A Comparison of Reading Comprehension Tests: 
Multiple-Choice vs. Open-Ended

Myong Hee Ko 
(Korea University)

This study compares the differences between a multiple-choice (MC) and an open-ended (OE) reading comprehension test in a within-subject design to investigate test validity by comparing the two methods. A total of 70 Korean undergraduates participated in the present study. After taking a cloze test, participants read a passage, returned it and took two types of reading comprehension tests. Results showed that there was a statistical difference between the two test methods in a within-subjects design, which agrees with previous between-subjects studies. The higher scores on the MC format were attributed to three reasons: memory constraint, the nature of the task, and test-taking strategy. Further analysis indicated that both proficiency level groups were significantly affected by the test methods. Moreover, both the MC and OE tests significantly separated the two proficiency levels. Information from a survey provides more information regarding differences between the two test methods.

I. INTRODUCTION

Reading is still the most extensively offered course in universities in Korea. This is probably because both educators and learners see it as the skill they can benefit most from as they do their best in a limited language learning environment, an EFL context. In fact, many people (especially, the educated) get information through the Internet, and thus have more chances to practice and use English. Given the fact that a great amount of information on the Internet is in English, reading ability is an essential tool for many English learners so that they can keep up with updated information or knowledge. Reading in English is crucial for those in science, medicine, or technology because they will need to read books and articles published in English throughout their careers. For these reasons,
reading has been regarded as an important skill. Accordingly, for the majority of the English learners in EFL instruction, reading is the most common among the four skills.

A reading-based English program has existed in universities for a long time and will continue to exist in the foreseeable future in Korea. There is no doubt that reading assessment will exist as long as reading courses exist. The need for valid and reliable measurement of reading comprehension in order to assess students’ ability accurately is unquestionable. In general, teachers in the English reading program use various reading assessment tools such as multiple-choice, written recall, sentence completion, open-ended, true/false, matching, and fill-in-the-blanks (Bernhardt, 1991).

Among these choices, based on my teaching experience, multiple-choice (MC) and open-ended (OE) tests appear to be the most widely used assessment tools among teachers in university English reading programs. The two different testing procedures seem to trigger different outcomes; however, both of them are still commonly used testing techniques employed by most language teachers. They will remain in use for the foreseeable future unless a more effective way to test learners’ reading comprehension is found. This means that teachers have to judge their learners’ language ability under limited conditions, which may display only a restricted part of students’ performances using a given test method. If examinees get different scores depending on different test methods, this calls the validity of the tests into question.

Thus, this study will investigate the effects of two test methods in measuring students’ reading comprehension. It also attempts to provide detailed information pertaining to these two test formats, using examinee opinion surveys. Awareness is essential. Teachers need to know about these two types of measurements since they will be using them as long as they teach English reading courses. Along with this awareness, relevant information regarding these two test formats will be useful to them when preparing reading comprehension tests.

Accordingly, this study is designed in the following way: First, considering that Shohamy (1984) and Wolf (1993) were based on a between-subjects design, this study incorporates the motive to look into how the same participants behave using the two test methods in a within-subjects design. By using a within-subjects design, it may be able to scrutinize the different assessment techniques with the same examinees. Second, in order to check the effect between proficiency levels, the levels of examinees were more clearly differentiated as stated in the procedure. By doing this, it may provide a clearer picture about the effect of proficiency levels on the test methods, for which Shohamy (1984) and Wolf (1993) have shown contradictory results. Third, a survey of participants’ perspectives on these two tests was included in order to provide additional information. Accordingly, the following research questions are addressed.

1. Is there a statistically significant difference between the two test methods?
2. Are there statistically significant differences between the test methods within each proficiency level?
3. Are there statistically significant differences between the proficiency levels within each method?
4. What are the participant perceptions and attitudes towards the different types of test format?

II. REVIEW OF LITERATURE

Measuring reading ability has been a major activity among teachers and researchers. In terms of obtaining observable protocols, assessing receptive skills such as reading and listening is more difficult than assessing productive skills such as speaking and writing. For example, it is not easy to obtain and test samples of examinees’ reading skills because the process of reading is unobservable (Anderson, Bachman, Perkins & Cohen, 1991; Wolf, 1993). On the other hand, measuring productive skills such as speaking and writing is relatively easy since these types of assessment are susceptible to observable protocols.

Two factors are involved when testing reading: the trait and the method. The trait is the specific knowledge that is assessed, and the method is the specific procedure or technique chosen for measuring the trait. Since multiple methods are available for a given trait, researchers have been curious about how different techniques influence measuring a specific trait (Shohamy, 1984). Although there is great variation among standardized reading tests, such as think-aloud protocols (Anderson et al., 1991), MC and OE questions still seem broadly accepted as a means for measuring learner reading comprehension. Along with true/false and matching items, a MC question is a kind of selection response task. A test-taker chooses an appropriate response among choices. MC is the most popular means of testing because learners are familiar with the procedure and it is easy to score, especially in a large class. An OE question is a kind of construction response task, which elicits one-word answers or short phrases. A test-taker needs to write the information on paper (Alderson, 2000; Brantmeier, 2005; Wolf, 1993).

