A Comparison of Constructed Response Formats as Measures of EFL Reading Comprehension*

Jeong-Won Lee
(Chungnam National University)


This study examined how four different constructed response formats of reading comprehension tests (cloze, recall, summary, and translation), developed based on the two texts with different reading difficulty, exert effects on EFL students’ reading test performance; and also investigated how those effects, if any, are differentiated with respect to the participants’ language proficiency levels. The total of 437 college students (220 male and 217 female students) participated in the current study. The test data was submitted to statistical treatment. The findings were that the four different response formats assessed different aspects of reading comprehension; the two texts with different text difficulty elicited different reading comprehension performance; and there were complicated interactions between the two variables of response format and text difficulty across the participants’ language proficiency levels. It is suggested that classroom teachers have to take it into consideration when planning to develop a reading comprehension test for EFL learners: how to choose reading texts that will be included in the test, considering an appropriate level of text difficulty and test formats to be engaged in order to assess students’ reading comprehension ability more accurately.

I. INTRODUCTION

Different test formats permit the assessment of different aspect of language ability (Alderson, 2000; Kintsch & Yarbrough, 1982; Kobayashi, 2002; Reder & Anderson, 1980; Shohamy & Inbar, 1991), especially when it comes to constructed response formats to assess EFL reading comprehension performance. According to Bernhardt’s (1991) constructivist model, reading comprehension is defined as a “constructive” and “active

---

* This study was financially supported by research fund of Chungnam National University in 2009.
process” (p. 191), and the process is an interplay between text-based components and reader-based ones. To measure such complicated, intriguing process precisely, it should be a prerequisite to identify the nature of measuring tools and then use them appropriately, since the tools of measuring reading comprehension influence and perhaps determine how readers perform on a reading comprehension test (Bernhardt, 1991; Lee, 1987; Shohamy, 1984; Wolf, 1993).

According to Bachman (1990), a variety of test facets can positively or negatively affect test-takers’ performance on the test tasks. In the area of assessment task type, he describes two such facets: the form of response necessary to complete a test item (verbal or non-verbal) and the format of response concerning how the test-taker responds to a test item. The response format can be classified into two: selected and constructed. In the selected response formats (e.g., multiple choice, true/false, matching, and checklists), the test-taker recognizes and chooses an appropriate response to the test item. The responses provide retrieval cues which can facilitate recall of passage information (Bransford, 1979). In the constructed response formats, the test-taker retrieves appropriate information from the text and produces responses themselves. Since the test-taker must generate his own retrieval cues, recall becomes more difficult (Bransford, 1979). Such tasks that require constructed responses from the test-taker vary from highly structured ones that elicit one-word answers or short phrases (e.g., open-ended questions and cloze tasks), to unstructured ones (e.g., summary, recall protocol, and translation).

Such variety in constructed response formats to assess reading comprehension ability may well exercise influence on assessment tasks in different ways through the interaction between the reader and the text. However, the issue of assessing reading by employing constructed response formats has not drawn significant attention of researchers and classroom teachers fully to try out these test types in the assessment of students’ reading comprehension performance.

One reason may have a definite bearing on that the multiple-choice format prevails in order to measure test-takers’ reading comprehension ability, in favor of reliability and practical convenience at the expense of validity. Despite its world-wide popularity as a test format for assessing reading in ESL/EFL contexts, however, the multiple-choice test type has its critical drawbacks in that test takers can guess the right answer without fully understanding the reading text (Nevo, 1989; Royer, 1990; Weir, 1993), and in that they can answer test items without reading the text (Johns, 1978; Perkins & Jones, 1985), mostly resorting to test-taking strategies (Cohen, 1984; Swaffar & Wältermann, 1988). Another can be found in the fact that testers, EFL teachers in particular, are not much aware of how differently the different formats measure test-takers’ reading performance—the formats that they try to develop and implement in their testing situation—though recent years have witnessed an increase in the number of those test formats used for assessing reading.
A couple of researchers have paid attention to the issue of the effects of different test formats on tests of reading performance in terms of the relationship between assessment tasks (multiple-choice, open-ended, cloze, recall) and language of assessment (Wolf, 1993), the effects of different test methods (recall and summary) on the readers’ performance (Riley & Lee, 1996), and the relationship between response formats (cloze, open-ended, summary) and text organization (Kobayashi, 2002). Still, there have been relatively few research studies that compared the four different constructed response formats (cloze, recall, summary, and translation) of reading comprehension tests so as to identify what and how they measure. It is timely, therefore, to make an exploratory attempt to gain a clear understanding of the relationship between those test formats and test performance in reading.

