Examining Cloze Tests as a Measure of Linguistic Complexity in L2 Writing

Eun Seon Chung[1] & Soojin Ahn[2†]
[1]Hankuk University of Foreign Studies, [2]University of Seoul

ABSTRACT

Given that cloze tests are strongly associated with learners' writing proficiency, the present study examines the relationship between cloze test scores and specific linguistic features in second language (L2) writing. We investigate whether cloze tests can adequately and reliably measure linguistic features of syntactic and lexical complexity in L2 writing. Furthermore, the present study investigates the effect of the scoring method and L2 proficiency level on the relationship between cloze test scores and linguistic complexity features. Analysis of 60 students' writing compositions found the syntactic and lexical complexity features in length-related measures, complex nominals, and lexical variation to be significantly correlated with their cloze test scores. There was no significant difference in the effects of the two scoring methods. Also, cloze tests were most informative for high-level learners and least informative for intermediate-level learners. Implications for the use of cloze tests as a diagnostic measure in L2 writing classrooms are provided.

Keywords: cloze test, syntactic complexity, lexical complexity, scoring method, L2 writing proficiency

1. Introduction

Cloze tests, originally developed by Taylor (1953) to measure the readability of texts for native speakers, require the examinee to fill in the blanks in a text in which some words are deleted at regular intervals. The cloze test is considered to be an integrative test because examinees must draw on their overall language knowledge and reconstruct the passage by observing relationships between sentences or within a sentence and filling in the blanks with appropriate answers. Numerous studies have found it to be a valid and reliable method to measure second language (L2) learners' global language proficiency (Alderson, 1979a; Brown, 1983; Eckes &

* We deeply appreciate the helpful suggestions and critiques of three anonymous reviewers.
† Corresponding author: soojina2013@uos.ac.kr

Grotjahn, 2006; Fotos, 1991; Lee, 1997), and it is widely used by practitioners and researchers alike as a measure of general language proficiency for L2 learners. Significant correlations have been reported between cloze test scores and various linguistic skills such as reading (Bachman, 1985; Gellert & Elbro, 2013), writing (Fotos, 1991; Hanania & Shikhani, 1986; Lee 1997), vocabulary (Harsch & Hartig, 2016; Ryoo, 2017), and grammatical competence (Markham, 1987; Shanahan et al., 1982). Moreover, cloze tests were found to significantly correlate with standardized tests such as the TOEFL and placement/entrance examinations of various institutions (Bachman, 1985; Brown, 1983; Fotos, 1991; Hanania & Shikhani, 1986; Lee, 1997; Stubbs & Tucker, 1974). With such body of evidence that points to the cloze test as a valid measure of linguistic skills and global proficiency, more and more researchers are adopting this tool for language assessment because it is quick, cost-effective, easy to administer and evaluate, and most of all, reliable. The cloze test is also a practical and useful tool for classroom teachers who teach a number of students at the same time.

Among the many linguistic skills that can be measured by cloze tests, previous research shows that cloze scores highly correlate with writing scores and can potentially be used as a substitute for writing compositions (Chapelle & Abraham, 1990; Fotos, 1991; Hanania & Shikhani,1986; Lee, 1997; Stubbs & Tucker, 1974). Fotos (1991) found essays to have a higher commonality with the cloze test than the TOEFL, which suggests that the cloze test can measure language skills such as discourse and cohesion that the TOEFL cannot measure. Likewise, Lee (1997) found a stronger relationship between the cloze test and essays than the Korean college entrance English exam and recommended cloze tests to be used as a measure of writing proficiency. Although it is clearly established that cloze tests are highly correlated with essay scores, most previous studies have examined this correlation using holistic writing scores and have rarely examined the relationship between cloze test scores and specific linguistic features in L2 writing. The purpose of this study is thus to investigate whether or not cloze tests can adequately and reliably measure linguistic features of syntactic and lexical complexity in L2 writing, and if so which features it can predict. Furthermore, two factors, scoring methods and L2 proficiency levels, were examined to investigate their effects on cloze test scores. By doing so, we aim to reveal what aspects of L2 learners' writing can be measured by cloze tests and what factors play a role in this relationship. The rest of the paper is organized as follows: First, the previous studies on linguistic complexity measures and the cloze test are reviewed, and specific research questions

are stated. Next, we describe the research method and explain how statistical data analysis was conducted. The results of the statistical data analysis are then presented and discussed in relation to the specific research questions. In the last part of the paper, we discuss and evaluate the use of cloze tests as a diagnostic measure in L2 writing classrooms.

2. Literature Review

2.1. Linguistic complexity

When assessing L2 writing proficiency and development, various measures in syntactic and lexical complexity have been found to be reliable attributes of writing quality (Crossley & McNamara, 2009; Crossley et al., 2011; Lu, 2011; Norris & Ortega, 2009; Ortega, 2003; Skehan, 2009; Wolfe-Quintero et al., 1998). Recently, studies on syntactic and lexical complexity have increased exponentially due to the wide availability of computational tools. Syntactic complexity is an important concept in L2 writing and is often measured in terms of coordination, subordination, length of production, and sophistication of particular syntactic structures (e.g., complex nominals per T-unit). Previous studies have found mean length of clauses (MLC), mean length of T-units (MLT), T-unit complexity ratio (C/T), dependent clause ratio (DC/C), dependent clauses per T-unit (DC/T), and verb phrases per T-unit (VP/T) to be strong predictors of syntactic complexity (Lu, 2011; Ortega, 2003; Wolfe-Quintero et al., 1998). In addition, lexical complexity measures have also been used as valid development measures in L2 writing. It is viewed in terms of lexical density (the proportion of content words to total words), lexical sophistication (the proportion of unusual/infrequent and advanced words), and lexical variation (the range of vocabulary) (Read, 2000), and studies have found measures of lexical variation, such as the number of different words (NDW) and transformations of Type-Token Ratio (TTR), as well as lexical sophistication measures to be highly indicative of L2 development (Crossley et al., 2011; Grant & Ginther, 2000; Jarvis et al., 2003; McNamara et al., 2010).

