English Teaching, Vol. 70, No. 3, Autumn 2015
DOI: 10.15858/engtea.70.3.201509.117

A Study of a Timed Cloze Test for Evaluating L2 Proficiency*

Eun-Young Chae**
(Kwangwoon University)
Jeong-Ah Shin
(Dongguk University)

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

This study examined a timed cloze test for evaluating English proficiency in second language (L2) experimental research. Forty-five Korean college students were randomly assigned to either a timed or untimed condition. In the timed condition, the participants read the sentences of the text one phrase at a time, using the self-paced, cumulative, moving-window reading paradigm, and their reading time (RT) and accuracy were measured. In an untimed condition, the participants carried out a typical pencil-and-paper cloze test. Linear mixed-effect models were used to analyze the data. Although the accuracy data did not indicate any significant results, the RT data showed that the participants responded faster when they selected accurate answers and they answered function words rather than content words; also, as the participants' TOEIC scores increased, a marginally significant RT difference was observed. A significant correlation was also found between the cloze test and high TOEIC scores in the timed condition. The results showed that the timed cloze test used in the study can provide useful data for L2 experimental research in measuring L2 proficiency.

Key words: timed cloze test, self-paced, accuracy, response time (RT), linear mixed-effects (LME) model

1. INTRODUCTION

Proficiency is considered a significant factor in second language (L2) experimental research, so researchers have used various assessment tools in evaluating L2 proficiency

* The experiment reported here was based upon a portion of the first author's master thesis.
** Eun-Young Chae: First author; Jeong-Ah Shin: Corresponding author

(Mackey & Gass, 2012). Many studies frequently refer to exams that measure college students' L2 English proficiency such as TOEIC (Test of English for International Communication) and TOEFL (Test of English as a Foreign Language) (Lee, 2014). However, these assessment tests are limited, in that those participating in the experimental study do not always have scores from identical tests or may not accurately recall their scores; thus, in-house placement tests such as cloze tests (e.g., Ionin, Montrul, & Crivos, 2013) or self-ratings (e.g., Bernolet, Hartsuiker, & Pickering, 2013) have been used as an alternative in L2 experimental studies.

Cloze tests are a type of fill-in-the-blank test with blanks evey six or seven word in the passage, and have been argued to be an efficient way to evaluate global and general language proficiency (Alderson, 1979; Babaii & Ansary, 2001; Connelly, 1997; Dörnyei & Katona, 1992; Eckes & Grotjahn, 2006; Klein-Braley, 1997; Klein-Braley & Raatz, 1984; Weir, 1990). The typical pencil-and-paper cloze tests, however, cannot measure participants' fluency (i.e., spontaneous responses) because it is not time-constrained per item. As suggested in studies on language testing (McGrath, 2009), a time-constrained test can measure participants' ability accurately because spontaneous responses are allowed within a time limit.

As noted above, researchers have utilized cloze tests as a measure of general language proficiency to compare it with various skills such as vocabulary and writing skills or to supplement other L2 experiments. However, no known studies with a self-paced and time-measuring cloze test have been conducted. Therefore, this study explored the timed self-paced cloze test performance by analyzing the accuracy and response time with a linear mixed-effects (LME) model. In particular, this study investigated whether L2 learners' performance in the timed cloze test had a correlation with their TOEIC scores and self-ratings compared with that in the typical cloze test. In addition, the study examined whether the types of target words (i.e., function or content words) affected the self-paced cloze-test performance in terms of response accuracy and response time.

2. LITERATURE REVIEW

2.1. Cloze Tests

Taylor (1953) first derived the term *cloze* from the concept of closure in the Gestalt school of psychology (Stansfield & Hansen, 1983), where people tend to fill in the missing parts by using their background knowledge or prior experiences; for example, people can perceive an imperfect circle as a whole circle. He applied the theory of closure to test readability for native speakers of English. Since then, other researchers have applied the

cloze procedure to non-native speakers of English (Hinofotis & Snow; 1980, Oller; 1973, Oller & Inal; 1971, Stubbs & Tucker; 1974).

Cloze tests require examinees to find out the main idea from the reading passage and help realize inter-sentential or intra-sentential relationships (i.e., relationships between sentences or within a sentence) in order to reconstruct the meaning of the passage by filling in the blanks with appropriate words. In this respect, it has been suggested that cloze tests are more communicative than discrete-point tests that measure independent factors—such as grammar, vocabulary, spelling and punctuation, pronunciation, intonation and stress—and language skills—listening, speaking, reading, and writing—separately (Weir, 1990).

The original deletion-method of the cloze (the so-called typical or standard cloze test) is the systematic deletion (fixed-ratio deletion) designed to delete words from the text either mechanically (every n^{th} word; a typical test uses $n = 6$) or selectively, depending on the purpose of the test. In this fixed-word deletion cloze test, the readers are required to fill in the blanks that have been removed from the text. Another method invented by Bachman (1985) is unsystematic deletion (rational deletion), which is also called "guided cloze," a modified version of the cloze test. It contains random deletion of words for a particular purpose such as testing grammar, reading comprehension, and vocabulary (Kim & Cho, 2015) or contains a list of all the deleted words as a supplement (Lee, 2002).