A collateral variable in the current study is the choice of text in reading comprehension tests that influences on test performance (Alderson, 2000; Shohamy, 1984). Obviously, this might be due to greater or lesser readability of the texts used. It would be worth investigating how readability of text in the four different constructed response formats of reading comprehension tests affects test performance.

The purposes of the current study are, therefore, first to examine how constructed response formats of reading tests, developed based on texts with different reading difficulty, exert effects on reading test performance, and second to investigate how those effects, if any, are differentiated with respect to the participants’ language proficiency levels. To fulfill the purposes of the study, the following questions were formulated:

- Do response formats of the English reading comprehension test and reading text difficulty affect the participants’ reading comprehension test performance? If so, how?
- Do response formats of the English reading comprehension test and reading text difficulty affect the participants’ reading comprehension test performance depending upon their language proficiency levels? If so, how?

II. ASSESSMENT FORMATS

Four constructed response formats of reading comprehension tests will be included in the present study: cloze, recall, summary, and translation. These formats are conceptualized as well-known representative constructed response formats for assessing reading formally.

Unlike the other constructed response formats of assessment tasks such as recall, summary, and translation, the cloze passage serves as the reading text, and test-takers read
a mutilated passage completing a cloze task (Wolf, 1993). It makes the cloze procedure operating incongruously with the normal course of reading. This has invited a great deal of research on the validity of the cloze test as a reading comprehension test. Some researchers (Alderson, 2000; Hughes, 2003; Kintsch & Yarbrough, 1982; Koda, 2005; Lado, 1986; Markham, 1985) claim that the cloze in general tackles more upon surface linguistic forms (e.g., grammatical and lexical knowledge) than upon overall reading comprehension because many cloze items are based more on constituents from the immediate environment around the blank rather than on discourse-level information from the whole text. Moreover, the same reader’s test performance may differ considerably across alternative test versions developed from the same text (Alderson, 2000; Hughes, 2003). The deviation clearly casts serious doubts on its construct validity. Others (Bachman, 1982; Clarke, 1980, 1981; Jonz, 1991; Nunan, 1985) contend, however, that the cloze is a valid measure of reading comprehension since it can assess learners’ sensitivity to semantic and syntactic constraints in local or sometimes wider context, and thus the sensitivity can be taken as a reliable gauge to measure reading ability. No matter how hot the debate is, there is general agreement that because of its relative ease in test construction, administration, and scoring, the cloze is widely used to assess reading.

In a recall test, a widely used method of assessing both L1 and L2 reading comprehension (Lee, 1990; Riley & Lee, 1996), learners are asked to read a text, and then write down everything they can remember from the text without referring back to the text. The virtue of the task lies in that it assesses reader-text interaction in the most straightforward manner “without procedural contamination from additional task requirements” such as test questions (Koda, 2005, p. 237). In other words, it is claimed to provide an overall picture of learners’ comprehension processes as a purer measure of comprehension: which information is stored and organized, and how; which retrieval strategies are exploited; and how the text message is reconstructed (Bernhardt, 1983, 1991). However, the recall protocols present problems: It resorts strongly to memory; it offers little information on what is not recalled; and the scoring procedure is so demanding and labor-intensive that its practical use is limited for most assessment purposes (Alderson, 2000; Bernhardt, 1991; Dole, Brown, & Trathen, 1996; Meyer, 1975).

As a variant of a recall test, a summary is believed to measure authentic overall comprehension of the reading text in the era of communicative language testing, since it tries to simulate real-world tasks in which readers are asked to read a text and write a summary of its main ideas (Cohen, 1993). Besides, it also requires learners to judge relevant ideas from irrelevant ones and to organize their thoughts about the ideas in order to construct a satisfactory and appropriate summary (Alderson, 2000). In other words, the merits of summarization are its close approximation of real-life activities, and simple construction and administration. According to Kozminsky and Graetz (1986), furthermore,
a summary is appropriate particularly with university students who have to read academic materials for comprehension requirements. However, the major disadvantages lie in scoring subjectivity and additional processing requirements such as higher level of writing proficiency (Alderson, 2000; Koda, 2005).