As such, much work in this line of research has focused on determining which linguistic measures are best indicators of L2 developmental levels in written productions that consistently increase in a linear relationship as the quality of writing increases. In most of these studies, the quality of writing or L2 proficiency

level was assessed using holistic or analytic scores of written samples, and linguistic measures were analyzed using automated computational tools, such as L2 Syntactic Complexity Analyzer (Lu, 2010, 2011), Coh-Metrix (McNamara et al., 2014), and Lexical Complexity Analyzer (Lu, 2012). A number of studies have provided robust evidence that support the validity of computational tools in measuring L2 writing development and have demonstrated that automated linguistic indices can not only help predict human ratings of L2 essays but also help understand various constructs of L2 writing proficiency (Crossley et al., 2012; Guo et al., 2013; Jung et al., 2019; Kyle & Crossley, 2015).

While previous research provides strong support for linguistic features (i.e., computational indices) as significant predictors of writing quality as measured by human raters, it is unclear whether computational indices of written samples can also be significant predictors of other proficiency measures that highly correlate with writing proficiency. Vice versa, we cannot be certain whether other measures of proficiency that highly correlate with writing proficiency can be predictive of these computational indices of linguistic complexity in L2 writing development. Given the high correlation between cloze tests and writing proficiency (Chapelle & Abraham, 1990; Fotos, 1991; Hanania & Shikhani, 1986; Lee, 1997; Stubbs & Tucker, 1974), it is worth examining the relationship between cloze test scores and various linguistic features in L2 writing which can be measured by computational tools. Ryoo (2017) examined the relationship between cloze tests and written productive vocabulary of Korean EFL learners using VocabProfile (Cobb, 2002), which analyzes lexical profiles from frequency word lists. She found that cloze test scores can serve as a reliable predictor of lexical knowledge and production of L2 learners. In fact, the cloze test was better able to distinguish lexical features in different proficiency groups than the TOEIC, a standardized proficiency test. Based on the high commonality between the cloze test and L2 writing proficiency, it can be predicted that cloze test scores will exhibit significant relationships with computational indices of syntactic and lexical complexity, but no study as of yet has examined this relationship.

2.2. The cloze test

As an integrative measure of general L2 proficiency (Fotos, 1991; Eckes & Grotjahn, 2006; Lee, 1997 among others), cloze tests have often been used as a supplementary placement procedure for students entering or leaving language

programs. In particular, cloze tests have been found to be highly predictive of L2 learners' writing scores (Chapelle & Abraham, 1990; Fotos, 1991; Hanania & Shikhani,1986; Lee, 1997; Stubbs & Tucker, 1974). Hanania and Shikhani (1986) reported on the substantial correlation between cloze tests and writing tests in addition to the significant correlation between cloze tests and standardized test for ESL students. Fotos (1991) also confirmed that cloze test scores are significantly correlated with the essay scores of Japanese EFL students. Lee (1997), who compared Korean EFL students' cloze test scores, essay test scores, and college entrance English test scores, confirmed that the cloze test could be used as a measure of EFL students' writing proficiency. These studies suggest that the cloze test can serve as a powerful and economical diagnostic tool for L2 writing proficiency that practitioners can easily use in the classroom (Stubbs & Tucker, 1974).

Designing and constructing a cloze test entails consideration of many variables such as scoring methods (exact-word and acceptable-word), deletion patterns (rational deletion and fixed ratio deletion), response types (open-ended, multiple-choice, and C-test), and item difficulty (the frequency of the word in the passage, word length, the number of occurrences of a test item word in the passage, the text difficulty, the length of a word to be restored, etc.). Alderson (1979a, b, 1980) observed that different formats in text difficulty, scoring procedure, and deletion frequency seemed to produce different cloze tests which measure different language abilities. Among the many variables, the scoring method may be of high interest to classroom teachers who want to apply the students' cloze test scores directly into their lesson design. Studies have often compared the acceptable-word scoring method which accepts alternative words in the context of the deleted words to the exact-word scoring method in which the responses must match the words deleted from the original text. Although the two scoring methods were found to be significantly correlated (Stubbs & Tucker, 1974), many have suggested that the acceptable-word scoring method is superior to the exact-word scoring method in that it provides better differentiation with L2 learners (Brown, 1980; Brown et al., 2001; Kobayashi, 2002; Oller, 1972; Porter, 1978). However, such favorable view of the acceptable-word scoring method has been challenged by several studies (Kim, 1994; Ryoo, 2017) that have not found significant differences between the two scoring methods. That is, it is still uncertain whether the scoring method indeed has an effect on the predictability of cloze tests, especially when it comes to predicting L2 writing proficiency.