Cloze tests can be further classified into three sub-types (Darwesh, 2010): the multiple-choice cloze test, the C-test, and the cloze elide. The multiple-choice cloze test is the most common type of cloze test, in which participants are provided with three to five choices for each blank, and it is easier than the typical cloze test (Chapelle & Abraham, 1990). Participants are required to choose one from the given options. It helps participants decrease their test anxiety, which will hopefully result in a better score; it is also considered reliable and objective. For language teachers, it helps them discover how knowledgeable students are of linguistic or content knowledge using syntactic and semantic cues to build meaning from the text, since it is easier to score, more economical, and less time-consuming (McNamara, 2000; Oller, 1979).

Second, the C-test, a modified version of the cloze test, was introduced by Klein-Braley and Raatz (1984), which contains blanks where the second half of a word is deleted. The first and last sentences are kept intact. An example of a C-test taken from the Rashid's study (2001) is as follows: *Once there was a merchant family who owned a dog and a donkey. One ni___ when th_ owner o_ the hou__ was fa__ asleep i_ his be_, a th___ broke in__ the ho___ to st___some o_ the own__ possessions.* Hosseini, Hassanzadeh, and Shayegh (2012) mentioned that some researchers have considered the C-test to be a highly valid, reliable and integrative measure of general language proficiency, and is easy to score and construct (Babaii & Ansary, 2001; Connelly, 1997; Dörnyei & Katona, 1992; Eckes &

Grotjahn, 2006; Klein-Braley, 1997; Klein-Braley & Raatz, 1984; Weir, 1990).

Last, in the cloze elide test, participants are required to cross out incorrect and superfluous words in the passage of the test, which is called "eliding" (Baker, 2011). This type of test requires observing long-range or immediate contextual constraints. Thus, it is considered as a test of text and style coherence and cohesion.

The aim of the cloze test is to measure not only reading comprehension of the content, but also readability (linguistic proficiency). Thus, several studies have compared cloze test performance with other linguistic skills. For example, Lee (1995, 1997) explored a correlation between standard cloze scores and L2 English writing proficiency. The study used a typical cloze test that deleted every fifth word of the passage and contained fifty blanks in total, and the deleted words covered grammatical and cohesive functions that were cued in the passages. The results showed that the cloze test could be an integrative measure of writing proficiency.

Also, Lee (2001) investigated the correlation between multiple choice cloze test scores and vocabulary, grammar, and background knowledge test scores in an EFL context. In this study, a modified version of the test was used with twenty blanks (70% function words and 30% content words among the target words). The results indicated modest positive correlations between the cloze test scores and the vocabulary, grammar, and background knowledge test, suggesting that the cloze test could be an economical and practical assessment tool for measuring overall language proficiency. He also suggested that the cloze test could be used as a substitution for vocabulary tests, multiple-choice reading comprehension tests, or background knowledge tests.

Moreover, cloze tests have been used as a supplemental assessment tool to conduct L2 experimental studies. For example, Ionin et al. (2013) measured L2 proficiency using a forced-choice (i.e., multiple-choice) cloze test, in which every seventh word was deleted, to divide participants into low and high proficiency groups and to examine their acquisition of plural noun phrase interpretation. In addition, Shin (2010) and Shin and Christianson (2012) administered a standard cloze test to determine proficiency effects in their main structural priming experiments. In these studies, the cloze test scores were used to divide participants into two or three proficiency groups or to function as a covariate or an independent variable in their statistical analyses.

2.2. Self-Paced and Timed Reading in Second Language

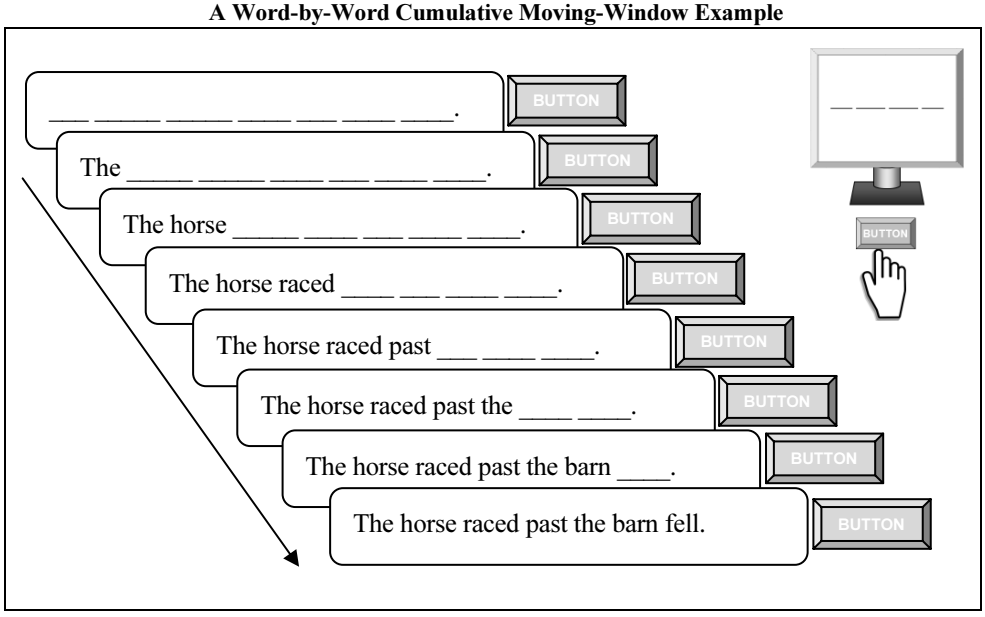
Success in L2 reading correlates with L2 proficiency. For the time factor in L2 reading, Browning (2003) argued that timed-reading is an effective exercise for learners to improve reading speed to become fluent L2 readers. In addition, from a psycholinguistic perspective on timed-reading, Juffs (2001) argued that the differences in reading times or response