The nature of translation is frequently misunderstood with its function in learning and testing not clearly identified in detail. Given that translation is strongly tied with the learning process of the grammar-translation method, it has long been recommended not to be used as a language learning device in the foreign language learning classroom, based on the argument that translation interferes an intermediate process of comprehension, and thus impedes learners to think directly in the foreign language. It can be argued, however, that even when they are taught by any forms of direct methods, they virtually interpose this intermediate step themselves, especially older students (Cordero, 1984).

While translation, as an L2 writing strategy, has been researched mostly to verify textual differences between L1 and L2 writing (Hedge, 2000; Uzawa, 1996), it has also been conceived as a learning tool or a convenient means of investigating comprehension and accuracy (Cordero, 1984). Comparing recall protocol and translation to investigate the impact of memory required in the recall task on EFL reading comprehension, Chang (2006) addressed feasibility of translation as a measure of reading comprehension that may provide a clearer, more complete picture about readers’ understanding of the text. Moreover, translation can provide an on-line tool for assessing comprehension in the foreign language (Donin & Silva, 1993; Hoover, Deffner, & Ericsson, 1989).

III. RESEARCH DESIGN

1. Participants

For this study, 437 students from a college were involved: 220 (50.3%) males and 217 (49.7%) females. They were taking one of the elective English reading courses offered by the college when asked to participate in the experiment. Among them, 77 students (17.6%) reported that they have experienced studying English abroad from 2 months to 2 years while most of the students (360: 82.4%) have not. The students in intact English language classes were randomly divided into 8 groups in total, and students in each group was tested by one of the four reading comprehension tests in different response formats—cloze, recall, summary, and translation—developed based on the two texts with different text difficulty. For further analysis, they were divided into three groups depending upon their English proficiency based on their TOEIC scores: 138 (31.6%) high-level, 178 (40.7%) mid-level, and 121 (27.7%) low-level students.
2. Materials

1) English Proficiency Test

The comparability of the eight groups of different response formats and text difficulty was confirmed using the participants’ TOEIC scores. No statistically significant difference was found among the eight groups in their English language proficiency ($F = 1.022, p = .415$). The scores were also used to divide the students into three different English proficiency levels in accordance with the rank order of their scores: high- (above 700), mid- (500-699), and low-level students (below 499).

2) Reading Texts

Two expository texts were used for the present study. For each text, the topic was chosen with care to minimize the potential effects of cultural bias or student familiarity with it (Alderson & Urquhart, 1985, 1988; Clapham, 1996; Hale, 1988). The two topics were “international aid” (called “Aid” henceforth) from Kobayashi (2002) and “college students’ writing ability” (called “Writing” henceforth) from Carrell (1991) respectively (see Appendices 1 and 2). The lengths of the two texts were similar (355 words for the Aid text and 341 words for the Writing text respectively). Text difficulty was determined by the Flesch Reading Ease Formula, one of the most widely accepted readability indices. The readability score for the Aid text was 63.011 while that for the Writing text was 35.435, which signifies that the latter is much more difficult than the former.

3) Reading Comprehension tests

For the two texts, test items were developed in the four different response formats: cloze, recall, summary, and translation. As for cloze, fixed-ratio deletion was adopted with every 13th word deleted, which resulted in 25 blanks in total. The deletion type is believed to be comparable, and equally reliable and valid with rational deletion (Bachman, 1985; Jonz, 1990). Great caution was taken not to delete proper nouns and numbers. For both texts, 25 blanks were made. The cloze test results were scored based on exact word keys. The exact word method was engaged in the present study in favor of practicality and reliability since most EFL teachers would choose this method rather than the appropriate word scoring in order to save their time and effort to determine whether each response is indeed appropriate, although the appropriate word method has been mostly advised and used to score the cloze test results particularly in terms of face validity of the test (Brown, 2004).

In the recall test, students were asked first to read the given text, then to turn the test
sheet over for them not to see the text while writing down information recalled from the
text, and finally to write down on it everything they can remember from the text. In order
to get the comprehension score in the recalls, idea units were counted in the original text.
Though it is rather hard to define idea units (Alderson, 2000), in this article the
propositions in the text based on pausal units (or breath groups) were analyzed (Johnson,
1970). Two native American English instructors made division of the units respectively,
and then they decided an agreed division for each text (53 units for the Aid text and 37
units for the Writing text). With these idea units in each text, the researcher and an
experienced Korean English instructor scored the students’ recalls.