In addition to the role of the scoring method, the role of proficiency in the relationship between cloze tests and linguistic features in syntactic and lexical complexity must also be further investigated. The effect of proficiency should be of interest to many teachers who have different levels of students in the classroom. Fotos (1991) observes that what cloze tests measure may vary depending on the proficiency level of the test-takers. According to this observation, the cloze test tends to measure basic skills and does not function well when conducted with beginners or low-intermediate proficiency learners. In contrast, it can assess advanced integrative skills and show significant correlations with other integrative measures such as the TOEFL at more advanced levels. Such observations must be further corroborated, and the relationship between proficiency (as measured by the cloze test) and linguistic features in L2 writing must be explored. With respect to linguistic features, many studies have found that the relationship between L2 writing proficiency and linguistic features vary depending on the L2 learners' proficiency level with different features correlating at varying degrees of strength according to levels of L2 development (Azizollah et al., 2012; Norris & Ortega, 2009; Schoonen et al., 2011). For example, coordination exhibits strong predictability at low proficiency levels, whereas subordination and phrasal complexity are more predictive with increasing proficiency (Norris & Ortega, 2009). Moreover, advanced learners produce more reduction phrases, nominalizations, and complex sentences with less coordination, which results in a slight decrease in production length (e.g., mean length of sentences, mean length of t-units) (Wolfe-Quintero et al., 1998). Whether such differentiation of linguistic features in L2 writing development can also be observed when proficiency levels are determined by cloze test scores is another question of interest in the present study.

The specific research questions for the study are addressed as follows:

1. Which linguistic features in syntactic and lexical complexity as measured by computational tools are most highly correlated with cloze test scores? What linguistic features and aspects of L2 writing can the cloze test examine?
2. Does the relationship between cloze test scores and linguistic features in L2 writing differ depending on the scoring method (acceptable-word vs. exact-word) and L2 proficiency level as determined by cloze test scores?

3. Method

3.1. Participants

The participants in this study were 60 Korean college students (26 males and 34 females) attending a public university in Seoul, Korea. All of them were freshmen taking the mandate general English course which is focused on English composition. Two-thirds of the participants were majoring in humanities and social sciences, while one-third of the participants were majoring in sciences. The participants were divided into three proficiency groups based on the cloze test scores in each scoring method. Students with scores less than 25th percentile of all the scores were included in the low proficiency group and students with scores greater than 75th percentile in the high proficiency group. The remaining others were included in the intermediate group. For the exact-word scoring method, there was a total of 17 high (M: 15.59, SD: 1.81), 22 intermediate (M: 10.46, SD: 1.87), and 21 low (M: 5.48, SD: 1.78) proficiency learners. For the acceptable-word scoring method, there was a total of 17 high (M: 24.18, SD: 2.58), 26 intermediate (M: 16.57, SD: 2.86), and 17 low (M: 8.18, SD: 2.79) proficiency learners. The proficiency groups were significantly different from each other in both scoring methods (exact $F(2, 57) = 145.439$, $p<.001$; acceptable $F(2, 57) = 142.626$, $p<.001$).

3.2. Instruments

Two types of instruments, a cloze test and a writing test, were administered at the beginning of the semester as diagnostic tests for the course. The text of the cloze test was taken from a passage in *American Kernel Lessons: Advanced Students' Book* (O'Neill et al., 1981) that has been adapted in several studies such as Chae and Shin (2015), Ionin et al. (2013), and Montrul (1997). Chae and Shin (2015) especially confirmed that this cloze test is an adequate measure of English proficiency for Korean learners of English. In this test, there were 40 blanks with blanks at every 7th word, and the participants were required to provide only one word as an answer per blank. The cloze test consisted of 23 function words (prepositions, pronouns, auxiliary verbs, conjunctions, determiners, relatives, pronouns, articles, and particles) and 17 content words (nouns, verbs, adjectives, and adverbs) and could be completed in 20 minutes (see Appendix). The writing test asked the students to write a descriptive paragraph about three characteristics

of a successful college life. Students were given 30 minutes and were not allowed to use the dictionary or other language references for both tests. The tests were given to the participants in a paper-and-pencil format.

3.3. Cloze test scoring

The cloze test was graded using two different scoring methods: the exact-word scoring method and the acceptable-word scoring method. The former does not allow any word other than the word deleted from the original text, while the latter allows alternative words to be acceptable in the context of the deleted words. For the acceptable-word scoring method, the two researchers cross-checked their scoring results and made an agreement on the alternative answers based on the criteria taken from Stubbs and Tucker (1974): Non-grammatical forms (e.g., she do for she does) were all excluded and no more than one word was allowed per blank. Responses with similar meaning were accepted if they were grammatically and contextually appropriate even if they differed from the original word in part of speech (POS). Spelling errors were penalized, but responses with capitalization errors were accepted. Examples of the acceptable answers based on these criteria are shown in Table 1.

Table 1. Examples of acceptable answers for two scoring methods

Test item	Exact-word answers	Acceptable-word answers	
		1) same POS	2) different POS
It was payday, but he wasn't (1) _____ excited about it.	even	very, so, really	feeling
He drove into a quiet country (18) _____.	road	town, field	soon, quickly, finally
The country sights made him feel (19) _____.	better	comfortable, relaxed, good	
His mind wandered as he drove (20) _____ small farms.	past	to, near, around, through	

The Cronbach α was 0.760 for the exact-word scoring method and 0.857 for the acceptable-word scoring method. The mean score of the acceptable-word scoring method (M = 13.35, SD = 6.66) was significantly higher than that of the exact-word scoring method (M = 10.17, SD = 4.42) ($t(59) = -17.68$, $p<.001$), but simple

bivariate correlation analysis via Pearson coefficient showed that the two scoring methods were highly correlated ($r = .96$) at a statistically significant level ($p<.001$).

