results are referenced to the Canadian Language Benchmarks (CLBs; Wu & Stone, 2013). The CLBs, a national standard for describing language proficiency in Canada, are a series of 12 benchmarks that describe English language ability on a scale from basic to advanced proficiency (Centre for Canadian Language Benchmarks, 2012). The CELPIP levels of performance and their CLB Equivalencies are reported in Table 1.

In the Speaking module, three raters are assigned to each test-taker's responses. Each rater evaluates half of the responses (4 tasks), with the result that 50% of the responses are double rated and 50% of the responses are single rated. In the Writing module, two raters are assigned to each test-taker's responses, with both of the writing tasks being double rated. Agreement between raters is defined as an exact adjacent rating on the CELPIP scale. In the 14-week period toward the end of 2014, percentage rater agreement on the Speaking module ranged between 0.75 and 0.81. In the same period, percentage rater agreement on the Writing module ranged between 0.76 and 0.84. In addition to a primary round of rating, a trusted pool of expert raters review a random selection of the primary ratings. Approximately 10% of responses to the speaking and writing tasks are benchmark rated, resulting in about 33% of test-takers having some of their tasks benchmark rated (J. Stone, personal communication, February 16, 2015).

Table 1
CELPIP-General Levels of Performance
(Paragon Testing Enterprises, 2015)

CELPIP Level	*CELPIP Descriptor*	*CLB Level*
12	Advanced proficiency in workplace and community contexts	12
11	Advanced proficiency in workplace and community contexts	11
10	Highly effective proficiency in workplace and community contexts	10
9	Effective proficiency in workplace and community contexts	9
8	Good proficiency in workplace and community contexts	8
7	Adequate proficiency in workplace and community contexts	7
6	Developing proficiency in workplace and community contexts	6
5	Acquiring proficiency in workplace and community contexts	5
4	Adequate proficiency for daily life activities	4
3	Some proficiency in limited contexts	3
M	Minimal proficiency or insufficient information to assess	0, 1, 2

One of the main motivations for taking the CELPIP-General Test is for Canadian immigration purposes, since the CELPIP-General Test is accepted by Citizenship and Immigration Canada (CIC) as evidence of English language proficiency (Citizenship and Immigration Canada, 2015). Namely, the CELPIP-General can be used as proof of English language proficiency for applicants applying to the Canadian Federal Skilled Workers Program, Canadian Experience Class, Canadian Federal Skilled Trades Program, some Provincial Nominees Programs, and Start-Up Visa Programs (Paragon Testing Enterprises, 2015).

Research Question

The analyses in the current study are grounded in corpus linguistic methodologies that explore the relationship between independent LFP measures and rater judgements on the CELPIP-General scale. As such, the overarching research question for this inquiry is as follows:

> What is the relationship between various Lexical Frequency Profiling measures and rater judgements of overall CELPIP-General levels of performance on the speaking and writing tasks of the CELPIP-General Test?

Method

Corpus Compilation

Upon receiving the appropriate behavioural research ethics board approvals from the researcher's institution, Paragon Testing Enterprises released pre-

viously rated CELPIP-General speaking and writing test samples. For this study, audio files for 1,786 speaking samples and text files for 802 writing samples were made available to the researcher. Rating information was included along with each sample; however, all data released to the researcher were anonymized with no personally identifying information included. Each sample consisted of the eight speaking tasks or the two writing tasks that test-takers were required to complete for the respective modules. In order to be included as a potential sample, test-takers had to have attempted all of the tasks in a module.

Out of the available data, two corpora of spoken and written test-taker-generated texts were created through random quota sampling to represent a balanced range of CELPIP levels of performance. From the pool of available data, this approach to sampling entailed randomly choosing 20 test-taker samples at each level of proficiency from CELPIP levels 3–12, so that each level would be equally represented. The figure of 20 samples was determined by the maximum number of samples available at the higher levels. Due to a lack of data, CELPIP levels 1 and 2 were not included for inclusion in the corpus. In order to analyse the spoken corpus, audio recordings were transcribed and digitized for computer analysis. In order to analyze the written corpus, all spelling errors were corrected so that the files could be read by the online LFP software. If spelling had not been corrected, lexical items would not have been recognized by the software. The result was two balanced corpora of 200 samples each, with 20 samples included at each CELPIP level of proficiency. These samples represented 11% of the available speaking samples and 25% of the available writing samples, resulting in a spoken corpus of 211,602 running words and a written corpus of 70,745 running words.

