A Comparison of Reading Test Formats in High School Exams: Multiple Choice vs. Open-ended

Yunhee Lee
(Seoul National University)

Lee, Yunhee. (2012). A Comparison of Reading Test Formats in High School Exams: Multiple Choice vs. Open-ended. ***Language Research*** **48.2, 343-367.**

Multiple choice test (MC) and open-ended test (OE) are the most common test formats used in high school exams. To examine the reliability of high school exams, this study investigated the effect of the two different test formats on high school students' performance. In addition, the students' preference and perception towards MC and OE were also looked into. The experiment was designed to reflect the real testing situation of Korean high schools and conducted to 129 students in the 10th grade. The participants took either MC or OE and completed the survey so that their scores and answers were compared and analyzed. The results showed that the students got higher scores in MC than in OE regardless of their proficiency levels, and that they preferred MC to OE while perceiving OE to be more valid. The reliability of high school exams was discussed based on the results and the implications for designing the exams and interpreting the scores were addressed.

Keywords: multiple choice test, open-ended test, test reliability, test validity, test formats

1. Introduction

High school students take English exams at least twice a semester which mostly consist of reading tests and the exams are created by teachers of each high school. The test items are usually made based on the textbooks authorized by the government and the teachers tend to measure what they actually taught in their classes. For that reason, the test items are different between schools and the objectivity or validity of the exams are often questioned (Kim 2009, 2010).

High school exams are, in principle, criterion-referenced tests to determine whether or not students have understood the course content

and the objectives of the course are achieved. Thus, the nature of high school exams is quite different from that of language proficiency tests like TOEFL or nation-wide exams like KSAT. Nonetheless, every test should be reliable and valid and the high school exams are not an exception. The high school exams should inform how well the students have mastered the course contents and what specific domain needs to be supplemented.

To ensure the appropriateness and usefulness of the high school exams, the reliability of the exams is prerequisite (Bachman 1990). Bachman mentioned "The primary concerns in examining the reliability of test scores are first, to identify the different sources of error, and then to use the appropriate empirical procedures for estimating the effect of these sources of error on test scores." (p. 24) Among the potential sources of error, the effect of test formats which require different types of responses (i.e., selected response or constructed response) have received lot of attention from researchers and the empirical findings on this topic revealed mixed results. With respect to high school exams, the most common test formats are multiple-choice test (hereafter MC) and open-ended test (hereafter OE). Students are expected to read and select an answer in MC while they read and write an answer in OE. Previous studies proved that the different test formats of MC and OE affected test-takers' performance and that MC or OE scores included the effect of the test format itself as well as the test-takers' true scores. The effect of the test formats in high school exams, however, has gone uninvestigated so far and it urges an empirical research. In this respect, the current study focuses on the effect of two test formats (MC & OE) on high school students' performance and it also investigates high school students' perception and attitudes toward the formats. The goal of the research is not to determine which format is more desirable because there would be no error-free measurements or absolutely reliable tests (Bachman 1990, Bachman & Palmer 1996). Neither will high school teachers be using one of the two test formats as long as KSAT is based mostly on MC format and the Office of Education encourages the use of OE format in school exams. Rather, the main purpose of the study is to provide the relevant and useful information in constructing the high school exams and interpreting the test scores in terms of the effect of test formats and students' attitudes. Hopefully, this study will shed some light on the validation of high school exams.

To serve the purpose, this study compares the scores of MC and OE to investigate the effect of test formats in measuring students' achievement in high school. Besides, students' proficiency level is taken into consideration to reflect the reality of high school education where students are taught in separate classes according to their test scores. Lastly, this study examines how high school students perceive the validity and difficulty of MC and OE formats and which test format they prefer. Accordingly, the research questions are addressed as follows:

1. Do students perform differently on the different test formats (MC / OE)?
2. Do students at different proficiency levels perform on the tests differently?
3. How do students perceive the different test formats?