In reality, the most common ways of assessing L2 learners’ receptive skills are through indirect procedures such as MC, OE, true/false, matching, summary, and cloze tests. These indirect techniques are of course limited in their capacity to explore L2 learners’ complex mental processes during reading because the answers given by the learners do not show in detail how they arrived at their interpretations (Alderson, 1984; Bachman, 1990). Accordingly, measuring reading skill using existing test methods leaves doubts and speculation. Some studies (e.g., Cheng, 2004; Shohamy, 1984; Wolf, 1993) have demonstrated that not only learner’s abilities but also a greater selection of test methods
could bring about different outcomes.

Cheng (2004) was interested in comparing three types of tests in assessing listening comprehension using a within-subject design: MC, multiple-choice cloze (MCC), and OE tests. One hundred and fifty-nine college students in Taiwan were divided into three treatment conditions where they listened to 30 minute dialogues and took three types of tests in a way that minimized practice effect (sequencing effect). The results showed that both MC and MCC tests indicated significantly higher scores than the OE test. A post-test survey revealed that 97% of the participants preferred the MC formats due to spoken stimuli and guessing possibility.

Two studies, Shohamy (1984) and Wolf (1993), compared MC and OE in assessing reading comprehension. Shohamy (1984) compared MC and OE along with languages (L1 and L2) and texts (easy and difficult). Four different tests (MC English, MC Hebrew, OE English, and OE Hebrew) with two different texts were given as part of a foreign language matriculation examination at a secondary school in Israel. The tests were designed so that no one would take the same test more than once; it was a between-subjects design. The findings showed that there were significant effects depending on method, language, and text. The students performed significantly better on the MC questions than on the OE questions, better on L1 than L2, and better on the easier text than on the difficult text. Shohamy also hypothesized that the low level students were likely to be affected by the three variables, whereas the high level students were hardly affected by any of them.

Wolf (1993) was interested in comparing the effects of three different methods (MC, OE, and cloze) along with language (L1 and L2) using a between-subjects design. The students who participated in that study were students learning Spanish at a university in the US. Each student read the same passage and took one of the six tests. The results indicated very similar findings to those of Shohamy (1984). The students performed significantly better on the MC than on the OE and cloze tests. Regarding the influence of students’ proficiency levels on the variables, Wolf found that even advanced learners were affected by test method, though they showed relatively less variation than did low-level learners. This is inconsistent with the findings reported by Shohamy (1984). One of the main reasons might be attributed to clearer level distinction. Wolf (1993) divided participants into two groups based on the number of hours of target language instruction (under 300 hours vs. about 700 hours), whereas Shohamy (1984) had relatively narrower intervals among three proficiency levels based on scores on the control part of test (1-3, 4-6, & 7-8 points, respectively).

In’nami and Koizumi (2009) conducted a meta-analysis on MC and OE formats on L1 reading, L2 reading, and L2 listening. Regarding L2 reading, they did not find any significant effect on the test formats although the MC format was found to be easier than the OE one. In the case of between-subjects design, only two published studies (i.e.,
Shohamy, 1984; Wolf, 1993) were included in a study of L2 reading. In the case of between-subjects design, one published study (i.e., Cheng, 2004) was used in a study of L2 listening. The authors stated more studies are needed in order to obtain more precise information on the effect of test methods.

As a consequence, more investigation is necessary in order to come to a better understanding of these two test formats. The current study attempts to investigate the test format effects on L2 reading using a within-subjects design since we need to have more research on this area, especially in L2 reading. This study is designed to find out how examinees behave when using these two test methods. The quantitative examinee data is supplemented by their opinions and attitudes about the two test formats. By knowing the advantages and limitations of each method, teachers may have available a clearer idea of the nature of these test methods, and the extent of their effects. Moreover, this information may make teachers wiser in minimizing negative impacts on their examinees while constructing the test. What is learned here may eventually provide a good data basis for searching for better ways to develop testing techniques in the future (Bachman, 1990; In’nami & Koizumi, 2009).

III. METHOD

1. Participants

University students taking English reading classes participated in the present study. The classes consisted of freshmen, sophomores, juniors, and seniors with various majors such as humanities, science, engineering, and education. Their ages ranged from 19 to 26. They mostly had had six years of English education prior to entering the university and no one had lived in English speaking countries. The participants were divided into two proficiency groups (low and advanced) based on the results of the cloze test. The mean score of the cloze test was about 15.5. Thirty-five of the examinees scored equal to or above 18 were assigned to the advanced level, and thirty-five who were equal to or below 15 were assigned to the low level. Eighteen was used as cutoff so that there was an equal number of test-takers at each level. Those who scored between 16 and 17 were excluded in order to have a meaningful interval between the two proficiency levels (Hudson, 2007). Thus, in sum, 70 examinees participated in this study.

2. Materials

An expository passage from the advanced level of the NorthStar reading series (Miller
& Cohen, 1998) was used. It is a piece of non-fiction, originally published in Reader’s Digest. It is a story about a Sarajevo cellist who played his cello on the street during the war in 1992. This particular story was selected in order to minimize possible activation of content schema of the participants because the content of the story was thought to be neutral. The text’s vocabulary was revised by paraphrasing with more commonly used words. The revised version was checked by a native speaker.

The length of text was adjusted to be neither too long nor too short. The length of the text was 912 words in total. It was thought that if the passage was too short, it would limit the number of questions and might complicate assessment of students’ reading comprehension. It was thought that if the passage was too long, then students would have much difficulty in retrieving the information. The passage scored 9.8 on the Flesch-Kincaid Grade Level readability formula tool in Microsoft Word. In other words, the level was appropriate for native speakers who are in the 8th month of the 9th grade. The appropriateness of readability level of the passage was checked with students with a similar background before giving it to the participants (see Appendix A).