Lexical Analysis

For both the spoken and written corpora, lexical frequency profiles were generated for each individual speaking or writing sample using the BNC-COCA Web VP vocabulary profile tool available on www.lextutor.ca (Cobb, 2015). The BNC-COCA Web VP employs an integration of British National Corpus (BNC) lists with a set of lists based on the Corpus of Contemporary American (COCA) English (Cobb, 2015). The BNC-COCA Web VP analyzes vocabulary output in a text by comparing the text to 25 vocabulary bands, each made up of 1,000 word families of decreasing frequency as found in the combined BNC and COCA lists. The resulting profile quantifies the percentage coverage of a text by each of the 1,000 word family bands of decreasing frequency. The 25 bands of the BNC-COCA Web VP can be understood as units of vocabulary frequency. Each band consists of 1,000 word families. As the bands increase in number, the word families decrease in their frequency of occurrence in the BNC-COCA corpus. Band 1 contains the 1,000 highest-frequency word families in the BNC-COCA lists. These word families are relatively common in English. Band 25 contains the 1,000 lowest-frequency word families in the

BNC-COCA lists. These word families are relatively uncommon in English. The result is that LFP measures can be generated indicating the percentage coverage of a text by an individual or set of frequency bands. LFP measures can also be generated indicating the lowest-frequency band represented by a lexical item in a text. For the purposes of this article, when discussing frequency bands and groups of 1,000 word families, these are the same measure, as generated by the BNC-COCA Web VP.

For each speaking and writing sample, eight lexical measures were generated by the BNC-COCA Web VP. The first two measures generated were the number of tokens (total number of running words) and the number of types (total number of different words) in each text. Next, the percentage coverage of each text by High Frequency Vocabulary (HFV) was calculated. HFV was defined as the 2,000 most frequent word families (Bands 1 and 2). This definition was based on Horst's (2013) assertion of the impact that 2,000 high-frequency word families have on English language proficiency. Next, the percentage coverage of each text by Mid-Frequency Vocabulary (MFV) was calculated. Schmitt and Schmitt (2014) define MFV as the vocabulary between 3,000 high-frequency and 9,000 low-frequency word families. Based on Douglas's (2014) contention that "word families found beyond the 10,000 word band are low-frequency vocabulary" (p. 10), for the purposes of the current study MFV was defined to include Bands 3–10. The percentage coverage of text for Low Frequency Vocabulary (LFV) was next calculated based on the percentage coverage of text by Bands 11–25 (including off-list items).

The next two measures involved what Douglas (2010, 2013) has called Lexical Stretch. The first measure of Lexical Stretch was that of the lowest frequency band accessed by a test-taker in order to cover 98% of a text. To calculate this measure, the percentage coverage of each frequency band is cumulatively added until 98% coverage of the overall text is reached. The band in which 98% coverage of the text is achieved is then recorded. For example, a test-taker with a high score for overall writing proficiency might deploy a lexical item from Band 15, a relatively low frequency band, with Bands 1–15 representing 98% of the lexical output of this particular test-taker. On the other hand, a test-taker with a low score for their overall writing proficiency might only deploy lexical items from the higher frequency bands, with lexical choices from Bands 1–3 (relatively high frequency bands) covering 98% of the test-taker's lexical output. The second measure of Lexical Stretch was the lowest frequency band overall accessed by a test-taker in producing a speaking or writing sample. Based on the examples above, the test-taker with the high score might access up to Band 15 (a relatively low frequency band) for 98% of the lexical output, but also deploy a lexical item from Band 20. Thus, Band 20 would represent the lowest frequency band accessed overall for this test-taker. The test-taker with the low score might access Bands 1–3 to cover 98% of the lexical output, but there may also be a lexical choice deployed from

Band 5. Thus, Band 5 would be the lowest frequency band overall accessed by this test-taker.

Finally, a new measure was introduced in order to calculate the number of frequency bands accessed by a test-taker in producing a text. When looking at the lexical stretch of a text, the analysis can reveal gaps between proficiency bands accessed by the producer of a text. For example, the lowest frequency band accessed by a test-taker might be Band 8. However, Bands 5, 6, and 7 might not have been accessed. While the top band reached is Band 8, the number of bands actually accessed is five. By considering the total number of bands accessed by test-takers, a more complete picture of their Lexical Stretch emerges.

In generating the lexical frequency profiles for the speaking and writing samples many proper nouns are typically categorized as not being found on the BNC-COCA 1–25k lists. As such, proper nouns were categorized in the first 1,000 word family frequency band. Proper nouns, which are typically marked by capitalization and identify unique people, places, products, organizations, or events, were included in the 1k frequency band based on the assumption that this type of word is "usually transparent" (Horst, 2013, p. 176) and is easily acquired and understood by learners (Nation, 2006; Schmitt, 2008; Webb & Rodgers, 2009). In addition, in order to avoid false measures of a test-taker's ability to deploy accurate low frequency lexical choices, semantic errors were identified and also categorized in the 1k frequency band based on the assumption that the test-taker did not have a correct understanding of the word in question (Douglas, 2010). However, the total number of semantic errors was minimal, with less than five errors being recategorized across both corpora.