2. Review of Literature

2.1. The Effect of Test Formats: Multiple Choice vs. Open-ended

In investigating the effect of the test formats of MC and OE (or constructed-response test in a broader term), there have been two distinctive focuses in researches: (a) the validity of the different test formats (b) the test reliability deteriorated by test formats. In terms of the first issue, Campbell and Fiske (1959) defined validity as the agreement between two attempts to measure the same trait through maximally different methods. Traditionally, the agreement is checked out by correlation analysis. That is to say, two test scan be considered to be congeneric, or to measure the same trait only when the scores of one test perfectly correlates with those of the other. In fact, confirming the trait equivalence is the prerequsite in the studies investigating the test format effect. Otherwise, the disparity of scores between different test formats can be attributed to the different trait that the tests might meaure rather than to the effect of test formats. However, previous correlation studies on different test formats reported mixed results and it was hard to assume that MC and OE measure the same trait. Facing the problem, Rodriguez (2003) tried to synthesize previous findings by a meta-analysis based on 29 articles reporting corre-

lations. The results revealed that when items were stem-equivalent,[1] the corrected correlation between MC and OE approached unity, while the correlation tended to be lower when the stems were different. Therefore, it can be tentatively assumed that stem-equivalent MC and OE measure the same trait and the difference of scores between the two tests is the evidence to prove the effect of test formats rather than the difference of the construct measured by each test.

On the other hand, other researchers have focused on the second issue and looked into the effect of test formats which deteriorates reliability. According to Bachman (1990), reliability is concerned with the extent to which an individual's test performance is affected by the measurement error, or by the factor other than the true language ability. The test format of MC or OE is one of the test method characteristics affecting the test-takers' performance. Previous studies usually showed how significantly different the test scores were when measured by different test formats but their findings were not consistent. In relation to this complexity, In'nami and Koizumi (2009) did a meta-analysis as Rodriguez (2003) did. The study synthesized the findings of 56 articles dealing with the effect of MC and OE on L1 reading, L2 reading and L2 listening and concluded that the overall effect of MC and OE did not exist as for L2 reading. However, MC was significantly easier than OE in the following conditions; between-subjects design, random assignment, stem-equivalent test, or high L2 proficiency. In Korea, one of the conditions, stem-equivalent test, was tested empirically. Go (2010) did an experimental research based on a within-subjects design to find the effect of MC and OE formats. The study showed that the test formats affected undergragudates' reading performance. Specifically, MC was significanlty easier than OE even though both formats could differentiate the low and high level students. Based on the analyses and survey results, Go suggested that OE should be more appropriate for high-level students and MC, for low-level students in that high-level students were able to comprehend the text and hold information to produce when they were asked later on paper. It is questioned, however, whether the suggestions can be applied to a high school situation considering the following two aspects. First, the

1) A stem and options in MC are as follows (Hughes 2009, p. 75):
Enid has been here _____ half an hour. (stem)
A. during B. for C. while D. since (options)

experiment in Go's study was very different from the testing situation in high school. In the experiment, the reading text was taken away before the test was given to prevent the participants from using search-and-match strategies in MC and from copying the language of the passage in OE. In high school, in contrast, students are allowed to refer to the given text while answering the questions and search-and-match strategies or using the same language of the given passage is not considered undesirable. With respect to the effect of the condition where student are allowed to refer to the text, Davey and LaSasso (1984) reported that there was no significant difference between MC and OE under lookback condition while under no-lookback condition, MC scores exceeded OE scores. Second, the subjects of Go's study were undergraduates and their proficiency levels were quite different from those of high school students. With the age and level difference, it's doubtful that OE would be more appropriate for high-level high school students or MC, for low-level ones as it was so for undergraduates.

2.2. Other Moderating Variables Related to Test Formats

As discussed above, previous studies led to inconsistent conclusions. As for the reasons of the heterogeneous results, above-mentioned In'nami and Koizumi's (2009) study extracted 15 moderating variables. Among them, only four variables turned out to be significant: stem equivalency, between-subjects design, random assignment, and learners' L2 proficiency.