times (RT) help educators to better understand learners' difficulties in L2 sentence processing. For nearly a century, RT data have been used to gain insight into mental processes and mental abilities in psychometric experiments (Gregory, 1996). Recently, RT measures have generally been used in the past dozen years in the L2 syntax acquisition of L2 syntax research using a self-paced reading paradigm. The self-paced reading paradigm (also called the "self-paced moving window" reading technique) has been frequently used in previous studies to examine native or L2 learners' real-time sentence processing.

In the self-paced moving window reading technique, each sentence appears in random order on a computer screen, one segment at a time. Only one segment is presented on the screen as the sentence is read. The participants control the reading speed by pressing a key to continue. When each successive segment appears, the previous segment disappears. Compared to the stationary-window technique, in which each segment overwrites the previous segment and stay in the center of the screen, the moving window paradigm demonstrates one segment at a time in the position it would normally have occupied in the passage. After the last segment appears, the participants are required to press a key to provide their responses either to comprehension questions or to grammatical judgment questions.

FIGURE 1

A Word-by-Word Cumulative Moving-Window Example



The moving-window technique has two versions: the cumulative moving-window and the non-cumulative moving-window versions (McDonough & Trofimovich, 2012). In the

cumulative moving-window, the subsequent segments accumulate with the previous one remaining on the screen. However, in the non-cumulative moving-window, the segment presents once and then covers up or removes the previous segment. The segments in the self-paced reading depend on the particular research question. Some studies present whole sentences as the segments, while other studies break down the sentence word by word (as illustrated in Figure 1) in measuring RT. The analysis of RT in the presentation of whole sentences was first employed with grammaticality judgments. In L2 research, grammaticality judgments have been used in order to gain insight into the learner's linguistic competence (White, 1987). In addition to off-line judgments, RT data functions as a supplementary measure to the learner's linguistic competence (Bley-Vroman & Masterson, 1989; Cook, 1990; White & Genesee, 1996; White & Juffs, 1998). On the other hand, some studies have used RT analysis in breaking down sentences into the word-by-word reading, providing insight into underlying competence (Fodor, 1998; Klein, 1998; White, 1987). In other words, RT analysis is an effective method to ascertain which segment the learner finds difficult while acquiring an L2 and reading L2 texts.

2.3. Linear Mixed Effects Model in L2 Research

Linear mixed-effects models have been extensively used in psycholinguistic research since the R package "lme4" was first developed in 2008. Mixed-effects models have been popular in psycholinguistic research because they have several advantages (Sonbul & Schmitt, 2013). They include both fixed-effects (independent variables) and random-effects. As random variables in one model, participants and items are treated as random variables, allowing for simultaneous generalization for results on new items and new participants (Gagné & Spalding, 2009). Mixed-effects models enable us to test main independent variables (item-related and participant-related). Mixed-effects models enable us to handle interval-scale (e.g., reaction time) measures through linear mixed-effects (LME; see Baayen, 2008) models and categorical (correct vs. incorrect) measures using mixed-logit models (the LME alternative for categorical data, see Jaeger, 2008). Lastly, mixed-effects models can deal with imbalanced designs and missing values. To include both correct and incorrect categorical measures in addition to an interval-scale RT measure and to test various item-related and participant-related factors, the mixed-effects analysis was used in this study.

To analyze the accuracy data, the method starts with the simplest (null) model, which includes only the dependent measure and the random variables. Fixed-effects are then added incrementally. The chi-square (χ^2) test is used to distinguish whether the inclusion of additional predictors contribute significantly to the model. Once the interim best-fit model is reached, variables are excluded one by one to check for any redundant predictor (i.e., a

present or missing variable in the model), leading to an insignificant difference.

To analyze the RT data, the mixed logit modeling is used, starting with a null model including the log RT (RTs were log-transformed to reduce skewness in the distribution) as a dependent variable, and participants and items as random effects. Predictor variables are then added: main effects (accuracy, function-word condition) and interaction effect (accuracy and function-word condition). Finally, variables are excluded one by one to check for redundancy to arrive at the final best-fit model.

Using this LME model, this study examined whether response accuracy and response time (RT) in the self-paced and timed cloze test can be explained by TOEIC scores, compared with the untimed cloze-test scores, and whether the types of target words (content or function words) affect the self-paced and timed cloze test performance in response accuracy and reading times. The research questions are addressed as follows:

1. Do response accuracy and response time (RT) in the self-paced and timed cloze test correlate with TOEIC scores, compared with the untimed cloze-test scores?
2. Do the types of target words (content or function words) affect the self-paced and timed cloze test performance in response accuracy and reading times?