3.4. Linguistic measures

The linguistic indices in the present study were measured using the L2 Syntactic Complexity Analyzer and the Lexical Complexity Analyzer.[1] We used a total of 38 indices that are predetermined by the computational analyzers with 2 indices measuring text length, 14 indices measuring syntactic complexity, and 22 indices measuring lexical complexity. A summary of the linguistic indices investigated in the study is presented in Table 2 below.

Table 2. Summary of linguistic measures investigated in the study

Category	Type	Measure (Code)	Formula
Text length	Text length	Word count (W)	# of words
		Sentence (S)	# of sentences
Syntactic complexity	Length of production	Mean length of sentence (MLS)	# of words/# of sentences
		Mean length of T-unit (MLT)	# of words/# of T-units
		Mean length of clause (MLC)	# of words/# of clauses
	Sentence complexity	Clause per sentence (C/S)	# of clauses/# of sentences
	Subordination	Clause per T-unit (C/T)	# of clauses/# of T-units
		Complex T-unit ratio (CT/T)	# of complex T-unit/# of T-units
		Dependent clause per clause (DC/C)	# of dependent clauses/# of clauses
		Dependent clause per T-unit (DC/T)	# of dependent clauses/# of T-units
	Coordination	T-unit per sentence (T/S)	# of T-units/# of sentences
		Coordinate phrase per clause (CP/C)	# of coordinate phrases/# of clauses
		Coordinate phrase per T-unit (CP/T)	# of coordinate phrases/# of T-units
	Particular structures	Complex nominal per T-unit (CN/T)	# of complex nominals/# of T-units
		Complex nominal per clause (CN/C)	# of complex nominals/# of clauses
		Verb phrase per T-unit (VP/T)	# of verb phrases/# of T-units

1) See Lu (2010, 2011) for a full description of each index for the L2 Syntactic Complexity Analyzer and Lu (2012) for the Lexical Complexity Analyzer.

Table 2. Continued

Category	Type	Measure (Code)	Formula
Lexical complexity	Lexical density	Lexical density (LD)	N_{lex} / N
	Lexical sophistication[2)]	Lexical sophistication-I (LS1)	N_{slex} / N_{lex}
		Lexical sophistication-II (LS2)	T_s / T
		Verb sophistication-I (VS1)	T_{sverb} / N_{verb}
		Verb sophistication-II (VS2)	T_{sverb} / $\sqrt{2N_{verb}}$
		Corrected VS1 (CVS1)	T^2_{sverb} / N_{verb}
	Lexical variation	Number of different words (NDW)	T
		(NDWZ)	T (first 50 words)
		(NDWERZ)	T (expected random 50)
		(NDWESZ)	T (expected sequence 50)
		Type/Token ratio (TTR)	T/N
		Corrected TTR (CTTR)	T/ $\sqrt{2N}$
		Root TTR (RTTR)	T/ $\sqrt{N}$
		Lexical word variation (LV)	T_{lex}/ N_{lex}
		Verb variation-I (VV1)	T_{verb}/ N_{verb}
		Squared VV1 (SVV1)	T^2_{verb}/ N_{verb}
		Corrected VV1 (CVV1)	T_{verb}/ $\sqrt{2N_{verb}}$
		Verb variation-II (VV2)	T_{verb}/ N_{lex}
		Noun variation (NV)	T_{noun}/ N_{lex}
		Adjective variation (ADJV)	T_{adj}/ N_{lex}
		Adverb variation (ADVV)	T_{adv}/ N_{lex}
		Modifier variation (MODV)	(T_{adj} + T_{adv})/ N_{lex}

Notes. N = the number of words; N_{lex} = the number of lexical words; N_{slex} = the number of sophisticated lexical words; N_{verb} = the number of verbs; T = the number of word types; T_{lex} = the number of lexical word types; T_s = the number of sophisticated word types; T_{sverb} = the number of sophisticated verb types; # = number; / = divided by; T-unit: one main clause + any subordinate clause.

2) Words are regarded as sophisticated when they are not only the list of the 2,000 most frequent words, as ranked by the American National Corpus.

3.5. Data analysis

A series of statistical analyses were performed using SPSS version 25.0 to analyze the complexity index scores of the students' compositions and the cloze test scores of the two scoring methods. Of the 38 linguistic indices (Table 2), nine indices that do not meet the normality assumption were removed from data analysis[3] resulting in one index for text length (W), 13 indices for syntactic complexity (MLS, MLT, MLC, C/S, C/T, CT/T, DC/C, DC/T, T/S, CP/C, CN/T, CN/C, VP/T), and 15 indices for lexical complexity (LD, LS1, LS2, NDWERZ, NDWESZ, CTTR, RTTR, LV, SVV1, CVV1, VV2, NV, ADJV, ADVV, MODV). First, a series of Pearson product-moment correlation coefficients were computed to examine whether there are significant and meaningful relations between the 29 normally distributed linguistic indices and the cloze test scores. Separate analysis was conducted for each scoring method, and significant correlations were compared. Second, participants were divided into three proficiency groups based on the cloze test scores in each scoring method in order to examine the effect of proficiency on the relationship between linguistic indices and cloze test scores. Pearson's bivariate correlations between linguistic index scores and cloze test scores were conducted again for each proficiency group and scoring method, and multivariate analysis of variance (MANOVA) was conducted to examine group differences between proficiency levels.