In terms of proficiency, In'nami and Koizumi (2009) defined high proficiency levels as the learners studying L2 for five or more semesters based on Norris and Ortega (2000), whose criteiron Wolf (1993) also used. In other studies, however, the distinction of proficiency levels was based on the obtained scores in an MC test (Shohamy 1984) or in a cloze test (Go 2010). Pointing out the dissimilarity of subjects and their levels between studies, many researchers were concerned about the lack of the generalizability and comparability of the findings (Wolf 1993, Norris & Ortega 2000, In'nami and Koizumi 2009). Specifically, J. F. Lee (1990, cited in Wolf 1993) indicated the difference between ESL and EFL learners. That is, most of the subjects in ESL studies could be considered intermediate and advanced levels while

those in EFL studies, beginning or at most intermediate levels in terms of the amount of L2 experience. J. F. Lee's point is very relevant in Korean educational situation, especially in high school testing situation. Korean high school students are being taught English as EFL and their English proficiency are considered beginning level in terms of the amount of L2 exposure. In reality, however, there exists proficiency difference within the beginning level, so students are often taught separately according to their test scores in intact graded-classes. None of the above-mentioned definitions are appropriate to describe these existing levels in Korean high school. To solve the problem in defining the levels of high school students, the current study uses the scores of a Nationwide Sample Test (hereafter NST) to incorporate the motive to generalize the findings of this study to Korean high school testing situation.

The next potential variable is the language of the questions and of the expected responses (Bachman 1990) and several studies reported its significant effect (Shohamy 1984, Wolf 1993, Cheng 2004). In terms of this language variable, the findings of studies were all consistent. That is, the participants did significantly better when the questions and expected answers were presented in L1 than in L2. Specifically, Shohamy (1984) discussed that the use of L1 in questions reduced the anxiety of the students and unneccesary source of difficulty and that presenting quesitons in L1 should be more natural for L2 learners in that they would utilize their L1 in processing L2 text. In relation to language in answering, Cheng (2004) insisted that freedom to choose the language of L1 or L2 would maximize the vailidity of the experiment, by which the effect of the language variable could be controlled. Agreeing to the suggestions, the current study also presents the questions in L1 and allows the participants to use either L1 or L2 in answering OE. By doing so, the paticipants are expected to demonstrate how well they understood the text without difficulty of understanding the questions or of producing the answers in L2.

Lastly, there is a possibility that MC and OE might affect the students' performance on individual test items. Currie and Chirammanee's (2010) study showed that the participants changed their answers in a grammar test according to the format of MC or OE. The present study examines this effect indireclty through a questionnaire by asking students if there are specific items to be easier or more difficult. If the

items the students report are different according to MC or OE, it can be suspected that there may exist some item types favoring a specific test format.

3. Method

3.1. Participants

This study involved 129 students in the 10th grade. All the students had taken NST in June, 2011. The students were divided into three groups according to their English scores of NST: high (HP), intermediate (IP), and low proficiency (LP). In the main study, half of each group (HP, IP, LP) would be randomly assigned to one of the two tests (MC or OE), which produced two subgroups in each level. The descriptive statistics of each group are shown in Table 1.

Table 1. Means and Standard Deviation of NST Scores

Level	Group	N	Mean	Std. Deviation
Whole	OE test	63	97.1	18.90
	MC test	66	97.5	16.55
High	OE test	17	124.1	12.95
	MC test	20	118.7	11.64
Intermediate	OE test	24	93.1	3.84
	MC test	23	94.4	4.28
Low	OE test	22	80.5	3.54
	MC test	23	82.3	3.53

To check the homogeneity of the subgroups, independent sample *t*-test was conducted. The result indicated that there was no significant difference between the subgroups in each level (t (35) = 1.342 in HP, t (45) = -1.104 in IP, t (43) = -1.667 in LP, $p < .05$).

3.2. Instrument and Pilot Study

3.2.1. Materials

Three reading texts were extracted from three different high school

textbooks which were authorized by Ministry of Education and Human Resources Development in 2002 and used in high schools until 2008. The selected passages were of general topics like a story about a survived baby with help of anonymous internet supporters, indoor air pollution, and cultural difference in conversation. These passages were not modified at all and their readability was checked out in terms of the Flesch-Kincaid Grade Level which was provided in Microsoft Word. The length of each text was 198, 200 and 179 and its grade level was 6.4, 6.8, 8.3, respectively, which suggested the third passage would be somewhat difficult compared to the other two passages.

Based on the selected passages, 22 test items were written into stem-equivalent MC and OE formats. The test items were created in order that they could be all passage-dependent and reflect different levels of understanding including questions asking implicit or explicit information and general or detailed information (Wolf 1993). All of the questions were written in Korean to make sure that the students fully understand the questions.