Given the length of the reading, a total of 22 OE questions were created from all parts of the passage in order to check overall understanding of the story. Originally, 24 items were created, but 2 of them were deleted after pilot testing. The questions were written in English and mostly queried factual information about the story, at both micro- and macro-levels. All questions were designed for short answers, such as a phrase or sentence. The scope of possible responses was limited for the sake of grading. The questions were revised after piloting. All items were passage dependent; any questions which could be answered without reading the text were removed. Ambiguous items were rewritten. In order to prevent them from failing to response due to the difficulty of vocabulary, none of the words used in the questions were unknown to them (Alderson, 2000; Brantmeier, 2003; Cheng, 2004).

Once the OE questions were completed, a MC format test was designed with the same question items in order to be equivalent to the OE test. Each MC question had four options: 1 answer and 3 distracters. The choices were revised and refined a couple of times, based on the pilot testing (see Appendix B).

A cloze test from Watanabe (1997) was given to the students to measure their general L2 language proficiency. It consisted of 277 words with 25 blanks. The passage is about aspirin. It was at a 9th grade level measured by the Flesch-Kincaid readability scale (see Appendix C).

3. Pilot Testing for Items

Using FACETS analysis, item fitness and consistency of the two tests were investigated
in order to check whether the test items were appropriate to assess participants’ reading comprehension. FACETS, a software program for multi-faceted Rasch analysis, offers detailed information on the facets and their interactions, along with an examinee’s ability and item difficulty (Lee, 1996; Lynch & McNamara, 1998). FACETS, version 3.0 (Linacre, 1996), was used to provide detailed information about each test item, such as whether it fit properly or not to the design of the specific test. This kind of information is useful since it indicates some degree of validity and reliability of the test being examined. One interesting feature of FACETS is that a bias analysis can generate the combination of two facets and provide information about an interaction between them.

Figure 1 shows the measures for examinees, proficiency levels, test methods, and items, respectively. The scale in the first column is the logit scale, which ranges from -2 to +2. The second column indicates the examinees’ ability, which varied from -1.5 to +2 level. The third column shows the examinees’ proficiency levels. Those at the advanced level performed better than those at the low level. The fourth column exhibits the difficulty of the two kinds of test methods. The OE type was more difficult than the MC type. The last column displays the difficulty of the original 24 items. They seemed to spread well along the continuum of the given range from -2 to +1.5.

FIGURE 1
FACETS Summary on Examinee Ability, Proficiency Levels, Test Difficulty, and Item Difficulty

<table>
<thead>
<tr>
<th>Measure</th>
<th>+Examinees</th>
<th>+Levels</th>
<th>-Tests</th>
<th>-Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>****</td>
<td>**</td>
<td>**</td>
<td>OE</td>
</tr>
<tr>
<td></td>
<td>***</td>
<td>*</td>
<td>*</td>
<td>***</td>
</tr>
<tr>
<td></td>
<td>***</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>0</td>
<td>*</td>
<td>*</td>
<td>Advanced</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>*</td>
<td>Low</td>
<td>*</td>
</tr>
</tbody>
</table>
Table 1 below presents the bias interaction between the items and the two test methods. It indicates that two out of 24 items were outside the acceptable range (from -2 to +2) as shown in Table 1 below. The result of chi-square of 110.9 with 48 df was significant at $p < 0.05$. Two items, 8 and 16 had z-scores lower than -2.0 in the OE test, and they had z-scores higher than 2.0 in the MC test. In other words, the examinees gained higher scores than expected on the OE test, whereas they got lower scores than expected on the MC test. These two items appeared relatively easy overall since most of the examinees wrote the answers correctly in the OE format. The reason for the lower scores in the MC was that examinees might have been misled due to very plausible distracters. Based on the findings, the two items were removed for the sake of validity of the test. As a result, 22 items were kept.

<table>
<thead>
<tr>
<th>Item</th>
<th>Test</th>
<th>Obs score</th>
<th>Exp score</th>
<th>Obs Count</th>
<th>Obs-Exp Ave</th>
<th>Biss+ measure S.E.</th>
<th>Infit mnsq</th>
<th>Mean (count:48)</th>
<th>SD</th>
<th>Fixed (all = 0)</th>
</tr>
</thead>
</table>

**TABLE 1**
Bias Interaction between Tests and Items
4. Procedures

Data collection took place during regular class time. Seventy students were given a cloze test as a means of checking their general proficiency levels. After that, the participants were informed that they would read a passage, hand it in, and take the OE and MC tests. Although the participants were familiar with these two test types, they were informed clearly about what these two tests were in order to make sure. Regarding the reading material, the participants reported that they had never read the text before; therefore, it was new to them.

They were given 18 minutes to read, 17 minutes to take the OE test, and 12 minutes to complete the MC test. The time limit was set based on the pilot test. Since this study is planned to be a within-subjects design, the students were supposed to take the two types of tests in sequence. They were given the OE type first and the MC type next in order to minimize the practice effect. It was expected that this sequence would minimize the learning effect on the MC format. If the tests had been given the other way around, the examinees would have remembered answers or cues from the choices of the MC items.

Examinees were allowed to write the answers in either their native language or the target language in the case of OE test so that they wouldn’t feel constrained from expressing their thoughts freely. It was thought that the language of the expected response might influence participants’ ability to demonstrate to what extent they understood (Bachman, 1990; Buck, 2001; Cheng, 2004). In order to have information for unanswered questions, they were also asked to briefly provide reasons in Korean if they were not able to write the answers for any of the questions in the OE format.