Correlational Analysis

The overall rater judgements of CELPIP levels for performance for each speaking or writing sample was obtained. First, scatter plots were generated as a preliminary investigation into the relationship between the LFP measures and overall rater judgements of test-takers' CELPIP levels for the speaking and writing samples. Next, LFP data were correlated with the module-specific CELPIP levels using the product moment correlation coefficient (Pearson *r*). In determining the effect size of the results, Ellis's (2009) guidelines for effect size thresholds in correlational research were followed, with $r = .10$ (small), $r = .30$ (medium), $r = 5.0$ (large), and $r = 70$ (very large). In order to further measure the strength of the relationship, *r* squared was also calculated. In correlational research, *r* squared can provide an additional representation of the effect size in that it expresses the magnitude of the relationship between two variables (Creswell, 2012). All results are reported sequentially, with results for the spoken corpus presented first, followed by the results for the written corpus. Statistical analyses were carried out with IBM SPSS Statistics 22.

Results

In general, meaningful relationships between the eight LFP measures and rater judgements of CELPIP levels of performance emerged from the data. The results inform a wider understanding of concurrent validity in standardized English language tests, such as the CELPIP-General, and underscore the usefulness of LFP measures, such as the lexical stretch and the number of bands accessed, for building a more complete picture of concurrent validity. The results also have a role in informing the teaching and learning of vocabulary and vocabulary's role as an important variable in overall general English language proficiency.

Spoken Corpus Correlational Analysis

The first set of correlational analyses were connected to the tokens (total number of words) and types (total number of different words) in the speaking samples and their relationship with rater judgements of CELPIP levels of performance. The first results point to a significant very strong positive correlation between tokens and CELPIP levels, $r(198) = .84$, $p < .001$, with $R^2 = .71$. As a result, 71% of the variance was explained by the number of tokens in a text. Increases in tokens were associated with increases CELPIP levels. There was also a significant very strong positive correlation between the number of types found in a speaking sample and CELPIP levels, $r(198) = .92$, $p < 0.001$, with $R^2 = .85$. As a result, 85% of the variance was explained by the number of types in a text. Increases in types were associated with increases in CELPIP levels.

The second set of analyses examined the relationship between the percentage coverage of the speaking sample texts by vocabulary frequency estimates of HFV, MFV, and LFV and rater judgements of CELPIP levels of performance. First, there was a significant moderate negative correlation between HFV and CELPIP levels, $r(198) = -.50$, $p < .001$, with $R^2 = .25$. As a result, 25% of the variance was explained by the percentage coverage of a text by HFV. Increases in percentage coverage by HFV were associated with decreases in CELPIP levels. The next analysis indicated that there was a significant moderate positive correlation between MFV and CELPIP levels, $r(198) = .48$, $p < .001$, with $R^2 = .23$. As a result, 23% of the variance was explained by the percentage coverage of a text by MFV. Increases in percentage coverage by MFV were associated with increases in CELPIP levels. Finally, there was also a significant moderate positive correlation between LFV and CELPIP levels, $r(198) = .38$, $p < .001$, with $R^2 = .14$. As a result, 14% of the variance was explained by the percentage coverage of a text by LFV. Increases in percentage coverage by LFV were associated with increases in CELPIP levels.

The next set of correlational analyses were carried out to examine the relationship between the lexical stretch—that is the lowest frequency band

(with lower frequency bands at the upper end of the scale of 25 bands, e.g., Band 1 represents the highest frequency word families and Band 25 represents the lowest frequency word families)—accessed to reach 98% coverage of a speaking sample and the lowest frequency band accessed overall, with rater judgements of CELPIP levels of performance. In examining the lowest frequency band to reach 98% coverage of a text and CELPIP levels, there was a significant moderate positive correlation, $r(198) = .40$, $p < .001$, with $R^2 = .16$. As a result, 16% of the variance was explained by the lowest frequency band reached in order to cover 98% of a text. Increases in the lowest frequency band reached to gain 98% coverage were associated with increases in CELPIP levels. Next, there was also a significant moderate positive correlation between the lowest frequency band reached overall and CELPIP levels, $r(198) = .60$, $p < .001$, with $R^2 = .36$. As a result, 36% of the variance was explained by the lowest frequency band reached overall in a text. Increases in the lowest frequency band reached overall were associated with increases in CELPIP levels.

Finally, the relationship between the number of frequency bands accessed by a test-taker in producing a speaking sample and rater judgements of CELPIP levels of performance was examined. There was a significant strong positive correlation between the number of bands accessed and CELPIP levels, $r(198) = .70$, $p < .001$, with $R^2 = .49$. As a result, 49% of the variance was explained by how many frequency bands were accessed in a text. Increases in the number of frequency bands accessed were associated with increases in CELPIP levels.

Written Corpus Correlational Analysis

The same sets of correlational analyses were carried out with the writing corpus data as with the speaking corpus data. The first set of correlational analyses were connected to the tokens (total number of words) and types (total number of different words) in the writing samples and their relationship with rater judgements of CELPIP levels of performance. The first results point to a significant moderate positive correlation between tokens and CELPIP levels, $r(198) = .54$, $p < .001$, with $R^2 = .30$. As a result, 30% of the variance was explained by the number of tokens in a text. Increases in tokens were associated with increases in CELPIP levels. There was also a significant strong positive correlation between the number of types found in a speaking sample and CELPIP levels, $r(198) = .77$, $p < 0.001$, with $R^2 = .59$. As a result, 59% of the variance was explained by the number of types in a text. Increases in types were associated with increases in CELPIP levels.