In the summary test, participants were required to read the given text and summarize
main ideas in full sentences in a half page in length. To score summaries, the two native
American English instructors wrote their own summaries of the texts, and decided the
main ideas to be included in the scoring rubric, comparing their summaries each other.
Eight and ten propositions were extracted from the summary of the Aid and Writing texts
respectively. The students’ summaries were scored by the researcher and the English
instructor based on the propositions.

To construct the translation test, the sentences including main ideas were underlined on
each text. Such sentences were chosen based on the summary rubric that the two native
English instructors agreed upon. Participants were asked to read the given text and translate the underlined sentences word for word in their mother tongue. The test was scored by the researcher and the Korean English instructor.

The participants were asked to respond in Korean, their mother tongue, to the three test
formats except the cloze format in order to eliminate any undesirable effects of the use of
English on representing their reading performance – so as not to make them the tests of
production (writing) as well as comprehension (reading) (Alderson, 2000; Chang, 2006;
Donin & Silva, 1993; Kobayashi, 2002; Lee, 1986; Wolf, 1993). Upon scoring, one point
was given for each unit (for the recall test) or proposition (for the summary and translation
tests) with no partial point in order to secure scoring accuracy and consistency. The
distribution of the eight groups of participants in accordance with the test format and text
difficulty is shown in Table 1.

### TABLE 1
The Distribution of the Eight Groups of Participants

<table>
<thead>
<tr>
<th></th>
<th>Cloze</th>
<th>Recall</th>
<th>Summary</th>
<th>Translation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aid (easy)</td>
<td>53</td>
<td>65</td>
<td>65</td>
<td>47</td>
<td>230</td>
</tr>
<tr>
<td>Writing (difficult)</td>
<td>57</td>
<td>41</td>
<td>58</td>
<td>51</td>
<td>207</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>110</td>
<td>106</td>
<td>123</td>
<td>98</td>
<td>437</td>
</tr>
</tbody>
</table>

Along with the four formats of assessment tasks, a simple questionnaire was given to
the participants to check background information such as year in college, gender, TOEIC score, and study/stay-abroad experiences in any English-speaking countries. The participants were asked to respond the questionnaire before taking the given reading comprehension test. The four formats of the reading comprehension tests were performed for two month in 2009.

3. Data Analyses

The test results were coded and analyzed using the statistical package of the SPSS 18.0 version. The mean scores were converted into percentages. For the tests, descriptive statistics were calculated, and ANOVAs (both one-way and two-way) and t-tests were conducted to examine the research questions.

So as to confirm the reliability of marking, the test papers were scored independently by the researcher and the experienced English instructor. The inter-rater reliability between the two raters on the two texts of each test format except the cloze was high enough for further analyses: $\alpha = .98$ for the recall test; $\alpha = .94$ for the summary test; and $\alpha = .98$ for the translation test.

IV. RESULTS AND DISCUSSION

1. Relationship between Test Performance, Text Difficulty, and Response Formats of the Test

The participants performed in the reading tests differently across the four types of response format and the two types of text difficulty. As shown in Table 2, the scores were lowest in the cloze test, followed by the recall and summary tests, and highest in the translation test. It signifies that the cloze test was the most difficult one for the students regardless of text difficulty, probably because the test results were scored by the exact word method and because the students had to come up with the exact English words to fill in the blanks, unlike other tests that required them to think and write in their mother tongue. The recall test was the second most difficult one conceivably because it asks students to construct their own ideas about the text in a more specific and detailed fashion, while the summary test, the third ranked, measures a more comprehensive, overall understanding of the text. The translation test was the easiest one in that they are provided with English prompt based on which they can work on corresponding translation in their mother tongue.

In addition, the Writing text (the difficult one) was significantly more difficult than the Aid text (the easy one) in the recall test ($t = 7.982, p = .000$) and in the translation test ($t =
3.045, \( p = .003 \) with the moderate effect sizes of .59 and .30 respectively (Cohen, 1988). It should be because the two types of test format require the test-takers’ detailed and thorough understanding of the reading passages, linguistic knowledge, and memory, to some extent. It may suggest that ESL/EFL readers’ linguistic proficiency matters when their reading comprehension ability is assessed through the recall and/or translation test formats (Kintsch, 1998).