4. Results

Of the 29 indices that demonstrated normal distributions, 11 indices (W, MLS, MLT, MLC, CN/T, CN/C, LD, NDWERZ, CTTR, RTTR, SVV1) significantly correlated with the cloze test scores, and one index (VP/T) showed a marginally significant correlation with the cloze scores. All significant indices were checked for multicollinearity, and one index (RTTR) was excluded from further analysis because it correlated with another index (CTTR) at $|r| \geq .90$. The linguistic complexity indices that significantly correlated with the cloze test scores were the same in both scoring methods, which indicates that the different methods of scoring (acceptable-word vs. exact-word) do not differ in their relationships with the linguistic measures

3) Indices that did not meet the normality assumption even after natural log transformation were removed from further statistical analysis.

examined. Moderate correlations[4] were found between cloze test scores and scores of text length (W) as well as measures of length of production (MLS, MLT, MLC), whereas weak correlations were found with indices that measure structures with complex nominals[5] (CN/T, CN/C), lexical density (LD), and lexical variation (NDWERZ, CTTR, SVV1). With the exception of LD, all of these indices were linearly correlated and increased as the cloze test score increased. The results are summarized in Table 3.

Table 3. Significant correlations between cloze test scores and linguistic complexity features

	Index	Correlation coefficient	
		Acceptable-word scoring	Exact-word scoring
Text length	W	.592**	.584**
Syntactic complexity	MLS	.457**	.457**
	MLT	.513**	.501**
	MLC	.490**	.506**
	CN/T	.399**	.368**
	CN/C	.349**	.331**
Lexical complexity	LD	−.292*	−.306*
	NDWERZ	.264*	.264*
	CTTR	.370**	.362**
	RTTR	.371**	.363**
	SVV1	.326*	.360**

Note. *p<.05, **p<.01; CN/C, CN/T, MLC, MLS, MLT, VP/T, W: Log transformed

To examine the effect of L2 proficiency, the participants were divided into three proficiency groups based on the cloze test scores in each scoring method. When correlations between linguistic index scores and cloze test scores were examined for each proficiency group and scoring method, differences could be found between

4) Correlations were characterized as high if the absolute value of coefficient r was greater than .65, moderate if between .45 and .65, and weak if between .25 and .45 (Wolfe-Quintero et al., 1998).

5) According to Cooper (1976), complex nominals are (1) nouns plus adjective, possessive, prepositional phrase, relative clause, participle, or appositive, (2) nominal clauses, and (3) gerunds and infinitives in subject position.

proficiency groups and between the two scoring methods.

Overall, the high proficiency group had the greatest number of significant correlations, albeit only in lexical variation. Moderate to high correlations were found between cloze test scores and scores related to lexical variation in both the acceptable-word (CTTR $r = .589$, $p = .013$; LV $r = .508$, $p = .037$; SVV1 $r = .696$, $p = .002$) and exact-word (LV $r = .568$, $p = .017$; SVV1 $r = .563$, $p = .019$; ADVV $r = -.657$, $p = .004$) scoring methods. As for the intermediate group, none of the linguistic features significantly correlated with the group's cloze test scores. In fact, the intermediate group's cloze scores only marginally correlated with lexical density and that only in the acceptable-word scoring method (LD $r = -.382$, $p = .054$). In the low proficiency group, cloze test scores were strongly correlated with text length (W: acceptable $r = .701$, $p = .002$; exact $r = .731$, $p<.001$) and moderately correlated with adverb variation (ADVV: acceptable $r = -.570$, $p = .017$; exact $r = -.612$, $p = .003$) in both scoring methods. A marginally significant correlation was found with lexical variation in the acceptable-word scoring method (CTTR: acceptable $r = .479$, $p = .052$). All of the significant correlations were in the positive direction except for the correlations with adverb variation. The significant correlations across proficiency groups are summarized in Table 4 below.

Table 4. Significant correlations between cloze test scores and linguistic features across proficiency groups

Index	Acceptable-word scoring			Exact-word scoring		
	High	Int	Low	High	Int	Low
W	.361	.009	**.701****	.289	.133	**.731****
CTTR	**.589***	−.048	.479	.457	−.046	.399
LV	**.508***	.141	.147	**.568***	.132	.014
SVV1	**.696****	.069	.312	**.563***	.175	.279
ADVV	−.326	.053	**−.570***	**−.657****	−.165	**−.612****

Note. $^{*}p<.05$, $^{**}p<.01$; W, MLT, MLC, T/S, CP/C: Log transformed

MANOVA was conducted separately for each scoring method to examine the effect of cloze proficiency level (high, intermediate, low) on linguistic complexity features in writing. There was a significant main effect of proficiency group on linguistic features in both the acceptable-word scoring (Wilk's Lambda = .059, F(54,

52) = 2.991, p<.001) and the exact-word scoring (Wilk's Lambda = .034, F(54, 52) = 4.284, p<.001). The indices for which the proficiency level had a main effect were the same in both scoring methods. Main effect of proficiency was mainly found in text length (W), length of production (MLS, MLT, MLC), and particular structures with complex nominals (CN/T, CN/C) with values increasing at higher proficiency levels. There was no main effect of proficiency for any lexical complexity measure. Post-hoc comparisons using the Tukey HSD test revealed significant differences between high and low proficiency groups in all of the aforementioned linguistic indices in both methods of scoring. The intermediate proficiency group showed significant differences in the measure of text length (W) with other groups but showed only marginally significant differences in the rest of the linguistic features. To sum up, proficiency group differences mainly occurred between low proficiency and high proficiency learners in text length, length of production, and complex nominal use. These results as well as the mean values and standard deviation of linguistic measures affected by proficiency are presented in Tables 5 and 6 below.