3. METHOD

3.1. Participants

The participants in this experimental study were forty-five Korean college students (twenty-one males and twenty-four females) who were studying in Seoul, Korea. The data from two non-Korean students and three Korean students who did not have a TOEIC score were excluded from the analysis. Prior to the test, the participants signed an informed consent form to allow their data to be used for this study. Next, participants were asked to fill out the background information questionnaire consisting of seventeen items in five minutes. It asked participants to describe their biodata such as gender and age, and to evaluate their general English proficiency (vocabulary, comprehension, writing, speaking, and listening skills) with the rating scale from one to ten. The questionnaire also informed them of the purposes of the study.

The participants constituted a homogenous group with regard to age and language background. The participants ranged in age from nineteen to twenty-seven (mean: 22.6 years old). Thirty- five participants (87.5%) started to learn English from at least the first grade in public elementary school; most of the participants had taken at least ten years of formal English courses in elementary, middle, and high schools prior to this study. The

participants have taken eleven English-related courses on average in college.

3.2. Materials

Among the sub-types of cloze tests, a multiple-choice cloze test was chosen for this study, since it can help lower participants' test anxiety compared to a typical cloze test. In this study, every 7th word was deleted from a passage and participants were required to choose the correct answer from three choices given. The reading material for the cloze test was originally taken from *American Kernel Lessons: Advanced Students' Book* published by O' Neill, Cornelius and Washburn (1981). Ionin et al. (2013) used this text passage in their L2 studies and reported that the Cronbach α for the reliability of the test was .817. It contained three possible choices for each blank and participants were required to choose the most appropriate answer.

The correct answer types of this test were made up of two components: content words and function words. Forty-five percent of the test blanks were for function words (prepositions, pronouns, auxiliary verbs, conjunctions, determiners, relatives, pronouns, articles, and particles), while the remaining blanks were for content words (nouns, verbs, adjectives, and adverbs). For function words, the questions asked about prepositions, which Koreans usually make mistakes with. Considering that participants had no prior experience with the cloze test, the researchers gave directions for taking the test, and provided them with one or two sample tests, before the actual test was administered. The full text of the cloze test is presented in the Appendix.

In order to check the test items, a pilot study was conducted. No serious problems were found, but their responses led to minor revision such as spelling corrections. The average correct answers was twenty-six (65%) out of forty.

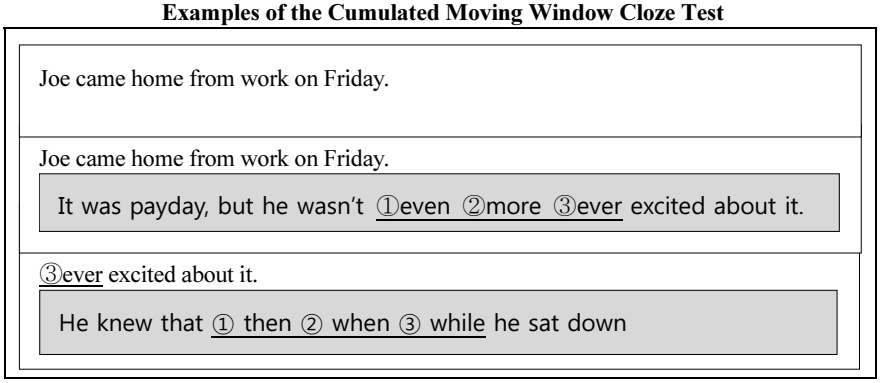
3.3. Design

In presentating the passages of the text, the self-paced moving-window reading technique was used in this study, specifically, the cumulative moving-window technique was used to help with comprehension. The reading passage was presented cumulatively as illustrated in Figure 2. Since some sentences contain more than five deleted blanks, the reading passage that they immediately read appeared again on the upper part of the screen. Participants were asked to read each question in a self-paced manner. They were asked to move on to the next question by pressing the "1", "2", or "3" button. Each question was given for a maximum of thirty seconds. If participants did respond to a question within the allotted thirty seconds, the screen automatically changed to the next question. Participants were not allowed to skip a question or go back to the previous question. Therefore, the test

could be completed within twenty minutes. Due to the unfamiliarity of this computerized program, a practice session was introduced before the main experiment. After this session, participants were asked to press the space bar to move onto the main experiment when they understood the procedure and felt ready.

FIGURE 2

Examples of the Cumulated Moving Window Cloze Test



The self-paced and timed condition was administered with E-prime software version 2.0 (Schneider, Eschman, & Zuccolotto, 2002), which was easy to use for a computerized experiment design, data collection, and analysis. The program also made it possible to obtain the participants' response time per question. This tool made it possible to determine how the time constraint on questions (thirty seconds) affects the test result. To measure accuracy as well as the response time (RT) for a correct answer, participants were also asked to respond as quickly as they could. In contrast, in an untimed control condition, participants were asked to answer all the forty questions printed on one sheet of paper within twenty minutes in total.

3.4. Procedure

The participants were randomly assigned to either the timed or untimed control condition. The participants in the former condition were asked to answer a question as quickly and correctly as possible and to take the test with the time constraint (maximum thirty seconds per question; after thirty seconds, it would automatically move on the next question), whereas the participants in the later condition were asked to complete the test within maximum twenty minutes. The participants in both conditions were not allowed to use an electronic dictionary while completing the test.

3.5. Scoring and Data Analyses

A correct response was coded as CORRECT. The exact word was worth one point for each item. A response was coded as INCORRECT if no response or the wrong response was given, for which zero was given. Illegible or multiple answers also were given no points. All descriptive data such as means and standard deviations were calculated. Independent variables were the experimental condition (the timed condition vs. untimed condition) and the target word group (content words vs. function words). As inferential statistics, a mixed log-effects modeling was used.