3.2.2. Pilot Study and Item Modification

The poliot study was conducted to 32 students, one class of 10^{th} graders in the same school to determine the familiarity of the text and the appropriateness of each item. The students were allowed to write their answers either in English or in Korean to minimize the effect of language in expected response as discussd above.

As a result, all the students reported that they were not familiar with any of the texts and some reported that the stem of one item (item number 5) was not clear and difficult to answer. The item fitness and consistency of the two tests were inverstigated using FACETS analysis. The results are shown in Figure 1.

In the figure, the third column exhibits the difficulty of the two test formats; OE was more difficult than MC. The last column displays the difficulty of the original 22 items; they spread along the continuum of the given range from -2 (easiest) to +1.5 (most difficult). On the other hand, Table 2 below presents the bias interaction between the items and the two test formats. It indicates that one item (number 9) out of 22 was not acceptable. In other words, as for item number 9, the students gained higher scores than expected in OE (p = .0134, $p < .05$), whereas they got lower scores than expected in MC

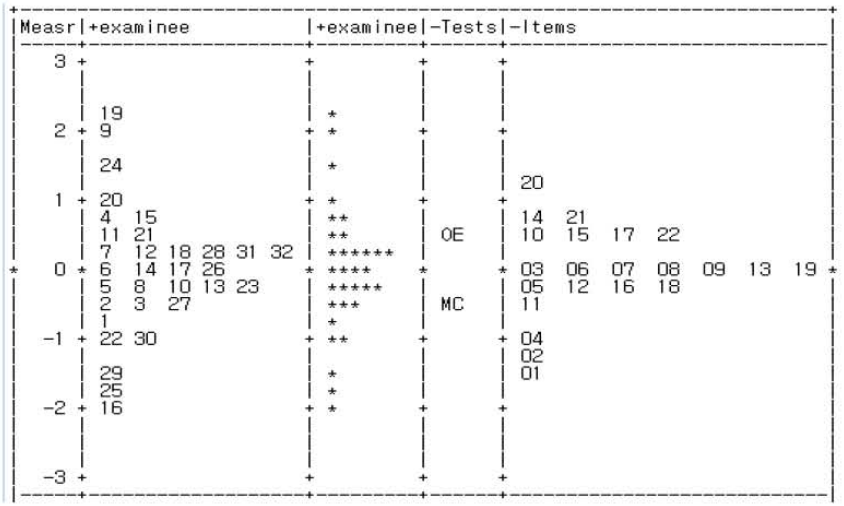

Figure 1. FACETS summary on examinee ability, test format difficulty, and item difficulty.

(p = .0159, p < .05). Based on the students' report and the statistical findings, two items (number 5 and 9) were removed for the sake of validity of the test. As a result, 20 items were kept. Besides, some options of MC test were modified based on the students' wrong answers given in OE to increase the plausibility of the options (Chon & Shin 2010, Currie & Chirammanee 2010). (See Appendix A and B)

Table 2. Bias Interaction Between Items and Test Formats

Obsvd Score	Exp. Score	Obsvd Count	Obs-Exp Average	Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Tests Sq	N	Te	Items measr	Nu	It	measr
11	17.9	32	-.21	1.03	.40	2.55	31	.0159	1.1	1.1	18	2	MC	-.39	9	09	.11
19	12.1	32	.21	-1.02	.39	-2.62	31	.0134	.8	.7	17	1	OE	.45	9	09	.11
15.7	15.7	32.0	.00	.00	.41	.00			1.0	1.0	Mean (Count: 44)						
5.4	4.8	.0	.07	.38	.04	.92			.2	.4	S.D. (Population)						
5.4	4.9	.0	.08	.38	.04	.94			.2	.4	S.D. (Sample)						

3.2.2. Survey Questions

To examine the effect of MC or OE on individual test items and to investigate 10th graders' perceived validity, difficulty and preference toward each format, five survey questions were developed by the researcher based on Go (2009) and Wolf (1993). (See Appendix C)

3.3. Procedure

The main study was conducted during regular class time. Two out of four intact classes were randomly given one of MC or OE. Time limit was not set to give the students plenty of time in answering and the students in OE test were told to write their answers in English or in Korean. After the test, the students completed the survey.