Table 5. Main effect of cloze proficiency group on linguistic complexity measures

Index	Acceptable-word scoring			Exact-word scoring		
	F	*p*	Post-hoc	*F*	*p*	Post-hoc
W	9.796	.000***	I < H* L < H***	8.620	.001**	L < H ** L < I *
MLS	3.857	.027*	L < H *	5.032	.010*	L < H **
MLT	5.909	.005**	L < H **	6.100	.004**	L < H **
MLC	5.575	.006**	L < H **	7.366	.002**	L < H **
CN/T	4.438	.017*	L < H *	3.519	.037*	L < H *
CN/C	4.056	.023*	L < H *	3.343	.043*	L < H *

Note. *p<.05, **p<.01, ***p<.001; H: high, I: intermediate, L: low; (log transformed: W, CN/C, CN/T, MLC, MLS, MLT).

Table 6. Mean values and standard deviation of linguistic measures affected by proficiency

Index	Acceptable-word scoring			Exact-word scoring		
	High	Int	Low	High	Int	Low
W	206.82 (77.18)	159.38 (37.88)	128.82 (33.13)	199.24 (82.06)	170.86 (35.92)	128.76 (30.36)
MLS	14.36 (3.56)	12.32 (2.78)	11.16 (3.56)	14.22 (3.28)	12.82 (3.15)	10.97 (3.22)
MLT	13.44 (3.45)	11.49 (1.83)	10.33 (2.46)	13.15 (3.31)	12.00 (2.19)	10.25 (2.21)
MLC	8.85 (1.58)	7.93 (1.21)	7.28 (1.21)	8.82 (1.58)	8.15 (1.21)	7.19 (1.12)
CN/T	1.61 (.61)	1.41 (.35)	1.21 (.58)	1.60 (.58)	1.42 (.40)	1.24 (.53)
CN/C	1.06 (.36)	.98 (.28)	.83 (.29)	1.07 (.36)	.98 (.29)	.85 (.28)

5. Discussion

The present study investigated the relationship between cloze test scores and linguistic features in syntactic and lexical complexity in L2 writing and observed the effect of different scoring methods (acceptable-word vs. exact-word) and proficiency level on this relationship.

5.1. The relationship between cloze test scores and linguistic features

Overall, we found that cloze tests scores can predict linguistic features that have consistently been found to be good indicators of L2 writing proficiency. Cloze test scores were significantly correlated with text length (W), length of production (MLS/MLT/MLC), complex nominals (CN/T, CN/C), lexical variation (NDWERZ, CTTR, SVV1), and lexical density (LD)[6]. Also, marginally significant correlation was found with the number of verb phrases (VP/T). These results suggest that students with higher cloze test scores are more likely to write longer texts and sentences/t-units/clauses and produce a greater number of complex nominalizations

6) Previous studies did not find a statistically significant relationship between lexical density and L2 writing proficiency, but LD was significantly correlated to cloze test scores in the negative direction in the present study. This could be interpreted as learners with low cloze test scores being more likely to omit function words, such as articles and prepositions, and predominantly using content words than learners with higher cloze test scores.

and verb phrases, a wider range of vocabulary, and a higher proportion of function words than those with lower cloze test scores. Previous research reveals robust evidence that text length and length of production measures can reliably distinguish L2 writing proficiency with higher scoring essays containing more words and longer words on average (Ferris, 1994; Grant & Ginther, 2000; McNamara et al., 2010; Wolfe-Quintero et al., 1998). Moreover, syntactic features such as complex nominal forms and phrases have often been used as valid measures of syntactic complexity in L2 writing (Lu, 2011; McNamara et al., 2010; Ortega, 2003; Park, 2012). Studies show that complex nominalization linearly increase across proficiency levels and are one of the best discriminators between proficiency levels. Moreover, the relationship between high-scoring essays and lexical variation/diversity is well attested and unquestioned (Grant & Ginther, 2000; Jarvis et al., 2003). Put together, linguistic features in syntactic and lexical complexity that have been found to be valid measures of L2 writing proficiency in previous research were significantly correlated with cloze test scores, and these results suggest that cloze test scores can reliably predict computational indices of linguistic complexity that are informative of important linguistic features in L2 writing. That is, the present study found that cloze tests can reveal specific linguistic aspects of L2 writing and vice versa—that computational index scores of written samples can predict not only writing quality as previously found but scores of another proficiency measure such as the cloze test.

5.2. The effect of L2 proficiency and the scoring method of cloze tests

When analysis was performed separately for three proficiency groups (proficiency as determined by cloze test scores), the linguistic indices that significantly correlated with cloze test scores were different from those of the whole group results and varied by proficiency group. Overall, the cloze test scores could reliably predict high proficiency learners' lexical complexity features especially in the area of lexical variation/diversity. In contrast, no linguistic feature showed significant correlations with the cloze test scores of the intermediate proficiency group, which suggests that cloze test scores that are neither high nor low are not very informative in predicting linguistic features in syntactic and lexical complexity. As for the low proficiency group, their cloze test scores could reliably predict text length in both scoring methods. Adverb variation (i.e., the proportion of adverb types to total lexical words), which was also significantly correlated in the negative direction, is speculated to be caused by a lower number of total lexical words, rather than a

higher number of adverb types, by those with lower cloze test scores. As such, cloze test scores correlated with different linguistic indices depending on the proficiency level and displayed strong correlations with high proficiency learners' lexical complexity features and low proficiency learners' text length, but almost no correlations for the intermediate level. These findings confirm Fotos' (1991) observation that what cloze tests measure can be variable depending on the proficiency level of the test-takers. When group differences were examined, a strong main effect of proficiency level was found in measures of text length (W), length of production (MLS, MLT, MLC), and particular structures with complex nominals (CN/T, CN/C) for both scoring methods. Significant group differences were found primarily between low proficiency and high proficiency levels, but not so much with the intermediate level, which is consistent with the lack of significant correlations for this particular group in the present findings.