To run mixed-effects analyses, the open-source statistical software package R (R development core team, 2011) was used. To analyze accuracy data, the method started with the simplest (null) model and added fixed-effects variables (TOEIC, Condition, WORD TYPE) incrementally. The best-fit model is reported in the Results section. Likewise, to analyze the RT data, the mixed logit modeling also started with a null model including the log RT as a dependent variable and random-effects variables (PARTICIPANTS and ITEMS), adding fixed-effects variables (ACCURACY, TOEIC, WORD TYPE) and interaction effect (ACCURACY and WORD TYPE) incrementally. The final best-fit model is reported in the next section.

4. RESULTS

The participants in the timed condition marked 27.8 (69%) on average and the participants in the untimed control condition marked 29.9 (75%) on average. The best-fit linear mixed-effects model was selected, which only included the CONDITION factor, indicating that there was a marginally significant difference in the accuracy of the cloze test (t = 1.792, p = .073). The model including the factor TOEIC was not selected, demonstrating that TOEIC was not a significant predictor of the cloze test accuracy score.

On the other hand, as for RT data, a linear mixed-effects model including the factors TOEIC as well as ACCURACY and WORD TYPE was selected as the best-fit model in a mixed-logit model, as presented in Table 1. Table 1 illustrates that there was a significant difference in the RT data when the participants selected the correct and incorrect answers (t = -2.816, p < .01). The participants took a shorter time to judge correct answers (mean: 9.82 sec) than incorrect answers (mean: 11.73 sec). Second, a statistically significant difference was obtained in the WORD TYPE (t = -2.308, p < .05). This indicates that participants selected answers for function words faster than for content words. Finally, a marginally significant difference was found in TOEIC (t = -1.833, p = .067). This shows that participants's TOEIC scores were a marginally significant predictor in explaining the

RT data since the participants with higher TOEIC scores took a shorter time to choose the answer (10.26 sec) in the cloze test compared to those with lower TOEIC scores (11.06 sec).

TABLE 1

Summary of Mixed-Effects Model in Log-Transformed Reaction Times

Factor	Estimate Std.	Error	*t*-value	*p*-value
(Intercept)	10.184	.518	19.662	.005
ACCURACY	-.108	.038	-2.816	.005
WORD TYPE	-.188	.082	-2.308	.021
TOEIC	-.001	.001	-1.833	.067

Prior to the experiment, the participants self-assessed their L2 proficiency in the questionnaire. Most of the participants reported that they were proficient at reading and listening (6.6 and 6.5 out of 10) compared to other skills (vocabulary, writing, and speaking) and also weak at speaking (4.5 out of 10). In order to examine more detailed language proficiency levels, the participants in the timed and untimed control condition were further subdivided into two groups by a median split on TOEIC scores, HIGH and LOW according to their TOEIC scores, and were analyzed accordingly.

TABLE 2

Correlations Between Cloze Test Scores and Other Measures

			Self-Ratings					
Condition	Group	TOEIC	Vocabulary	Grammar	Comprehension	Writing	Speaking	Listening
Timed	HIGH	.791**	.373	.317	.509	.130	-.021	.730*
	LOW	.000	.487	.125	-.339	-.170	-.406	-.181
Control	HIGH	.272	.091	.107	.365	.444	-.149	.097
	LOW	-.217	-.198	.251	-.199	.143	-.447	-.235

* $p < .05$, ** $p < .01$

Table 2 displays correlation coefficients between cloze test accuracy scores and other measures such as TOEIC and sub-categories of self-ratings depending on the cloze test conditions. A significant correlation ($r = .791$, $p < .01$) was found between the timed cloze test scores and TOEIC scores in the HIGH proficiency level. There was also a significant correlation ($r = .730$, $p < .05$) between the cloze test and listening, but other significant correlations were not found with the cloze test.

In order to see which items were correctly answered and responded fast, we further carried out a qualitative item analysis (see APPENDIX for items). Most of the participants (98%) correctly answered the questions for content words such as *road* (#18 He drove into a quiet country ①*road* ②house ③air) and *better* (#19 The country sights made him feel ①as good ②*better* ③best) and function words such as *he* (#13 Finally, ①*he* ②she ③it

got into his car) and *on* (#31 a solar energy panel ①at ②out ③*on* the roof). In particular, all participants correctly answered the question about a content word, *driving* (#14 started ①drive ②driven ③*driving*).

However, the participants tended to incorrectly answer the questions for content words such as *past* (#20 His mind wandered as he drove ①*past* ②in ③to small farms) and *land* (#22 living on his own piece of ①house ②*land* ③farm and becoming self-sufficient), and a function word, *in* (#32 to heat the house ①*in* ②for ③over winter and power a water heater). Fifteen out of forty participants (37.5%) answered the question #20 correctly (36.4% in the timed condition; 30.4% in the untimed condition). On the other hand, only 25% (ten out of forty participants) answered the question #22 correctly; only 18.2% in the timed condition and 26.1% in the untimed condition marked a correct answer for the question #22. A percentage of choosing the correct answer for question #32 was 32.5% (thirteen out of forty participants). Only 30% in the time-constrained condition and 20% in the control condition marked a correct answer in the question #32.