They were told that they would hand in the reading material before the test since it is thought that their reading comprehension ability could be better checked in this way. The passage was relatively short. If the text had been available, the examinees might have used search-and-match strategies instead of actual comprehension to complete the test. Besides, in the case of the OE test, there is a possibility that they would have answered the questions using the language from the passage rather than their own words (Alderson, 2000). After completion of the test, the participants completed a survey. The survey was created by the author in order to check participant opinions about and preferences toward the two tests: (a) knowledge about the content, (b) discrimination between proficiency levels, (c) appropriate test methods, and (d) personal preference (see Appendix D).

5. Analysis

The cloze test was scored by giving 1 point for each correct answer and 0 points for each wrong answer since the answers were clear-cut, that is, either correct or wrong. The
same grading system is also employed regarding the MC and OE tests. With respect to the OE questions, partial points were not given because it was thought that the possible answers were clear-cut; they were short answers using phrases or short sentences. Since there were 22 items for each test, the total points for the OE and MC were each 22 points. Twenty percent of OE items were scored by both the author and her colleague for the sake of interrater reliability. The interrater reliability was 0.948, based on the percentage agreement by the two raters. The items on which the two raters disagreed were discussed and negotiated to reach the final decision.

IV. RESULTS

The reliability of the cloze, the OE and the MC tests were 0.85, 0.77, and 0.77, respectively measured by Cronbach’s alpha. The descriptive statistics for the scores of the examinees are presented in Table 2 below. A total of 70 students took the cloze test, the MC and the OE test, respectively. The table displays their mean scores and standard deviations.

<table>
<thead>
<tr>
<th>TABLE 2</th>
<th>Descriptive Statistics of Examinees</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
</tr>
<tr>
<td>Cloze Test</td>
<td>70</td>
</tr>
<tr>
<td>MC Test</td>
<td>70</td>
</tr>
<tr>
<td>OE Test</td>
<td>70</td>
</tr>
</tbody>
</table>

The first research question concerned whether there was a significant effect between the MC and the OE tests in a within-subjects design. A repeated t-test indicated that there was a statistical difference between the two test methods ($t(69) = 19.279, p < 0.025$). Accordingly, there was a significant difference between the test methods. In other words, the same students behaved significantly differently depending on the test formats.

The second research question considered whether there were significant differences between the methods within each proficiency level. A repeated two-way ANOVA was conducted to examine the two methods used to test the two proficiency levels (low and advanced). The results indicated that the interaction of levels and methods was not significant, but there was a significant main effect of the levels and methods as shown in Table 3 below.
TABLE 3

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Sq.</th>
<th>df</th>
<th>Mean Sq.</th>
<th>F</th>
<th>Sig.</th>
<th>note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>1632.029</td>
<td>1</td>
<td>1632.029</td>
<td>369.786</td>
<td>0.00*</td>
<td>0.421</td>
</tr>
<tr>
<td>Method*Level</td>
<td>2.857</td>
<td>1</td>
<td>2.857</td>
<td>.647</td>
<td>0.424</td>
<td>0.0007</td>
</tr>
<tr>
<td>Error (Method)</td>
<td>300.114</td>
<td>68</td>
<td>4.413</td>
<td></td>
<td></td>
<td>0.077</td>
</tr>
<tr>
<td>Between-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Levels</td>
<td>787.314</td>
<td>1</td>
<td>787.314</td>
<td>46.407</td>
<td>0.00*</td>
<td>0.203</td>
</tr>
<tr>
<td>Subject Error</td>
<td>1153.657</td>
<td>68</td>
<td>16.966</td>
<td></td>
<td></td>
<td>0.297</td>
</tr>
</tbody>
</table>

*Mean difference is significant at the 0.05 level.

Since the repeated two-way ANOVA does not provide post hoc analysis, a follow-up contrast analysis was performed in order to locate the source of the difference. The p-value for the subsequent repeated t-test was adjusted to 0.025, dividing 0.05 by the number of analyses. The low MC and low OE groups were compared by two consecutive repeated t-tests to locate the significance. The results pointed out that both the low and advanced level groups were significantly affected by the test methods ($t(34) = 15.857, p < 0.025$, $t(34) = 11.884, p < 0.025$, respectively). In other words, both low and advanced level students behaved differently depending on the test methods. The effect size was 0.421, indicated by eta squared.

The third research question involved whether there are significant differences between the proficiency levels within each method. The repeated two-way ANOVA also indicated that there was a significant main effect of methods as shown in Table 3 above. The effect size was 0.203. Another follow-up analysis, two consecutive t-tests with an adjusted p-value at 0.025, was conducted to locate the significance. The results indicated that there was significant difference in both MC method ($t(68) = -5.689, p < 0.025$), and OE method ($t(68) = -6.449, p < 0.025$). This implies that both MC and OE tests differentiated the low level from the advanced level examinees.

The fourth research question concerned learner opinions about and preferences between the test methods. Seventy students participated in the survey. They were asked about four areas: (a) knowledge about the content; (b) discrimination between proficiency levels; (c) appropriate test methods for examinees; and (d) personal preferences as to test methods. The results are presented in Tables 4, 5, 6, and 7, respectively.