The second set of analyses examined the relationship between the percentage coverage of the writing samples texts by vocabulary frequency estimates of HFV, MFV, and LFV and rater judgements of CELPIP levels of performance. First, there was a significant strong negative correlation between HFV and CELPIP levels, $r(198) = -.73$, $p < .001$, with $R^2 = .54$. As

a result, 54% of the variance was explained by the percentage coverage of a text by HFV. Increases in percentage coverage by HFV were associated with decreases in CELPIP levels. The next analysis indicated that there was a significant strong positive correlation between MFV and CELPIP levels, $r(198) = .73$, $p < .001$, with $R^2 = .53$. As a result, 53% of the variance was explained by the percentage coverage of a text by MFV. Increases in percentage coverage by MFV were associated with increases in CELPIP levels. Finally, there was a significant weak positive correlation between LFV and CELPIP levels, $r(198) = .28$, $p < .001$, with $R^2 = .08$. As a result, 8% of the variance was explained by the percentage coverage of a text by LFV. Increases in percentage coverage by LFV were associated with increases in CELPIP levels.

The next set of correlational analyses were carried out to examine the relationship between the lexical stretch—that is, the lowest frequency band accessed to reach 98% coverage of a writing sample—and the lowest frequency band accessed overall, with rater judgements of CELPIP levels of performance. In examining the lowest frequency band to reach 98% coverage of a text and CELPIP levels, there was a significant moderate positive correlation, $r(198) = .57$, $p < .001$, with $R^2 = .32$. As a result, 32% of the variance was explained by the lowest frequency band reached in order to cover 98% of a text. Increases in the lowest frequency band reached to gain 98% coverage were associated with increases in CELPIP levels. Next, there was also a significant moderate positive correlation between the lowest frequency band reached overall and CELPIP levels, $r(198) = .45$, $p < .001$, with $R^2 = .20$. As a result, 20% of the variance was explained by the lowest frequency band reached overall in a text. Increases in the lowest frequency band reached overall were associated with increases in CELPIP levels.

Finally, the relationship between the number of frequency bands accessed by a test-taker in producing a writing sample and rater judgements of CELPIP levels of performance was examined. There was a significant moderate positive correlation between the number of bands accessed and CELPIP levels, $r(198) = .58$, $p < .001$, with $R^2 = .34$. As a result, 34% of the variance was explained by how many frequency bands were accessed in a text. Increases in the number of frequency bands accessed were associated with increases in CELPIP levels.

Spoken and Written Corpus Comparisons

LFP data reveal interesting differences between the spoken and written corpora. In the spoken corpus, the strongest correlations were found for tokens, types, LFV, lowest frequency band accessed, and the number of bands accessed. In the written corpus, the strongest correlations were found for HFV, MFV, and the lowest frequency band accessed in covering 98% of a text. These differences are summarized in Table 2.

Table 2
Comparison of Pearson *r* Correlations Between
Lexical Profiling Measures and CELPIP Levels

Lexical measure	Spoken corpus (n = 200)	Written corpus (n = 200)
Tokens	.84*	.54*
Types	.92*	.77*
HFV (K1+K2)	-.50*	-.73*
MFV (K3–K10)	.48*	.73*
LFV (K11–K25 + off-list)	.38*	.28*
98% band coverage	.40*	.57*
Top band accessed	.60*	.45*
# of bands accessed	.70*	.58*

*$p < .001$

Discussion

This study set out to explore the relationship between eight productive vocabulary measures obtained with the BNC-COCA 1–25k LFP tool and rater judgements of overall CELPIP-General levels of performance on the Speaking and Writing modules of the CELPIP-General Test. The goal was to examine if LFP measures might be used to contribute validity evidence for large-scale standardized English language testing. Generally, there appears to be a meaningful relationship between the eight LFP measures explored in this study and rater judgements of CELPIP levels of performance. These relationships contribute to a broader understanding of concurrent validity in the CELPIP-General test, and point to the value of using LFP measures, including measures such as lexical stretch and the number of bands accessed, to gather evidence contributing to a wider picture of concurrent validity in standardized English language testing.

Tokens and Types

The first set of measures generated through LFP were tokens and types. In the spoken corpus, these two measures had very strong positive correlations with rater judgements of CELPIP levels of performance, and they represented the strongest relationships overall in the speaking corpus. It appeared that the ability to speak at length, thus producing a larger number of words, and the ability to use a wide variety of words were salient features of advancing spoken English language proficiency as judged by the CELPIP raters. These results recall findings by Read and Nation (2006), who found in the IELTS Speaking module that, in terms of overall type and token production, test-takers at lower IELTS band scores produced less vocabulary on average than test-takers at higher IELTS band scores. For Read and Nation (2006), increas-

ing IELTS band scores were accompanied by an increasing ability to tap into the lexical resources required to speak at greater length. Higher ratings of judged English language proficiency accompanying an increased ability to produce greater overall output also held true for the Speaking module of the CELPIP-General test.