Similar to the accuracy data as illustrated above, the consistent results of RTs in the item analysis were obtained. Most participants (mean: thirty-nine participants) answered the questions #13 (*he*), #14 (*driving*), #19 (*better*), and #31 (*on*) quickly (average RT: 8.41 sec), while few participants (mean: twelve participants) answered and took a longer time (average RT: 14.54 sec) in choosing the answers for the questions #22 (*land*) and #32 (*in*). Given that the average of all data was 10.67 seconds, participants spent a shorter time answering the questions in which most participants answered correctly. The question answered the fastest was #31 (*on*), and the average RT was 5.51 seconds. The question answered the slowest was #20 (*fast*), and the average RT was 17.45 (sec). In particular, participants spent more than fifteen seconds to choose *gas* (# 4), *past* (# 20), and *land* (# 22). Otherwise, the participants quickly selected the answers, *on* (# 31), *he* (# 33), *money* (# 38), and *out* (# 39) within seven seconds. As reported above, note that there was a significant difference in RTs between function words and contents words. For content words, participants took ten seconds (RT: 10.70 sec) on average, but nine seconds (RT: 9.64 sec) on average for function words. Therefore, for content words such as *gas* (# 4), *past* (# 20), and *land* (# 22), participants generally took a longer time judging them, and the average RT was higher than that for function words.

However, there were two questions (#27 and #35) worth paying attention to. First, the correct answer to the question #27 (his logical side was scoffing at his ①favorite ② practical ③*impractical* imaginings) was *impractical*. 75% of the participants in the untimed condition (fifteen participants) marked the correct answer, but in a timed condition, only 30% of them (six participants) marked the correct answer for the same question. Next, the correct answer for the question #35 (①Whether ②Even ③*If* the crops had a good yield) was *if*. 95% of the participants in the untimed condition marked the correct answer,

but only 60% of them in the timed condition did so. This discrepancy between the timed and untimed condtions might occur because it was hard for the participants in the timed condition to instantly incorporate the previous contents and apply them to answering the question (because it was technically impossible for them to go back to the previous contents).

5. DISCUSSION AND CONCLUSION

This study explored a self-paced and timed cloze test as an experimental research tool to evaluate English proficiency in second language (L2) experimental research by measuring accuracy and RT. The accuracy data did not exhibit any significance, but the RT results indicated that performance in the self-paced and timed cloze test was explained by several factors: ACCURACY, TOEIC, and WORD TYPE. That is, higher TOEIC scores correlated with greater response accuracy, and participants tended to respond more quickly to function words than to content words. Also, we investigated whether the cloze-test performance had a correlation with self-ratings for English proficiency and TOEIC scores. The results showed that TOEIC was significantly associated with cloze test scores in the high proficiency group. In particular, a significant correlation ($r = .791$, $p < .01$) was found between the cloze test and the TOEIC score from the HIGH group in the timed condition. However, no other significant correlations with the cloze test were observed in the untimed condition. Finally, a qualitative item analysis was additionally carried out, and its results were also consistent with the accuracy and RT results from the inferential statistics as illustrated above.

These results suggest that the participants' response time in the self-paced and timed cloze test can be useful as a proficiency measure; thus, it can be used as a supplemental measure for other experimental studies. In this respect, the results are also consistent with S.-Y. Lee's (2011) findings that the cloze test could be an economical and practical assessment tool for measuring overall language proficiency, and that the cloze test could be used as a substitute assessment tool for vocabulary tests, multiple-choice reading comprehension tests, and background knowledge tests.

This study has several limitations. First, the untimed condition was confounded with paper-based vs. computer-based test conditions. The participants in the untimed condition were given a pencil-and-paper test and were asked to complete the test within a maximum of twenty minutes, and they completed the test before the allowed twenty minutes. Since they were familiar with the pencil-and-paper test, their performance was marginally better than that in the timed condition. However, the TOEIC test administered in Korea is also a pencil-and-paper test. If the type of test (paper-based vs. computer-based) had really

mattered, a correlation between the untimed, pencil-and-pater cloze test and TOEIC scores would have been found; however, this was not the case. As another possible explanation for the marginal difference between the two conditions, there was one test question (#35 ①Whether ②Even ③*If* the crops had a good yield) worth noting. The correct answer for this questions was '*if*.' In a control condition, 95% of the participants (nineteen out of twenty participants) marked the correct answer, but in the time-constrained condition, only 60% of them (twelve participants) did so with the same question. This is probably because the participants were not allowed to see the next clause and to go back to the question. This issue is also relevant to the second limitation. That is, there were technical problems in the self-paced and timed cloze test. For example, the participants were not allowed to go back to previous questions they had already answered, even though they wanted to change the answers. In addition, they were not allowed to see the entire text, although the subsequent passages were presented cumulatively with the previous one remaining on the screen. Finally, this study did not consider whether the participants knew the cloze test or were trained to prepare for the time-constrained condition. If they had been familiar with the cloze test technique, they would have reduced their test anxiety and received higher scores.