TABLE 4

<table>
<thead>
<tr>
<th>Knowledge about the Content</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MC format</strong></td>
<td></td>
</tr>
<tr>
<td>● If I know a certain amount of the content, I still can answer.</td>
<td>55</td>
</tr>
<tr>
<td>● I need to know the requested content accurately in order to provide an answer.</td>
<td>10</td>
</tr>
<tr>
<td>● Even if I do not know the content at all, I still can choose an answer by guessing.</td>
<td>5</td>
</tr>
<tr>
<td><strong>OE format</strong></td>
<td></td>
</tr>
</tbody>
</table>
Table 4 shows students’ perceptions of MC and OE formats. In the case of MC format, 55 students reported that they still could answer if they knew a certain amount of the content. However, in the case of OE format, 64 students responded that they needed to know the requested content accurately in order to provide an answer. It appears that the OE format demands more on the knowledge about the content than the MC format.

<table>
<thead>
<tr>
<th>TABLE 5</th>
<th>Discrimination between Proficiency Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Formats &amp; Proficiency Levels</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>MC format</strong></td>
<td></td>
</tr>
<tr>
<td>• Between advanced and high-intermediate levels</td>
<td>38</td>
</tr>
<tr>
<td>• Between high-intermediate and low levels</td>
<td>60</td>
</tr>
<tr>
<td>• Between advanced and low levels</td>
<td>70</td>
</tr>
<tr>
<td><strong>OE format</strong></td>
<td></td>
</tr>
<tr>
<td>• Between advanced and high-intermediate levels</td>
<td>68</td>
</tr>
<tr>
<td>• Between high-intermediate and low levels</td>
<td>68</td>
</tr>
<tr>
<td>• Between advanced and low levels</td>
<td>70</td>
</tr>
</tbody>
</table>

Table 5 presents how the two test methods discriminate between proficiency levels. Most students indicated that, in general, the OE format is better at separating among levels. In the case of the MC format, 32 students reported that it may not differentiate advanced level from high-intermediate level.

<table>
<thead>
<tr>
<th>TABLE 6</th>
<th>Appropriate Test Methods for Examinee Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appropriate Test Method for Examinee Levels</td>
<td>Frequency</td>
</tr>
<tr>
<td><strong>Low level</strong></td>
<td></td>
</tr>
<tr>
<td>MC format</td>
<td>61</td>
</tr>
<tr>
<td>OE format</td>
<td>7</td>
</tr>
<tr>
<td>Mixed format</td>
<td>2</td>
</tr>
<tr>
<td><strong>Intermediate Level</strong></td>
<td></td>
</tr>
<tr>
<td>MC format</td>
<td>46</td>
</tr>
<tr>
<td>OE format</td>
<td>15</td>
</tr>
<tr>
<td>Mixed format</td>
<td>9</td>
</tr>
<tr>
<td><strong>Advanced Level</strong></td>
<td></td>
</tr>
<tr>
<td>MC format</td>
<td>9</td>
</tr>
<tr>
<td>OE format</td>
<td>57</td>
</tr>
<tr>
<td>Mixed format</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 6 displays an appropriate test method for each proficiency level. Sixty-one students indicated that a MC format is the most appropriate reading assessment tool for
low level students. Forty-six students also selected a MC format for intermediate level. However, fifty-seven students chose an OE format for advanced level students.

<table>
<thead>
<tr>
<th>Preferred Test Format</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC format</td>
<td>45</td>
</tr>
<tr>
<td>OE format</td>
<td>17</td>
</tr>
<tr>
<td>Mixed format</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 7 presents personal preference for test methods. Forty-five reported that they preferred a MC format. Seventeen students favored an OE format over a MC format. Eight students suggested a mixed format approach.

V. DISCUSSION

The first research question was whether there is a statistically significant difference between the two test methods. The findings of the present study showed that there was a statistically significant difference between the two test methods. These results may stem from three reasons: memory constraint, the different nature of the task, and test-taking strategies. First, learners’ memory constraint appeared to play a role as one of the variables, as Wolf (1993) indicated. The difference between Wolf’s (1993) and the present study is students’ pre-knowledge of the reading assessment. She wondered whether the results of OE test were due to her not informing participants about the test. Wolf indicated that, in the case of her OE test, students might not have tried to memorize or rehearse the information in the text while reading because they were not told that their comprehension would be tested after reading the passage. She suspected that this might have made it difficult for them to recall the information. However, students in the present study were explicitly told they would be tested in the forms of MC and OE after reading. Thus, memory constraints seemed an important moderating factor. Unlike the MC format, some blanks were left when examinees could not answer in the OE format. As students were invited to report in Korean, some of them indicated that they simply did not focus on those facts while reading. Some said that they could partially remember the relevant part, so they were confused and not sure how to express their thoughts. Others said that there was so much information in the passage, that it was not easy to remember everything. Recalling information was difficult because in general no retrieval cues were provided.

Second, as Shohamy (1984) and Wolf (1993) claimed, the MC format is an easier task than the OE format since the former requires comprehension and selection while the latter
demands comprehension and production. Even if an examinee does not know the answer for sure, he/she can at least guess from among the given choices in the MC test, whereas a test-taker can hardly produce at all if he/she does not know the answer in the OE test. A difference between selection and production appeared to exist in assessing reading comprehension. Selection (e.g., MC) may require surface level processing when reading since students need to recognize what is an appropriate response among choices. On the other hand, production (e.g., OE) may require relatively deeper level processing while reading since students retrieve ideas from working memory and produce them independently on the paper. A great deal of retention of the reading material may be needed in the case of OE format.