As for the effect of different methods of scoring (acceptable-word vs. exact-word), the linguistic indices that were significantly correlated with cloze test scores were the same in both scoring methods. Even when the students were divided into three proficiency levels, results for both scoring methods displayed similar patterns in significant correlations and differed only in marginally significant relationships. This is consistent with the findings of Kim (1994) and Ryoo (2017) in which different scoring methods did not significantly affect the predictability of the cloze test. Although the mean score of the acceptable-word scoring method was significantly higher than that of the exact-word scoring method, it can be said that both scoring methods are equally effective in predicting linguistic features that measure L2 writing proficiency.

6. Conclusion

The present findings provide support for cloze tests as an appropriate assessment tool that can predict linguistic features of syntactic and lexical complexity in L2 writing. More specifically, cloze test scores have a linear relationship with text length, length of production, the number of complex nominals, and the range of vocabulary, all of which are thought to be indicative of L2 writing proficiency. Moreover, cloze tests can effectively tease apart differences between high and low proficiency levels especially in syntactic complexity measures of length and nominalization that are reliable measures of L2 writing development. Considering

the relative ease and efficiency of administering and evaluating cloze tests when compared to writing tests, educators in L2 writing classes can use the cloze test in the beginning of the semester to be informed of the above linguistic features in writing. The information that can be gained by cloze tests is by no means comprehensive enough to evaluate overall writing proficiency, but these features can be used as an initial preliminary guide for proficiency placement or for assigning teams or peer review groups. However, it must be pointed out that the cloze test may not be effective for all proficiency levels. We found that cloze test scores have almost no predictability for the intermediate proficiency group and can be used to predict text length with the low proficiency group but not much else. The cloze test seems to measure different aspects of L2 writing depending on the proficiency level, and thus practitioners using this assessment tool must be careful not to jump to conclusions based on the cloze test scores alone. In fact, we recommend that cloze tests be used in addition to writing compositions; that is, the cloze test can supplement but not substitute writing compositions.

The present study is not without limitations. The low mean scores for the cloze tests in both scoring methods (acceptable-word 13.35, exact-word 10.17 out of a total of 40 points) suggest that either (1) the overall proficiency level of the learners in the study was fairly low or (2) the cloze test used in the study in the present format was too difficult for L2 learners. In fact, previous studies that have used the same cloze test usually adapted it in the multiple-choice format with answer choices (Chae & Shin, 2015; Ionin et al. 2013) whereas only blanks were given in the present study. Also, dividing proficiency groups based on percentiles and the learners' relative scores may not accurately reflect the learners' actual proficiency level. Therefore, the present findings must be corroborated with different populations using cloze tests of different texts, difficulty level, and format. Moreover, the writing samples in the present study were relatively short descriptive paragraphs (100-200 words), and therefore the study must be replicated with longer essays of different genres. Also, proficiency effect must be examined using other independent proficiency assessments in order to fully explore the effect of proficiency. Using cloze test scores as a measure of proficiency as well as the primary variable is less than ideal in that it could cause a confound in the findings as pointed out by one of the reviewers. Lastly, the present study is only concerned with linguistic complexity of L2 writing and fails to address other important aspects such as accuracy, content, and organization. Such critical aspects of writing must be additionally addressed in future studies, and future researchers must closely examine the various factors that determine the predictability

of cloze tests in L2 writing classrooms.

References

Alderson, J. C. (1979a). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly, 13*(2), 219-227.

Alderson, J. C. (1979b). The effect on the cloze test of changes in deletion frequency. *Journal of Research in Reading, 2*(2), 108-119.

Alderson, J. C. (1980). Native and nonnative speaker performance on cloze tests. *Language Learning, 30*(1), 59-76.

Azizollah, D., Reza, Z., & Mohsen, R. (2012). Argumentative and narrative written task performance: Differential effects of critical thinking. *International Journal of Research Studies in Language Learning, 24*(1), 1-12.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly, 16*, 61-70.

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *The Modern Language Journal, 64*(3), 31-317.

Brown, J. D. (1983). A closer look at cloze: Validity and reliability. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 237-250). Rowley, MA: Newbury House.

Brown, J. D., Yamashiro, A. D., & Ogane, E. (2001). The emperor's new cloze: Strategies for revising cloze tests. In T. Hudson & J. D. Brown (Eds.), *A focus on language test development* (pp.143-161). Honolulu, HI: University of Hawaii.

Chae, E., & Shin, J. (2015). A study of a timed cloze test for evaluating L2 proficiency. *English Teaching, 70*(3), 117-135.

Chapelle, C., & Abraham, R. G. (1990). Cloze method: What difference does it make? *Language Testing, 7,* 121-46.

Cobb, T. (2002). *VocabProfilers* [computer software]. Retrieved from https://www.lextutor.ca/vp/

Cooper, T. C. (1976). Measuring written syntactic patterns of second language learners of German. *Journal of Educational Research, 69*(5), 176-183.

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.

Crossley, S. A., Weston, J. L., Sullivan, S. T. M., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28*(3), 282-311.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing, 29,* 243-263.

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing, 23*(3), 290-325.

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.

Fotos, S. S. (1991). The cloze test as an integrative measure of EFL proficiency: A substitute for essays on college entrance examinations? *Language Learning, 41*(3), 313-336.

Gellert, A. S., & Elbro, C. (2013). Cloze tests may be quick, but are they dirty? Development and preliminary validation of a cloze test of reading comprehension. *Journal of Psychoeducational Assessment, 31*(1), 16-28.