In the written corpus, while they accounted for less of the variance in rater judgements of CELPIP levels compared to the spoken corpus, relationships between tokens and CELPIP levels were moderate, and between types and CELPIP levels were strong. This echoed the findings of Banerjee et al. (2007) in which Writing module IELTS test-takers at lower IELTS band scores produced less tokens and types than test-takers judged at higher IELTS band scores.

It is interesting to note how the correlations between the number of words and the number of different words were stronger in the spoken corpus than the written corpus. It appears that the ability to speak at length was more salient for raters than the ability to write at length. What's more, the ability to deploy a wider variety of words also seemed to be more salient for raters in the spoken corpus compared to the written corpus. In the written corpus, length did not have as strong a relationship with overall rater judgements as it did in the spoken corpus. In Douglas (2010), the findings also pointed to longer texts not necessarily being an indicator of improved writing quality. Shorter writing texts seemed to be able to deploy a greater lexical precision to convey meaning, whereas longer texts seemed to be made up of a more general vocabulary that required a greater number of higher frequency lexical choices to convey meaning, possibly because of strategies such as circumlocution to make up for missing lexical items.

HFV, MFV, LFV

The increasing ability of test-takers to rely less on high-frequency vocabulary and deploy larger amounts of low-frequency vocabulary also seemed to have a salient relationship with rater judgements of English language proficiency in the Speaking module of the CELPIP-General test. In the speaking corpus, the relationships between HFV, MFV, and LFV were represented by moderate correlations with rater judgements of CELPIP levels. As judgements of English language proficiency advanced, test-takers had less need to rely on high-frequency vocabulary for their spoken output. In addition, as judgements of English language proficiency advanced, test-takers deployed more instances of MFV and LFV. Test-takers who were consistently able to access lower-frequency vocabulary in their spoken output were judged to have increasing English language proficiency. Similar patterns of less reliance on high-frequency vocabulary for more highly rated speakers were found in test-taker output on the IELTS Speaking module by Read and Nation (2006). High-frequency vocabulary accounted for a greater percentage of speaker output for lower IELTS bands than higher IELTS bands. The reverse was also

true with low-frequency vocabulary, which accounted for less output in the lower IELTS bands than the higher IELTS bands.

Similar lexical patterns were found in the CELPIP-General written corpus as well. As rater judgements of test-takers' writing abilities increased, test-takers' reliance on high-frequency vocabulary decreased, with a strong negative correlation between the percentage coverage of text by HFV and rater judgements of CELPIP levels. There was also a strong correlation between the percentage coverage of text by MFV and rater judgements of CELPIP levels. More highly rated test-takers appeared to have greater facility for deploying vocabulary from the BNC-COCA mid-frequency bands. These findings are similar to Banerjee et al.'s (2007) results that demonstrated IELTS Writing module test-takers' IELTS band scores increased as their reliance on high-frequency words decreased.

In comparing the two corpora, a decreasing reliance on HFV and an increasing ability to deploy MFV had a stronger relationship with rater judgements of language ability in the writing corpus. These strong correlations point to the importance of having an adequate lexical command through the MFV bands for writers. Having a good command of MFV is a necessary component of English language proficiency (Schmitt & Schmitt, 2014), particularly so for the general writing tasks represented on the CELPIP-General test. Higher quality writing appeared to be accompanied by an increased use of MFV, a factor that was not as important in the speaking corpus.

Lexical Stretch and the Number of Bands Accessed

Information regarding the lexical stretch of test-takers related to the lowest frequency band accessed to cover 98% of a text, the lowest frequency band accessed overall, and the number of frequency bands accessed appear to have great potential for contributing independently obtained lexical information related to the concurrent validity of a standardized English language proficiency test. For both the spoken and written corpora, moderate to strong relationships were found for all of the lexical stretch measures, with the numbers of bands accessed being the strongest relationship in both corpora. Thus, it seems that fuller lexical frequency profiles in which test-takers access a greater number of frequency bands overall are related to increased rater judgements of language skills more than the ability to access low-frequency vocabulary that is isolated from the bulk of the lexical frequency choices employed by a test-taker. Simply using a low-frequency vocabulary choice in isolation is not enough to delineate higher levels of English language proficiency. Rather, a fuller and more balanced access through the frequency bands goes hand in hand with increasing rater judgements of general English language proficiency.

On balance, relationships for the lowest frequency band accessed overall and the number of frequency bands accessed in total with rater judgements of CELPIP levels of performance were stronger in the spoken corpus. Lower-

frequency lexical choices often equate with a more precise lexical choice, and being able to access a wider variety of frequency bands may be indicative of being able to deploy a wider variety of lexical choices in general. Being able to tap into a lower-frequency band as well as more of the bands in total appeared to be more salient in the spoken corpus.