### TABLE 2

<table>
<thead>
<tr>
<th></th>
<th>Aid (easy) M (SD)</th>
<th>Writing (difficult) M (SD)</th>
<th>Total M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloze</td>
<td>28.6 (12.6) (N = 53)</td>
<td>28.4 (12.6) (N = 57)</td>
<td>28.5 (12.6) (N = 110)</td>
</tr>
<tr>
<td>Recall</td>
<td>52.3 (28.3) (N = 65)</td>
<td>31.4 (17.0) (N = 41)</td>
<td>46.4 (25.1) (N = 106)</td>
</tr>
<tr>
<td>Summary</td>
<td>58.0 (25.6) (N = 65)</td>
<td>54.1 (23.8) (N = 58)</td>
<td>56.2 (24.8) (N = 123)</td>
</tr>
<tr>
<td>Translation</td>
<td>77.6 (27.9) (N = 47)</td>
<td>62.7 (22.6) (N = 51)</td>
<td>69.9 (25.1) (N = 98)</td>
</tr>
</tbody>
</table>

The first research question was examined using the two ANOVAs: one about the relationship between the students’ reading test performance and the different four response formats of the reading comprehension test, and the other about the relationship between their reading test performance, response formats, and text difficulty.

The first ANOVA revealed that the scores gained from the four response formats were significantly different (\( F = 62.799, p = .000 \)) with the very large effect size of .77. In addition, according to the post-hoc results, the differences were statistically significant in all the combinations of the response formats (recall-summary: \( p = .006 \); the others: \( p = .000 \)). In other words, the mean differences found among the four test formats in Table 2 were statistically confirmed that the participants of the current study performed significantly differently across the four different test formats. It means that the different test formats may assess different aspects of reading comprehension ability in the EFL setting (Alderson, 2000; Kobayashi, 2002), although the test formats were developed based on the same text and quite a similar scoring rubric for the three formats except the cloze procedure.

In the second ANOVA, the two-way interaction (text difficulty * response formats) proved to be statistically significant (\( F = 10.714, p = .000 \)), along with the two main effects (\( F = 74.692, p = .000 \) for response format; \( F = 36.199, p = .000 \) for text difficulty). This means that the effects of test response format on EFL students’ reading test performance should be interpreted with respect to text difficulty because of the interaction between the two variables as shown in Figure 1. In other words, it can be argued that reading test performance is affected by complicated interaction between text difficulty and response format.
FIGURE 1
Effects of Text Difficulty and Response Format on Reading Test Performance

2. Relationship between Test Performance, Text Difficulty, and Response Formats of the Test Depending upon Language Proficiency Levels

First, text difficulty effects for the three language proficiency levels were examined with the independent t-tests as in Table 3. Interestingly, while no significant effect of text difficulty was found on the three response formats by students’ language proficiency, the results of the recall test only were statistically significant between the two texts irrespective of students’ language proficiency with the moderate effect size of .49 for high, .69 for mid, and .52 for low. This shows that the scores of the recall test based on the Aid text were significantly higher than those on the Writing text in all proficiency levels. This may imply that the recall test format measures different aspects of reading comprehension from the others in that it is typically associated with memory along with comprehension of the text (Alderson, 2000).

| Table 3: Text Difficulty Effects by Response Formats and Proficiency Levels |
|---------------|-----|-----|-----|
|               | High| Mid | Low |
| Cloze         | -1.519 (.139) | 1.456 (.152) | 0.136 (.893) |
| Recall        | 2.931 (.007)  | 6.646 (.000)  | 3.669 (.001)  |
| Summary       | 0.480 (.634)  | 0.688 (.495)  | 0.244 (.809)  |
| Translation   | 1.767 (.086)  | 0.898 (.384)  | 1.694 (.108)  |
| Total         | 2.148 (.033)  | 3.410 (.001)  | 2.080 (.040)  |

Second, Table 4 reveals the results of one-way ANOVAs to check response format effects on text difficulty depending upon students’ language proficiency levels. In all three language proficiency levels, the F values were significant with mostly high effect sizes regardless of text difficulty. This means that students in each level performed differently in the different test formats when dealing with texts with different text difficulty.
A Comparison of Constructed Response Formats as Measures of EFL Reading Comprehension