Third, test-taking strategy also appeared to be an important variable since it contributed to higher scores in the MC format. Examinees well trained in MC testing may be good at finding an answer even if they do not know the answer for sure. In general, Korean students are well trained and accustomed to MC tests since childhood. Most achievement tests, including a college entrance examination, are in the form of MC. As a consequence, the MC test might have been much easier for them than the OE equivalent.

The second research question was whether there are statistically significant differences between the test methods within each proficiency level. Students in both the low and advanced proficiency levels were affected significantly by the different test methods. Shohamy (1984) hypothesized that low level learners would be more sensitive to the different test methods, showing a 20% variance, whereas advanced level learners would be hardly affected by this variable displaying only a 4% variance. However, the findings of this present study did not support her findings. Instead, this study was congruent with the results of Wolf (1993) in that both levels were affected by the test methods. As mentioned earlier in the literature review, the degree of distinction of proficiency level might have been the main attribution to this result. Wolf (1993) had relatively clearer separation between the two levels (low and advanced) according to the amount of target language instruction, whereas Shohamy (1984) had relatively narrower intervals among the three proficiency levels. The present study also differentiated the two proficiency levels, leaving a noticeable interval between the low and the advanced. Thus, it may be reasonable to conclude that students could be affected by test methods regardless of their proficiency levels.

The third research question was whether there are statistically significant differences between the proficiency levels within each method. There were statistically significant differences between the advanced and low levels within each method. Both MC and OE tests were able to separate the low level from the advanced level. However, the results of the survey provided additional insights into these two test methods.

The fourth research question was to investigate examinee perceptions and attitudes
towards the two formats: (a) knowledge about the content; (b) discrimination between proficiency levels; (c) appropriate test methods for examinee levels; and (d) personal preferences as to test methods. Regarding the amount of knowledge of the content for assessment, they indicated that the MC format was less of a burden than the OE format. Fifty-five out of 70 responded that, with a certain degree of understanding the content, they still could answer. Three main reasons mentioned by these examinees were: (a) I can get contextual cues from the distracters; (b) I can choose an answer by getting rid of distracters; and (c) I can still choose an answer if we know the main ideas and the atmosphere of the content. They reported that the MC format provides retrieval cues, i.e., they could get some cues from the choices. They also mentioned that even though some parts of their information on the content were vague or ambiguous, once they had some knowledge of the general atmosphere of the story, they could perform relatively easily based on the given cues from distracters. It seemed that they somehow could answer without knowing detailed information from the passage. On the contrary, in the case of the OE format, 64 students indicated that they needed to know the requested content accurately in order to produce a written answer. They reported that since there are no choices to select, they have to know the content for sure. It appeared that the OE format revealed their ability to answer more overtly compared to the MC format. Interestingly, most of the examinees were able to provide the main idea of the story. The macro-level questions seemed easy for them to remember though they did not make an effort to memorize the information.

When asked about discriminating ability between levels, students provided interesting results. As shown in Table 5, in the case of MC format, 32 students out of 70 responded that this method may not separate the advanced level from the high-intermediate level. They further explained that the amount of understanding between the two levels may not be distinguished clearly. For example, the MC format may not be able to discriminate between those with 100% comprehension and those with 70-80% comprehension. They see that, compared to the OE, MC is less able to discriminate between the advanced and high-intermediate levels. In other words, they seemed to see that the MC test could be a rather poor measurement in discriminating those who are able to perform really well and those who perform moderately well. The MC format may not clearly separate those who know the answer from those who only know it partially. Because of some retrieval cues and a chance to select among listed choices, those who knew the information partially could succeed in getting the answer just as well as those who completely knew the answer. The participants reported that they would read thoroughly for the OE type, but they would not pay much attention to the detailed information for the MC test because there is the possibility of choosing right answers without knowing everything. Conversely, they perceived that the OE format was better at discriminating examinees among the levels.
Examinees could not guess or get any assistance from any cues since those aids were not available. Those who knew the answer could write it, and those who did not know the answer could not provide it at all. It seemed there was not any gray (vague) area between those who knew or those who did not know the answer.

When asked about an appropriate test method for each level, most students (61 and 46 examinees, respectively) indicated that a MC format was the most appropriate way to assess low level and intermediate level students. They reported that since these students may not fully understand the material, they may find it difficult to produce answers. The reason that the OE format was regarded as inappropriate was that it is seen as unable to distinguish between those who have some partial knowledge and those who do not have any knowledge. If the examinees do not know the information well enough, it may be difficult for them to produce anything due to the nature of the task, just like those who have no idea at all. Students also mentioned that if the test is difficult, they may lose confidence. Several students proposed that the OE format may motivate them to read more accurately and to study more. They seemed to be looking for a positive washback effect of the test. In the case of intermediate level, 46 students proposed a MC format. They see that an OE format is difficult because most students in this level may not be good at understanding the content. Interestingly, 15 students proposed an OE format because they think that this may motivate them to read more accurately. In contrast to this, 59 students proposed an OE format for advanced level students. Three main reasons were: (a) since advanced level students understand material well, they may not need choices to select; (b) this format requires detailed information, so it may be appropriate for advanced level students; and (c) it may discriminate better among advanced students. In fact, those who were in the advanced level group performed quite well even in the questions which asked for detailed information.