Implications for Teaching and Learning

There are a number of implications for English language teaching and learning based on the results of this study. First of all, the results underline the importance of bringing vocabulary instruction solidly into classes that focus on the productive skills of speaking and writing. As an underlying variable of overall general English language proficiency, the ability not only to understand vocabulary encounters but also to deploy an apt and varied vocabulary in spoken and written English is central to overall language ability. Thus, sound strategies and techniques for learning new vocabulary (Nation, 2008) can be of great benefit in both speaking and writing instruction. In addition to strategies and techniques for teaching and learning vocabulary, the results start to point the way to lexical goals for learners. Building on work examining vocabulary thresholds for effective academic writing (Douglas, 2013), the current results point to the importance of using vocabulary that lies beyond high-frequency word choices for general writing tasks. A focus on MFV can be brought into the classroom, using the frequency principle (i.e., vocabulary being taught in order of frequency from high to low) as a means of organizing vocabulary instruction. Students and teachers can invest more time and effort into learning higher-frequency vocabulary and then mid-frequency vocabulary, in that order. The results further point to the lack of necessity for mastering low-frequency vocabulary items for general English language proficiency purposes. Excessive time in the classroom need not be devoted to LFV instruction. In the face of the time constraints typical of classroom instruction, resources can be focused on the HFV and MFV that support improved general English language proficiency.

The findings also point to a number of implications for teachers preparing students to take standardized English language proficiency tests such as the CELPIP-General test. For the Speaking module, the stronger correlations between tokens and types and rater judgements in the speaking corpus compared to the writing corpus demonstrate the importance of speaking at length and being able to use a variety of word choices as an important quality of spoken output. Students would be advised to make full use of the time allotted for each of the speaking tasks, and to vary their vocabulary. The token count might be related to fluency, in that more fluent speakers are able to utter a greater number of words in a given time period. Thus, automatic recall and deployment of lexical items within a time limit are important factors for improved judgements of overall spoken English language proficiency. The strong relationship between types and rater judgements of CELPIP levels

of performance might also be related to the coherence achieved in a text by varying lexical output through hypernyms, hyponyms, synonyms, and judicious pronoun usage. Students can be encouraged to make a conscious effort to avoid overuse of the same lexical choices and to use appropriate alternatives. Stronger relationships between the lowest frequency band accessed and the total number of frequency bands accessed were also found in the spoken corpus. While the ability to tap into a lower-frequency band points to an ability to deploy an apt word choice, the ability to access a higher number of frequency bands overall supports the conclusions drawn in relation to the strong correlation between types and rater judgements of CELPIP proficiency levels. Making use of a wide variety of lexical choices has a salient relationship with overall spoken general English language proficiency.

In preparing students for the Writing module of the CELPIP-General test, the strong negative correlation between the percentage coverage of text by HFV and CELPIP levels and the strong positive correlation between the percentage coverage of a text by MFV and CELPIP levels demonstrate the importance of encouraging students to rely less on high-frequency word choices and to strive to use word choices coming from the mid-frequency bands. In writing for general purposes, conveying meaning with a solid set of vocabulary includes using mid-frequency word families.

Limitations and Future Studies

The focus of the current study was on LFP measures generated from CELPIP-General test-takers' speaking and writing samples. These measures provide valuable information related to the breadth of test-takers' lexical output, in other words how many words test-takers deploy. However, the depth of knowledge (i.e. quality of word use) associated with that output is not assessed. Although what vocabulary choices test-takers are deploying can be uncovered, how well test-takers use those vocabulary choices is not examined in the current study. For example, the current study does not examine test-takers' use of vocabulary in terms of meaning, appropriacy, derivation, form (including spelling), omission, or style. A future study exploring the relationship between rater judgements of CELPIP levels of performance and test takers' lexical depth of knowledge, in particular in areas such as form and meaning, could contribute to a fuller understanding of the role independent measures of vocabulary knowledge play in the concurrent validity of standardized English language tests.

The CELPIP-General test is a test of general English language proficiency in the Canadian context using the varieties of English typically used in Canada (Paragon Testing Enterprises, 2015). The LFP tool employed for the current study, the BNC-COCA 1–25k (Cobb, 2015) consists of frequency bands representing North American and UK varieties of English (Cobb, 2015). However, there does not appear to be specific reference to Canadian varieties of English. This raises the question of whether Canadian usages found in the

corpus, such as *toque* or *washroom*, might skew the lexical frequency profiles generated from the test-taker responses. Although this question can be raised, it appears that the overall patterns emerging from the data are not overly affected by Canadian word choice. However, the building of a Canadian version of the BNC-COCA 1–25k tool would be a worthwhile future endeavour.