### TABLE 4
Response Format Effects by Proficiency Levels

<table>
<thead>
<tr>
<th>Format</th>
<th>High (F(p))</th>
<th>Effect Size</th>
<th>Mid (F(p))</th>
<th>Effect Size</th>
<th>Low (F(p))</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aid (easy)</td>
<td>17.450 (.000)</td>
<td>.97</td>
<td>16.047 (.000)</td>
<td>1.25</td>
<td>6.480 (.001)</td>
<td>.67</td>
</tr>
<tr>
<td>Writing (difficult)</td>
<td>9.393 (.000)</td>
<td>.83</td>
<td>47.644 (.000)</td>
<td>1.49</td>
<td>9.138 (.000)</td>
<td>.80</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>24.196 (.000)</td>
<td>.84</td>
<td>35.627 (.000)</td>
<td>.94</td>
<td>10.596 (.000)</td>
<td>.58</td>
</tr>
</tbody>
</table>

According to the post-hoc results as in Table 5, in the Aid text, the high group showed a bit different performance from the other two groups in that the former revealed more various significant differences in their scores across the four response formats than the latter who performed in the same fashion that the cloze was significantly harder to respond than the other three formats. It is interesting that the high group performed best in the translation format that requires overall comprehension ability as well as specific linguistic knowledge, while, for the mid and low groups, it was not significantly easier than recall and summary conceivably because they are not equipped enough with linguistic knowledge although the text was easy. This finding appears to echo the argument made elsewhere in the study about the threshold level hypothesis for EFL readers (Lee & Schallert, 1997).

### TABLE 5
Post-hoc Results of Response Formats by Proficiency Levels

<table>
<thead>
<tr>
<th>Format</th>
<th>High</th>
<th>Mid</th>
<th>Low</th>
</tr>
</thead>
</table>

* C: cloze; R: recall; S: summary; T: translation

** The slash denotes statistically significant group division.

The similar pattern in the group difference emerged in the Writing text as in the Aid text that the high group performed somewhat differently from the mid and low groups. One of the outstanding differences between the high group and the other two groups is where the summary belongs: For the high group, the cloze and recall were significantly difficult than the translation, with the summary indecisive in its comparative difficulty, whereas the mid and low groups performed clearly better in the summary and translation than in the cloze and recall. The more difficult the reading text is, the less distinguishable the difference between high and low level learners’ test performance is.

The comparison of the results in terms of text difficulty indicates that the recall format, which asks test-takers to concentrate on overall understanding while memorizing specific information, was as difficult as the cloze test in the Writing text, but not in the Aid text, regardless of the participants’ language proficiency levels. It can be safely said that there should be a close relationship between text difficulty and performance in the recall test.
Third, the post hoc results were submitted to scrutiny in order to glean more detailed information on the effects of the response formats on text difficulty depending upon students’ language proficiency level. With regard to the Aid text, the easy one, as shown in Table 6, there was little variation across the language proficiency levels in each response format. In other words, there was no statistically significant mean differences in students’ reading comprehension test performance depending upon their language proficiency levels across the four different response formats, although the low group was scored lower than the other two groups overall. Such results represent that the participants of the current study reveal a similar fashion of reading test performance when tackling the easy reading text across the four different response formats, no matter what language proficiency level they are in. That is, the easy text ("Aid") is not likely to be conducive to differences of reading performance in accordance with the students’ language proficiency levels.

<table>
<thead>
<tr>
<th></th>
<th>High M (SD)</th>
<th>Mid M (SD)</th>
<th>Low M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloze</td>
<td>30.3 (14.0)</td>
<td>30.9 (13.0)</td>
<td>24.0 (10.0)</td>
</tr>
<tr>
<td>Recall</td>
<td>59.3 (20.3)</td>
<td>61.8 (21.4)</td>
<td>50.3 (26.7)</td>
</tr>
<tr>
<td>Summary</td>
<td>59.3 (26.8)</td>
<td>63.0 (22.0)</td>
<td>48.7 (17.9)</td>
</tr>
<tr>
<td>Translation</td>
<td>85.6 (27.2)</td>
<td>77.6 (28.3)</td>
<td>63.0 (29.0)</td>
</tr>
</tbody>
</table>

When it comes to the Writing text (the difficult one), the two response formats (cloze and translation) affected the students’ reading test performance according to their language proficiency levels (see Table 7 and Figure 2). In the cloze format, which was proven as the most difficult test format overall, the high group performed significantly differently better than the other two groups ($F = 7.196, p = .002$) with the moderately high effect size of .66. This signifies that higher linguistic ability is required to perform better in the cloze test developed based on the difficult text, since the test is often claimed to measure general language proficiency more than reading ability (Alderson, 2000).