Nonetheless, this study can be expanded to further research. Since one of the most crucial backgrounds of language learners is their L2 proficiency, as highlighted in studies by Dörnyei (2007) and Mackey and Gass (2012), participants' L2 proficiency should always be controlled for in experimental research. In this respect, the self-paced and timed cloze test can be used in evaluating English proficiency in L2 studies if participants cannot be accommodated with assessment tools such as TOEFL and TOEIC. The participants' mean RT can be used as a covariate or an independent variable in the data analysis. This economical and practical tool would lead to more solid experimental results by measureing L2 proficiency. Furthermore, this timed cloze test can be useful for educators or teachers to measure their students' proficiency in a classroom equipped with computers and to efficiently exploit it as a placement test or a review test.

REFERENCES

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

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press.

Babaii, E., & Ansary, H. (2001). The C-test: A valid operationalization of reduced redundancy principle? *System, 29*, 209-219.

Bachman, L. F. (1985). Performance on cloze test with fixed-ratio and rational deletions.

TESOL Quarterly, 19, 535-556.

Baker, B. A. (2011). Use of the cloze-elide task in high-stakes English proficiency testing. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 9*, 1-16.

Bernolet, S., Hartsuiker, R. J., & Pickering, M. J. (2013). From language-specific to shared syntactic representations: The influence of second language proficiency on syntactic sharing in bilinguals. *Cognition, 127*(3), 287-306.

Bley-Vroman, R., & Masterson. D. (1989). Reaction time as a supplement to grammaticality judgments in the investigation of second language learners' competence. *University of Hawai'i Working Papers in ESL, 8,* 207-237.

Browning, J. (2003). Why teachers should use timed reading in ESL classes. *The Internet TESL Journal*, *9*(6). Retrieved on February 4, 2014 from the World Wide Web: http://iteslj.org/Articles/Browning-Timed Reading.html.

Chapelle, C., & Abraham, R. G. (1990). Cloze method: What difference does it make? *Language Testing, 7*, 121-146.

Connelly, M. (1997). Using C-Test in English with postgraduate students. *English for Specific Purposes*, *16*, 139-150.

Cook, V. J. (1990). Timed comprehension of binding in advanced L2 learners of English. *Language Learning, 40*, 557-599.

Darwesh, A. J. A. (2010). Cloze tests: An integrative approach. *Journal of the College of Basic Education, 64*, 105-116.

Dörnyei, Z. (2007). Research methods in applied linguistics: Quantitative, qualitative and mixed methodologies. Oxford: Oxford University Press.

Dörnyei, Z., & Katona, L. (1992). Validation of the C-Test amongst Hungarian EFL learners. *Language Testing, 9*(2), 187-206.

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

Fodor, J. D. (1998). Triggers for parsing with. In E. C. Klein & G. Martohardjono (Eds.), *The development of second language grammars: A generative approach* (pp. 363-406). Philadelphia, PA: John Benjamins.

Gagné, C. L., & Spalding, T. L. (2009). Constituent integration during the processing of compound words: Does it involve the use of relational structures? *Journal of Memory and Language*, *60*, 20-35.

Gregory, R. J. (1996). *Psychological testing: History, principles and applications* (2nd ed.). Boston, MA: Allyn and Bacon.

Hinofotis, F. B., & Snow, B. G. (1980). An alternative cloze testing procedure: Multiple choice format. In J. W. Oller & K. Perkins (Eds.), *Research in language testing* (pp. 129-133). Rowley, MA: Newbury House.

Hosseini, F., Hassanzadeh, N., & Shayegh, K. (2012). A comparative study of the C-test

and the NC-test with Iranian EFL students. *Journal of Academic and Applied Studies, 2*, 8-31.

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.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434-446.

Juffs, A. (2001). Psycholinguistically oriented second language research. *Applied Linguistics, 21*, 207-220.

Kim, J., & Cho, Y. (2015). Proficiency effects on relative rules of vocabulary and grammar knowledge in second language reading. *English Teaching, 70*(1), 75-96.

Klein, E. (1998). Just parsing through. In E. C. Klein & G. Martohardjono (Eds.), *The development of second language grammar: A generative approach* (pp. 197-216). Philadelphia, PA: John Benjamins.

Klein-Braley, C. (1997). C-test in the context of reduced redundancy testing: An appraisal. *Language Testing, 14*(1), 47-84.

Klein-Braley, C., & Raatz, U. (1984). A survey of research on the C-test. *Language Testing 1*(2), 134-146.

Lee, E.-J. (2001). *The correlation between the cloze test scores and the vocabulary, grammar, background knowledge test scores in EFL situation*. Unpublished master's thesis, University of Sogang, Seoul, Korea.

Lee, J.-W. (2002). An analysis of test-taking strategies for the cloze. *English Teaching, 57*(1), 213-237.

Lee, J.-K. (2014). Self-assessment of second language proficiency as a research tool. *English Language and Linguistics, 20*(2), 125-137.

Lee, S.-Y. (1995). *Cloze test as an integrative measure of Korean students' English writing proficiency.* Unpublished doctoral dissertation, University of Texas, Austin.

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

Mackey, A., & Gass, S. (2012). (Eds.). *Research methods in second language acquisition: A practical guide*. Malden, MA: Wiley–Blackwell.

McDonough, K., & Trofimovich, P. (2012). How to use psycholinguistic methodologies for comprehension and production. In A. Mackey & S. M. Gass (Eds.), *Research methods in second language acquisition* (pp. 117-138). Oxford, UK: Wiley-Blackwell.

McGrath, J. (2009, June 5). *Do time constraints affect test takers?* Retrieved on February 4, 2014, from the World Wide Web: http://voices.yahoo.com/do-time-constraints-affect-test-takers-3356716.html.