In the current study, each speaking or writing sample consisted of all of the tasks associated with their respective module. As a result, the speaking samples consisted of eight tasks pooled together and the writing samples consisted of two tasks pooled together. In addition, different forms of the CELPIP-General test exist, but all of the forms were considered together in the corpus. Future studies could disaggregate the tasks and test forms to investigate whether the patterns uncovered in this study hold true for specific tasks and separate test forms.

Another line of future research could extend the current correlational analysis by employing a multiple regression analysis to explore the predictive strength of the combined LFP measures in order to examine if the LFP measures significantly predict rater judgements of CELPIP levels of performance in the Speaking and Writing modules.

Finally, the current study points the way to explorations of the lexical aspects of concurrent validity in other standardized tests of English language proficiency. Examining independently obtained lexical evidence in tests such as the Michigan English Language Assessment Battery (MELAB), the Pearson Tests of English, the TOEFL iBT, and the Cambridge English examinations, amongst others, will contribute to establishing a well-rounded understanding of concurrent validity in these standardized measures of English language proficiency.

Conclusion

If vocabulary is an underlying variable of overall English language proficiency, evidence related to LFP measures can contribute to a more complete understanding of a standardized English language test's concurrent validity. In the Speaking and Writing modules of the CELPIP-General test, there seemed to be a meaningful relationship between independently calculated LFP measures and rater judgements of test-takers' overall CELPIP levels of performance. However, some differences did emerge between the two corpora. It appeared that overall text length and the number of different words in a text have stronger relationships with rater judgements of spoken output compared to written output. In addition, a decreasing reliance on HFV and an increasing ability to deploy MFV had stronger relationships with rater judgements of written output compared to spoken output. Overall, however, it can be generally stated that increasing rater judgements of CELPIP levels of performance in both corpora were accompanied by test-takers demonstrating an increasing ability to produce greater numbers of words, deploy a greater

variety of words, rely less on high-frequency vocabulary, tap into lower-frequency word choices, and access a greater number of frequency bands, all of which can be understood as important pieces of lexical evidence contributing to the concurrent validity of a standardized test of general English language proficiency.

Acknowledgements

The author would like to express his appreciation to the anonymous reviewers who offered valuable feedback on previous drafts of this article. The author would also like to thank the editor, the guest editor, and everyone else involved for helping bring this work to publication. Funding for this project was provided through a grant from Paragon Testing Enterprises.

The Author

Scott Roy Douglas is an assistant professor in the Faculty of Education on the University of British Columbia's Okanagan campus. His research includes various aspects of English as an additional language teaching and learning, with a particular focus on assessment, vocabulary, and pathways to higher education.

References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Banerjee, J., Franceschina, F., & Smith, A. M. (2007). Documenting features of written language production typical at different IELTS band score levels. *IELTS Research Reports, 7*. British Council/IELTS Australia.

Brynildssen, S. (2000). *Vocabulary's influence on successful writing*: ERIC Digest D157 (ERIC Document Service No. ED446339). Bloomington, IN: ERIC Clearinghouse on Reading, English, and Communication. Retrieved from http://www.eric.ed.gov/PDFS/ED446339.pdf

Centre for Canadian Language Benchmarks. (2012). *Canadian language benchmarks for adults: English as a second language for adults*. Citizenship and Immigration Canada. Retrieved from http://www.cic.gc.ca/english/pdf/pub/language-benchmarks.pdf

Citizenship and Immigration Canada. (2013a). *Facts and figures 2012—Immigration overview: Permanent and temporary residents: Permanent residents*. Retrieved from http://www.cic.gc.ca/english/resources/statistics/facts2014/permanent/index.asp

Citizenship and Immigration Canada. (2013b). *Facts and figures 2012—Immigration overview: Permanent and temporary residents: Temporary residents*. Retrieved from http://www.cic.gc.ca/english/resources/statistics/facts2013/temporary/index.asp

Citizenship and Immigration Canada. (2015). *Designated language testing agencies*. Retrieved from http://www.cic.gc.ca/english/resources/tools/language/agencies.asp

Cobb, T. (2003). Analyzing late interlanguage with learner corpora: Quebec replications of three European studies. *Canadian Modern Language Review 59*(3), 393–423.

Cobb, T. (2015). VocabProfile Home. *Compleat Lexical Tutor.* Retrieved from http://www.lextutor.ca/vp

Cobb, T., & Horst, M. (1999). Vocabulary sizes of some City University students. *Journal of the Division of Language Studies of City University of Hong Kong, 1*, 59–68. Retrieved from http://www.lextutor.ca/cv/CitySize.html

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson.

Douglas, S. R. (2010). *Non-native English speaking students at university: Lexical richness and academic success* (Unpublished doctoral dissertation). University of Calgary, Calgary, Canada. Retrieved from http://prism.ucalgary.ca/bitstream/1880/48195/1/2010_Douglas.pdf

Douglas, S. R. (2013). The lexical breadth of undergraduate novice level writing competency. *Canadian Journal of Applied Linguistics, 16*(1), 152–170. Retrieved from http://journals.hil.unb.ca/index.php/CJAL/article/view/21176/24432

Douglas, S. R. (2014). *Academic inquiry: Writing for post-secondary success*. Don Mills, ON: Oxford University Press.