On the other hand, in the translation test, the scores of the low group were significantly differently lower than the other two groups in their test performance ($F = 13.960, p = .000$) with the very high effect size of .91. Although the translation test was proven as the easiest one overall in the present study, it still asks the students to understand the text passage in a more specific fashion especially in terms of exploiting linguistic information such as vocabulary and grammar (Chang, 2006; Cordero, 1984). These findings seem to be somewhat indirectly conformed to the threshold hypothesis that “before L2 readers can transfer L1 reading skills (strategies) and background knowledge to improve comprehension, they must reach a threshold level of language proficiency” (Lee &
Schallert, 1997, p. 717). In addition, the findings also imply that it is critical to take text difficulty of the reading passages into consideration when the cloze and translation tests are to be used in order to assess students’ reading comprehension in EFL settings.

### TABLE 7
Reading Test Performance in the Writing Text by Proficiency Levels

<table>
<thead>
<tr>
<th></th>
<th>High M (SD)</th>
<th>Mid M (SD)</th>
<th>Low M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloze</td>
<td>25.8 (10.8)</td>
<td>22.9 (11.7)</td>
<td>23.5 (10.8)</td>
</tr>
<tr>
<td>Recall</td>
<td>25.9 (13.5)</td>
<td>22.3 (11.7)</td>
<td>22.9 (11.7)</td>
</tr>
<tr>
<td>Summary</td>
<td>58.6 (22.3)</td>
<td>56.5 (22.3)</td>
<td>58.6 (22.3)</td>
</tr>
<tr>
<td>Translation</td>
<td>70.1 (11.5)</td>
<td>70.1 (11.5)</td>
<td>46.9 (18.3)</td>
</tr>
</tbody>
</table>

### FIGURE 2
Reading Test Performance in the Writing Text by Proficiency Levels

Table 8 shows the results of two-way ANOVAs that explore the effects of response format and text difficulty in accordance with the participants’ language proficiency levels. The F value of the mid group was higher in both effects of response format and text difficulty than the other two groups. In other words, the two variables exerted greater effects on students with mid-level proficiency. The findings clearly show that, across the language proficiency levels, response format and text difficulty respectively and their combination influence students’ achievement in reading comprehension, mid-level students’ achievement in particular.

### TABLE 8
Relationship Between Text Difficulty and Response Format by Proficiency Levels

<table>
<thead>
<tr>
<th></th>
<th>High F (p)</th>
<th>Mid F (p)</th>
<th>Low F (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main effect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response format</td>
<td>23.900 (.000)</td>
<td>45.691 (.000)</td>
<td>12.548 (.000)</td>
</tr>
<tr>
<td>Text difficulty</td>
<td>4.920 (.028)</td>
<td>21.390 (.000)</td>
<td>8.655 (.004)</td>
</tr>
<tr>
<td>Interaction effect</td>
<td>2.691 (.049)</td>
<td>7.887 (.000)</td>
<td>2.513 (.062)</td>
</tr>
</tbody>
</table>
V. CONCLUSION

The current study was aimed to examine how four different constructed response formats of reading comprehension tests, developed based on texts with different reading difficulty, exert effects on reading test performance of 437 college students, and to investigate how those effects, if any, are differentiated with respect to their language proficiency levels.

As for the first research issue, there was relationship between the students’ reading comprehension test performance, text difficulty, and the four response formats of the test to some extent. The students performed in the reading comprehension tests differently across the four different formats of response format and the two types of text difficulty. In other words, the test tasks with different response formats exerted influence on the readers’ performance on the test of reading comprehension (Bernhardt, 1991; Lee, 1987; Shohamy, 1984; Wolf, 1993). The second research issue was also positively confirmed that the students revealed complicated reading test performance in different test formats with different text difficulty according to their language proficiency levels: The high level group performed differently from the other two groups across the formats; significant difference of test performance according to the language proficiency levels was found with the difficult text; and there was complicated interaction between the two variables (response formats and text difficulty) in the linguistic levels.