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing, 9*(2), 123-145.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18,* 218-238.

Hanania, E., & Shikhani, M. (1986). Interrelationships among three tests of language proficiency: Standardized ESL, cloze, and writing. *TESOL Quarterly, 20*(1), 97-109.

Harsch, C., & Hartig, J. (2016). Comparing C-tests and yes/no vocabulary size tests as predictors of receptive language skills. *Language Testing, 33*(4), 555-575.

Ionin, T., Montrul, S., & Crivos, M. (2013). A bidirectional study on the acquisition of plural noun phrase interpretation in English and Spanish. *Applied Psycholinguistics, 34*(3), 483-518.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing, 12*(4), 377-403.

Jung, YJ., Crossley, S., & McNamara, D. (2019). Predicting language writing proficiency in learner texts using computational tools. *The Journal of Asia TEFL, 16*(1), 37-52.

Kim, S.-S. (1994). The cloze procedure and EFL proficiency. *English Teaching, 48*, 127-150.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing, 19*(2), 193-220.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*(4), 757-786.

Lee, S. (1997). Cloze test as a measure of EFL writing proficiency. *English Teaching, 52*(3), 151-172.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474-496.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*(1), 36-62.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*(2), 190-208.

Markham, P. (1987). Rational deletion cloze processing strategies: ESL and native English. *System, 15*(3), 303-311.

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

McNamara, D. S., Graesser, A.C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation*

of text and discourse with Coh-Metrix. Cambridge: Cambridge University Press.

Montrul, S. A. (1997). *Transitivity alternations in second language acquisition: A crosslinguistic study of English, Spanish and Turkish*. Unpublished doctoral dissertation. Montreal, Canada: McGill University.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555-578.

Oller, W. (1972). Scoring methods and difficulty levels for cloze tests of ESL proficiency. *The Modern Language Journal, 56*(3), 151-158.

O'Neill, R., Cornelius, E. T., & Washburn, G. N. (1981). *American kernel lessons: Advanced students' book*. London: Longman.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492-518.

Park, S-Y. (2012). A corpus-based study of syntactic complexity measures as development indices of college-level L2 learners' proficiency in writing. *Korean Journal of Applied Linguistics, 28*(3), 139-160.

Porter, D. (1978). Cloze procedure and equivalence. *Language Learning, 28*, 333-340.

Read, J. (2000). *Assessing vocabulary*. Oxford: Oxford University Press.

Ryoo, Y. (2017). Predictability of the cloze test as a measure of written productive vocabulary. *Modern English Education, 18*(4), 25-45.

Schoonen, R., Gelderen, A. V., Stoel, R. D., Hulstijn, J., & Glopper, K. D. (2011). Modeling the development of L1 and EFL writing proficiency of secondary school students. *Language Learning, 61*(1), 33-79.

Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly, 17*(2), 229-255.

Skehan, P. (2009). Modeling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510-532.

Stubbs, J. B., & Tucker, G. R. (1974). The cloze test as a measure of English proficiency. *The Modern Language Journal, 58*(5/6), 239-241.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly, 30*, 414-438.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu, HI: University of Hawaii Press.

Eun Seon Chung
Research Professor
Language Research Institute
Hankuk University of Foreign Studies
Yimun-ro 107, Dongdaemun-gu, Seoul 02450, Korea
E-mail: prolingesc@gmail.com

Soojin Ahn
Visiting Professor
Division of General English
University of Seoul
163 Seoulsiripdaero, Dongdaemun-gu, Seoul 02504, Korea
E-mail: soojina2013@uos.ac.kr

Received: October 31, 2019
Revised version received: December 2, 2019
Accepted: December 18, 2019

Appendix

Cloze Test

Please fill in the blanks in the following passage. Each blank must have one and only one word.

Joe came home from work on Friday. It was payday, but he wasn't **(1)** even excited about it. He knew that **(2)** when he sat down and paid his **(3)** bills and set aside money for groceries, **(4)** gas for the car, and a small **(5)** deposit in his savings account, there wouldn't be **(6)** too much left over for a good **(7)** life.

He thought about going out for **(8)** dinner at his favorite restaurant, but he **(9)** just wasn't in the mood. He wandered **(10)** around his apartment and ate a sandwich. **(11)** For a while, he couldn't stop himself **(12)** from worrying about the money situation. Finally, **(13)** he got into his car and started **(14)** driving. He didn't have a destination in **(15)** mind, but he knew that he wanted **(16)** to be far away from the city **(17)** where he lived.

He drove into a quiet country **(18)** road. The country sights made him feel **(19)** better. His mind wandered as he drove **(20)** past small farms and he began to **(21)** imagine living on his own piece of **(22)** land and becoming self-sufficient. It had always **(23)** been a dream of his, but he **(24)** had never done anything to make it **(25)** a reality. Even as he was thinking, **(26)** his logical side was scoffing at his **(27)** impractical imaginings. He debated the advantages and **(28)** disadvantages of living in the country and **(29)** growing his own food. He imagined his **(30)** farmhouse equipped with a solar energy panel **(31)** on the roof to heat the house **(32)** in winter and power a water heater. **(33)** He envisioned fields of vegetables for canning **(34)** and preserving to last through the winter. **(35)** If the crops had a good yield, **(36)** maybe he could sell the surplus and **(37)** buy some farming equipment with the extra **(38)** money.

Suddenly, Joe stopped thinking and laughed **(39)** out loud, "I'm really going to go **(40)** through with this?"