Ellis, P. D. (2009). *Thresholds for interpreting effect sizes*. Retrieved from http://www.polyu.edu.hk/mm/effectsizefaqs/thresholds_for_interpreting_effect_sizes2.html

Engber, C. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing 4*(2), 139–155. Retrieved from http://www.sciencedirect.com/science/article/pii/1060374395900047

Gay, L. R., Mills, G. E., & Airasian, P. (2012). *Educational research: Competencies for analysis and applications* (10th ed.). Boston, MA: Pearson.

Grabe, W. (1984). Written discourse analysis. In R. B. Kaplan, A. d'Anglejan, J. R. Cowan, B. Kachru, G. R. Tucker, & H. Widdowson (Eds.), *Annual review of applied linguistics* (Vol. 5, pp. 101–123). New York, NY: Cambridge University Press.

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly, 37*(2), 275–301.

Horst, M. (2013). Mainstreaming second language vocabulary acquisition. *Canadian Journal of Applied Linguistics, 16*(1), 171–188.

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal, 25*, 21–33. Retrieved from http://rel.sagepub.com/content/25/2/21.full.pdf+html

Laufer, B. (2005). Lexical frequency profiles: From Monte Carlo to the real world. A response to Meara (2005). *Applied Linguistics, 26*(4), 582–588.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307–322.

Laufer, B., & Paribakht, T.S. (1998). The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning, 48*(3), 365–391.

Lee, S., & Muncie, J. (2006). From receptive to productive: Improving ESL learners' use of vocabulary in a postreading composition task. *TESOL Quarterly, 40*(2), 295–320.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57–86.

Meara, P. (2005). Lexical frequency profiles: A Monte Carlo analysis. *Applied Linguistics, 26*(1), 32–47.

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect: An Australian Journal of TESOL, 16*(3), 5–19.

Morris, L. (2003). Linguistic knowledge, metalinguistic knowledge and academic success in a language teacher education programme. *Language Awareness, 12*(2), 109–123.

Morris, L., & Cobb, T. (2004). Vocabulary profiles as predictors of the academic performance of teaching English as a second language trainees. *System, 32*(1), 75–87.

Muncie, J. (2002). Process writing and vocabulary development: Comparing lexical frequency profiles across drafts. *System, 30*, 225–235.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge, UK: Cambridge University Press.

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*(1), 59–82.

Nation, P. (2008). *Teaching vocabulary: Strategies and techniques*. Boston, MA: Heinle Cengage Learning.

O'Loughlin, K. (2013). *Research summary: Investigating lexical validity in the Pearson Test of Academic English*. Retrieved from http://pearsonpte.com/wp-content/uploads/2014/07/OLoughlin_K_2014.pdf

Paragon Testing Enterprises. (2015). *CELPIP-General Test*. Retrieved from http://celpiptest.ca

Raimes, A. (1983). Tradition and revolution in ESL teaching. *TESOL Quarterly, 17*(4), 535–552.

Raimes, A. (1985). What unskilled ESL students do as they write: A classroom study of composing. *TESOL Quarterly, 19*(2), 229–258.
Read, J., & Nation, P. (2006). An investigation of the lexical dimension of the IELTS speaking test. *IELTS Research Reports, 6*. The British Council/IELTS Australia.
Roessingh, H. (2006). BICS-CALP: An introduction for some, a review for others. *TESL Canada Journal, 23*(2), 91–96. Retrieved from http://www.teslcanadajournal.ca/index.php/tesl/article/viewFile/57/57
Roessingh, H. (2008). Variability in ESL outcomes: The influence of age on arrival and length of residence on achievement in high school. *TESL Canada Journal, 26*(1), 87–107.
Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research, 12*(3), 329–363.
Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47*(4), 484–503.
Smith, C. (2003). *Vocabulary's influence on successful writing: ERIC topical bibliography and commentary* (ERIC Document Reproduction Service No. ED480633). Bloomington, IN: ERIC Clearninghouse on Reading, English, and Communication. Retrieved from http://www.eric.ed.gov/PDFS/ED480633.pdf
Spack, R. (1984). Invention strategies and the ESL college composition students. *TESOL Quarterly, 18*(4), 649–670. Retreived from http://www.jstor.org/stable/3586581
Webb, S., & Rodgers, M. (2009). Vocabulary demands of television programs. *Language Learning, 59*(2), 335–366.
Wu, A., & Stone, J. (2013, June). *A standard setting design for the CELPIP-G speaking test.* Language Assessment Validation: Diversity in Method and in Stakeholder Perspectives Symposium. Canadian Association of Language Assessment. Paper presented at the annual conference of the Canadian Society for the Study of Education. University of Victoria, Victoria, British Columbia.
Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics, 31*(2), 236–259.