With the findings of the present study, it can be clearly articulated that different response formats assess different aspects of reading comprehension (Alderson, 2000; Kobayashi, 2002), that different text difficulty elicits different reading comprehension performance, and that there are complicated interactions between the two variables of response format and text difficulty across the students’ language proficiency levels. Such results suggest that classroom teachers (or testers) have to take it into consideration how to choose texts with an appropriate level of difficulty and which test formats to be engaged so as to measure students’ reading comprehension ability more accurately.

Lastly, this study was an exploratory attempt to figure out any influence of different test formats with different text difficulty on EFL learners’ reading test performance. It was performed rather with limited number of students, only two sample of reading texts, and four constructed response formats of reading tests which are quite different in their nature and function. Accordingly, it would be recommended to carry out similar studies with more extended as well as more various factors of the participants, test formats, text types, text organization, etc. In so doing, future research is needed to identify the nature and function of different constructed response formats of reading comprehension tests and to suggest any ideas to bridge the gap between testing practice and language learning/teaching in classroom contexts.
REFERENCES


**APPENDIX A**

International Aid

The industrialized countries between them possess 78% of all existing wealth in the world. This means that the other countries, which are usually called the 'Third World,' have
about 22%, even though their population is about 76% of the world's total. Many rich industrialized countries give aid to poorer Third World countries. The intention is simple—giving aid in this way should help the poorer countries to improve their situation. Of course, they hope that aid will no longer be necessary in the end, since the Third World countries will have become able to look after themselves.

However, many people argue that much of the aid given to Third World countries does more harm than good. One example of this is 'tied aid.' Money or machinery is given to a Third World country, but on certain conditions. These usually mean, for example, that the receiving country has to spend the money on what is produced in the giving country. As a result, the Third World country may have to buy products it does not need, or at a higher price.

At the same time, Third World countries become dependent on industrialized countries. They need them more and more. For example, a Third World country may be given expensive tractors. Agricultural productivity may improve enormously, but when the tractors go wrong, they will require skilled mechanics or expensive spare parts. Either way, the poor country needs to pay money to the richer country to repair the tractors.

Moreover, most aid has been used in cities. This makes life there look more attractive, offering jobs which are highly paid and which are not available in rural areas. So people leave the countryside and move to cities. As a result, the countryside becomes empty and the country can no longer produce enough food for its people. At the same time, cities become overcrowded and there are all sorts of problems, from housing shortages to poor health facilities. Worse still, there may not be enough jobs for all the people who come to the cities hoping that they will become richer: Many of them, in fact, become poorer than before.

**APPENDIX B**
College Students' Writing Ability

Cries of dismay are heard in the halls of American business and industry, where employers charge that American colleges are graduating students who have failed to master effective written communication.

There are several reasons why graduates of even America's best colleges cannot write well. First, freshmen are entering American colleges without the proper previous education. Colleges complain that even the brightest freshmen, in some ways more sophisticated than ever before in history, are seriously deficient when it comes to organizing their thoughts on paper. By the time they reach college, the professors complain, it is almost too late to overcome years of inadequate study of the basics of grammar and style which high schools
no longer require of students. Second, students today do not read as widely as they should in order to learn to write clearly. Educators are convinced that a student who does not read widely and who cannot read with true comprehension will never learn to write well. Finally, these deficiencies which college students bring with them into American colleges cannot be overcome in only one year of required freshmen composition.

In order to stop the alarming decline of writing skills of the next generation of American college graduates, a number of solutions have been proposed. These include placing greater emphasis on the secondary school curriculum to ensure that the basics of effective written communication are taught, and assigning more reading in both elementary and secondary schools. At the college level emphasis needs to be given to requiring student writing not only in one year of freshmen composition, but also in courses for all four undergraduate years. Many colleges have established programs in 'writing across the curriculum' to call attention to the importance of students mastering writing in the content areas.

Is American television to blame for why Johnny can't write? Although this is a hotly debated point, it is really irrelevant. Whether it is or isn't, the important point is that school systems from elementary through college must refocus attention on the teaching of the written language.

Applicable levels: secondary and tertiary levels
Key words: measurement of EFL reading comprehension, constructed response tasks of assessment

Jeong-Won Lee
Dept. of English Education
College of Education
Chungnam National University
99 Daehak-ro, Yuseong-gu
Daejeon 305-764, Korea
Tel: (042) 821-5333
Fax: (042) 821-8730
Email: jeongwon@cnu.ac.kr

Received in April, 2010
Reviewed in May, 2011
Revised version received in June, 2011