McNamara, T. (2000). *Language testing.* Oxford, UK: Oxford University Press.

Oller, J. W. (1973). Cloze tests of second language proficiency and what they measure. *Language Learning, 23*(1), 105-118.

Oller, J. W. (1979). *Language tests at school: A pragmatic approach*. London: Longman.

Oller, J. W., & Inal, N. (1971). A cloze test of English prepositions. *TESOL Quarterly, 5*, 315-326.

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

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rashid, S. M. (2001, May). *Validating the C-test among Malay ESL learners*. Paper presented at the fifth Melta Biennial International Conference. Petaling Jaya, Malaysia.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime 2.0*. Pittsburgh, PA: Psychological Software Tools.

Shin, J.-A. (2010). Structural priming and L2 proficiency effects on bilingual syntactic processing in production. *Korean Journal of English Language and Linguistics, 10*(3), 499-518.

Shin, J.-A., & Christianson, K. (2012). Structural priming and second language learning. *Language Learning*, *62*(3), 931-964.

Sonbul, S., & Schmitt, N. (2013). Explicit and implicit lexical knowledge: Acquisition of collocations under different input conditions. *Language Learning, 63*(1), 121-159.

Stansfield, C., & Hansen, J. (1983). Field dependence-independence as a variable in second language cloze test performance. *TESOL Quarterly, 17*(1), 29-38.

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

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

Weir, G. (1990). *Communicative language testing*. Hemel Hempstead, UK: Prentice Hall.

White, L. (1987). Against comprehensible input: The input hypothesis and the development of second language competence. *Applied Linguistics, 8*, 95-110.

White, L., & Genesee, F. (1996). How native is near native? The issue of ultimate attainment in adult second language acquisition. *Second Language Research, 12*, 233-265.

White, L., & Juffs, A. (1998). Constraints on Wh-movement in two different contexts of non-native language acquisition: Competence and processing. In S. Flynn, G. Martohardjono, & W. O'Neill (Eds.), *The generative study of second language acquisition* (pp. 111-130). Hillsdale, NJ: Lawrence Erlbaum Associates.

APPENDIX

Joe came home from work on Friday. It was payday, but he wasn't (1) ①*even* ②more ③ever excited about it. He knew that (2) ①then ②*when* ③while he sat down and paid his (3) ①checks ②*bills* ③salary and set aside money for groceries, (4) ①driving ②pay ③*gas* for the car and a small (5) ①*deposit* ②withdrawal ③money in his savings account, there wasn't be (6) ①quite ②not ③*too* much left over for a good (7) ①pleasure ②leisure ③*life*. He thought about going out for (8) ①eat ②*dinner* ③eating at this favorite restaurant, but he (9) ①*just* ②only ③very wasn't in the mood. He wandered (10) ①*around* ②at ③in his apartment and ate a sandwich. (11) ①In ②*For* ③After a while, he couldn't stop himself (12) ①for ②*from* ③about worrying about the money situation. Finally, (13) ①*he* ②she ③it got into his car and started (14) ①drive ②driven ③*driving*. He didn't have a destination in (15) ①head ②*mind* ③fact, but he knew that he wanted (16) ①be ②*to be* ③being far away from the city (17) ①which ②there ③*where* he lived. He drove into a quiet country (18) ①*road* ②house ③air. The country sights made him feel (19) ①as good ②*better* ③best. His mind wandered as he drove (20) ①*past* ②in ③to small farms and he began to (21) ①try ②think ③*imagine* living on his own piece of (22) ①house ②*land* ③farm and becoming self-sufficient. It had always (23) ①being ②*been* ③be a dream of his, but he (24) ①having ②have ③*had* never done anything to make it (25) ①*a* ②one ③some reality. Even as he was thinking, (26) ①their ②*his* ③her logical side was scoffing at his (27) ①favorite ②practical ③*impractical* imaginings. He debated the advantages and (28) ①cons ②*disadvantages* ③problems of living in the country and (29) ①*growing* ②breeding ③building his own food. He imagined his (30) ①*farmhouse* ②truck ③tractor equipped with a solar energy panel (31) ①at ②out ③*on* the roof to heat the house (32) ①*in* ②for ③over winter and power a water heater. (33) ①She ②*He* ③They envisioned fields of vegetables for canning (34) ①either ②*and* ③but preserving to last through the winter. (35) ①Whether ②Even ③*If* the crops had a good yield, (36) ①*maybe* ②possible ③may he could sell the surplus and (37) ①store ②save ③*buy* some farming equipment with the extra (38) ①economy ②cost ③*money*. Suddenly, Joe stopped thinking and laughed (39) ①at ②*out* ③so loud, "I'm really going to go (40) ①*through* ②away ③in with this?"

Applicable levels: Tertiary

Eun-Young Chae
Department of Early Childhood English Education
Graduate School of Education
Kwangwoon University
20 Gwangwoon-ro Nowon-gu
Seoul 139-70, Korea
Email: eyc@kw.ac.kr

Jeong-Ah Shin
Department of English Language and Literature
Dongguk University
30 Pilfong-ro 1gil Jung-gu
Seoul 100-715
Email: jashin@dongguk.edu

Received in June 1
Reviewed in July 15
Revised version received in August 15