For personal preference, as shown in Table 7, most of the participants preferred a MC format over an OE format because the former burdened them less when preparing for a test. Those who chose a MC format reported that (a) selecting answers is easier than producing them; (b) they had been used to this type of testing since they were young; and (c) they do not belong to advanced level. Those who preferred an OE format indicated that this method may motivate them to study more and discriminate better than the MC type. Interestingly, eight of them preferred a mixed approach because they thought it would be better for the sake of effective assessment since either type of test was not perfect.

VI. CONCLUSION

The present study has shown that learner test performance can be affected by the
method used to elicit their responses. The examinees scored significantly higher in the MC type than in the OE type. The causes were thought to be memory constraint, the different degree of task difficulty, and test-taking strategies. Moreover, learners at a low level as well as at an advanced level were influenced by the effects of different test methods when the two levels were differentiated clearly. Furthermore, the two test methods were able to separate low level learners from advanced level learners within each method. Findings from the survey also provided useful information to teachers. When adopting these two test types, teachers need to consider (a) learner level (b) cultural and educational backgrounds of learners, and (c) the purpose of the assessment.

Regarding learner level, if the learners are advanced, the OE format might be more appropriate since their reading ability is good enough to comprehend the text and also much better at holding information to produce when they are asked later on paper. Most examinees proposed an OE format for advanced level students in the opinion surveys. In fact, those at an advanced level showed that they could answer detailed questions in the OE test format. On the other hand, if the learners are low level, the MC format might be more appropriate since most of the time they may have difficulty in producing their partially gained information due to less competent comprehension ability. In the opinion surveys, most students indicated that a MC format was the most appropriate way to assess low level and intermediate level students.

Another factor to be considered is the cultural and educational backgrounds of the learners. The reason that the participants in the present study preferred the MC test, regardless of their level, could be attributed to their test-taking strategies. Given that they had been well trained for MC type tests in preparation for the highly competitive college entrance examination while in high school, they might have been good at taking this kind of test. That is probably why many students indicated that the MC test might not discriminate well enough between advanced level and high-intermediate level students. However, students may not be aware of that a well-made MC test can discriminate between these two proficiency levels as well as the OE test format can.

Lastly, the purpose of assessment has to be taken into consideration. If a teacher wants to make students understand reading material thoroughly, an OE format may be appropriate. According to the opinion surveys, most of the examinees indicated that, in the case of OE format, they needed to know the requested content accurately in order to produce a written answer. They reported that since there are no choices to select, they have to know the content for sure. However, when the purpose of reading is to retain general ideas with some distinct facts, a MC format may be better. Most of them reported that, with a certain degree of understanding the content, they still could answer. They indicated that once they had some knowledge of the general atmosphere of the story, they could perform relatively easily due to the given cues from distracters. If the two tests are selected
appropriately, a positive washback effect may bring about improvement in student study behaviors.

Overall, it appears that different formats of tests gauge different aspects of learners’ reading ability (Shohamy, 1997; Urquhart & Weir, 1998). Since readers’ comprehension may vary according to their proficiency level, background knowledge, topic, and genre, there may, in reality, be no perfect test to measure interactions of all these variables (In'nami & Koizumi, 2009; Wolf, 1993). We may be measuring a limited set of skills in reading comprehension with a narrow range of test types (Sternberg, 1991). Alderson (2000) claimed that there is no best method to assess learners’ reading comprehension because “no single test method can fulfill all the varied purposes for which we might test” (p. 203). Therefore, teachers need to be aware of these facts in order to construct more effective reading comprehension tests. Depending on their students’ levels, their educational and cultural backgrounds, and purpose of the assessment, teachers need to able to choose or mix the two types of tests.

For future research, it would be worth investigating more among the lines of learners’ proficiency levels in the two different reading comprehension test methods. Follow-up studies with various proficiency levels may be necessary in order to scrutinize student behavior in each level in detail. For example, it may be worth checking examinees’ behaviors between advanced and high-intermediate in the MC and OE methods. Furthermore, studies are also required to compare other types of reading comprehension tests such as matching, true/false, and summary as well, in order to come up with an overall understanding of the positive and negative aspects of various test methods which are being used among teachers in the present day. Findings in these areas may contribute a good knowledge base to teachers in becoming better evaluators of their students.

ACKNOWLEDGMENT

I wish to thank Thom Hudson for his insightful comments on the previous draft of this article. I also wish to thank three anonymous reviewers for their constructive feedback for revising this article.

REFERENCES

A Comparison of Reading Comprehension Tests: Multiple-Choice vs. Open-Ended

APPENDIX A
A Sample Text

As a pianist, I was invited to perform with cellist Mike Chang at the International Cello Festival in England. Every two years, a group of the world's greatest cellists and others devoted to the not-so-popular instrument such as bow makers and collectors gather for a week of workshops, seminars, recitals and parties. Each evening, the 600 or so participants gather together for a concert. The opening-night performance at the Royal College of Music consisted of works for unaccompanied cello. There on the stage in the big concert hall was a solitary chair: No piano, no music stand, no conductor's podium. This was to be music in its purest, most intense form. The atmosphere was filled with anticipation and concentration. The world-famous cellist Yo-Yo Ma was one of the performers that April night in 1994, and there was a moving story behind the musical composition he would play. On May 27, 1992, in Sarajevo, one of the few bakeries that still had a supply of flour was making and distributing bread to the starving, war-shattered people. At 4 p.m. a long line stretched into the street. Suddenly, a large bullet of a big gun fell directly into the middle of the line, killing 22 people and splattering flesh, blood and bone.

APPENDIX B
A Sample Multiple-Choice Test

*Choose the best answers.

1. Why is the author invited to the festival?
   a. To play the violin
   b. To help to organize the festival
   c. To perform with cellist Mike Chang
   d. To watch and report on Yo-Yo Ma’s concert

2. What kinds of people came (gathered together) for the concert?
   a. world famous musicians and music fans
   b. a group of the world’s greatest cellists and pianists
   c. people who work for popular music and who use various musical instruments
   d. famous cellists and people who work for instruments that are not well known

3. What was something unique (special) about the stage on the opening-night performance at the Royal College of Music?
   a. Only the cello was played.
   b. There were only two chairs.
   c. The concert was organized to help Sarajevo.
   d. There was everything except piano and podium.

4. What was the atmosphere like at the festival on that night before the concert had started?
APPENDIX C

A Cloze Test

Directions: Read the following passage and fill out the blanks choosing the right word from the box. You may use the word more than once if necessary.

Ordinary Aspirin Is Truly a Wonder Drug

Aspirin is the most popular drug 1._________ the world today. Many people take aspirin when they have a headache. It is effective in relieving other pains too. Americans this year will swallow 15,000 tons of aspirin, one of 2._________ safest and most effective drugs invented by man. The 3._________ popular medicine in the 4._________ today, it is an effective pain reliever. Its bad effects are relatively mild, and it is cheap.

For millions of people suffering 5._________ arthritis, it is the only thing that works. Aspirin, in short, is truly the 20th-century wonder drug. It is also the 6._________ largest suicide drug and is the 7._________ cause of poisoning among children. It 8._________ side affects that, although relatively mild, are largely unrecognized among users.

Although aspirin 9._________ first sold by a German company 10._________ 1899, it has been around much 11._________ than that. Hippocrates, in ancient Greece, 12._________ the medical value of the leaves 13._________ tree bark which today are known 14._________ contain salicylates, the chemical in aspirin. 15._________ the 19th century, there was a 16._________ deal of experimentation in Europe with this chemical, and it led to the 17._________ of aspirin. By 1915, aspirin tablets 18._________ available in the United States.

A 19._________ quantity of aspirin (two five-grain tablets) 20._________ pain and inflammation. It also reduces 21._________ by interfering with some of the 22._________ reactions. Specifically, aspirin seems to slow 23._________ the formation of the acids involved 24._________ pain and the complex chemical reactions 25._________ cause fever. The chemistry of these acids is not fully understood, but the slowing effect of aspirin is well known.

and body’s but by down during introduction fever
from great are at most pulls understood small
that were was in has second relieves world
longer of is to the leading
**APPENDIX D**

Questionnaire

High level refers to those who understand the passage nearly 100%.
Upper-Intermediate level refers to those who understand the passage around 70-80%.
Low level refers to those who understand the passage around 20-40%.

*Choose items that agree with you and provide reasons below.*

I. Knowledge about the Content for Assessment

**MC Format**

- a. Even if I do not know the content, I still can choose an answer.
- b. If I know a certain amount of the content, I still can answer.
- c. I need to know the requested content accurately in order to answer.

Reason: ________________________________________________________________

---

**OE Format**

1. Knowledge about the Content for Assessment

- a. Even if I do not know the content, I still can choose an answer.
- b. If I know a certain amount of the content, I still can answer.
- c. I need to know the requested content accurately in order to answer.

Reason: ________________________________________________________________

---

II. Discrimination Between Proficiency Levels

**MC format**

- Between high and high-intermediate levels
  - Yes
  - No
  - Reason: ________________________________________________________________

- Between high-intermediate and low levels
  - Yes
  - No
  - Reason: ________________________________________________________________

- Between high and low levels
  - Yes
  - No
  - Reason: ________________________________________________________________

**OE format**

- Between high and high-intermediate levels
  - Yes
  - No
  - Reason: ________________________________________________________________

- Between high-intermediate and low levels
  - Yes
  - No
  - Reason: ________________________________________________________________
### A Comparison of Reading Comprehension Tests: Multiple-Choice vs. Open-Ended

- Between high and low levels
  - Yes
  - No
  
  **Reason:**

<table>
<thead>
<tr>
<th>III. Appropriate Test Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Low level</strong></td>
</tr>
<tr>
<td>a. MC format</td>
</tr>
<tr>
<td>b. OE format</td>
</tr>
<tr>
<td>c. Mixed format</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Reason:</strong></th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th><strong>2. Intermediate Level</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>a. MC format</td>
</tr>
<tr>
<td>b. OE format</td>
</tr>
<tr>
<td>c. Mixed format</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Reason:</strong></th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th><strong>3. High Level</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>a. MC format</td>
</tr>
<tr>
<td>b. OE format</td>
</tr>
<tr>
<td>c. Mixed format</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Reason:</strong></th>
</tr>
</thead>
</table>

### IV. Personal Preference

What is your preferred test method? Why?

---

**Applicable levels:** secondary education, tertiary education

**Key words:** test method, test format effect, test validity

Myong Hee Ko  
Korea University  
1203-ho 102-dong  
Shinil Yootobil Plus Apartment  
461-bunji Howon-dong  
Uijeongbu-si, Kyonggi-do, Korea  
Email: myongheeko@yahoo.co.kr

Received in December, 2009  
Reviewed in January, 2010  
Revised version received in February, 2010