3.4. Analysis

The students' answers of ME and OE were scored by giving 1 point for each correct answer and 0 point for wrong one so that the perfect points were 20. In case of the OE questions, partial points were not given and some problematic answers were scored through the discussion of three raters who have taught English in secondary school at least for 7 years to increase the reliability. With respect to the items which the raters diagreed, the scoring was done corresponding to the agreement of the two raters. The scores of MC or OE, then, were compared using a two-way ANOVA and independent sample *t*-test.

4. Result

4.1. The Effect of Test Methods

To answer the first research question, the students' scores of MC and OE were compared. As shown in Table 3, the mean score of MC was higher than that of OE in the whole and in each level. On the other hand, the standard deviation of OE was larger than that of MC, which was more clearly seen in HP than in IP or LP.

To check whether the mean difference between OE and MC was significant, a two-way ANOVA was conducted by setting test formats and proficiency levels as variance between groups. As shown in Table 4, the mean difference between MC and OE ($F = 102.526$, $p = .000$) and the effect of proficiency ($F = 165.619$, $p = .000$) were significant. The consecutive post hoc analysis of Tukey confirmed that the mean differences between three levels were all significant in MC and OE.

Table 3. Means and Standard Deviation of MC and OE Tests

Level	Test format	N	Mean	Std. Deviation
Whole	OE	63	6.3	6.09
	MC	66	11.8	4.43
High	OE	17	14.8	3.20
	MC	20	16.7	1.45
Intermediate	OE	24	4.8	2.78
	MC	23	11.4	3.24
Low	OE	22	1.4	2.73
	MC	23	7.9	2.81

Table 4. Results of ANOVA for the MC/OE Tests

Source	Type III sum of squares	Df	Mean square	F	Sig.
Level	2548.419	2	1274.210	165.619	.000
Test format	788.794	1	788.794	102.526	.000
level * test	142.438	2	71.219	9.257	.000
Error	946.315	123	7.694		

a R Squared = .791 (Adjusted R Squared = .783)

On the other hand, the interaction between test formats and proficiency levels was also significant ($F = 9.257$, $p = .000$). The following figure shows the interaction more clearly.

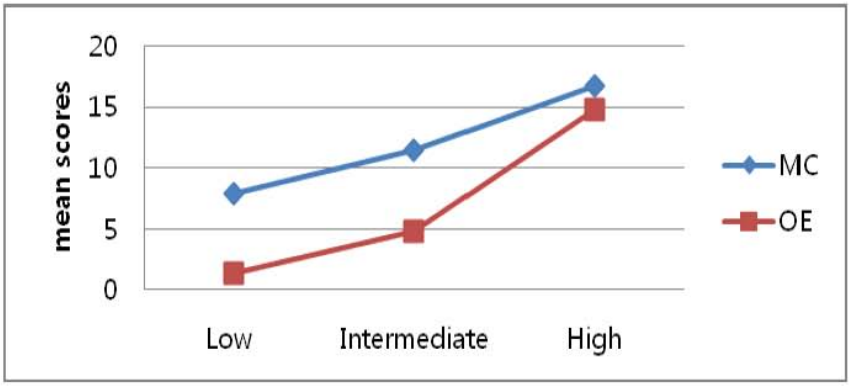

Figure 2. Interaction between test formats and proficiency levels.

The figure indicates no conflicting relationship between test formats and proficiency levels but the difference between MC and OE became smaller when the students' proficiency levels got higher.

In sum, it can be said that both MC and OE could divide the students into different levels but MC is significantly easier than OE.

4.2. The Effect of Proficiency

The second research question concerned how the students at different proficiency levels performed in MC or OE. In each level of proficiency, the students got higher scores in MC than OE as shown in Table 3 above. In addition, the effect of proficiency turned out to be significant as already shown in Table 4. However, the result just indicates that both MC and OE discriminated three proficiency levels and they did not show whether the mean difference between MC and OE in each level was significant. Accordingly, independent sample *t*-test was conducted again. As shown in Table 5, there was significant difference between MC and OE in all levels (t (35) = -2.352 in HP, t (45) = -7.494 in IP, t (43) = -7.803 in LP, $p < .05$). That is, MC was significantly easier than OE regardless of the students' levels.

Table 5. Results of Independent Sample *T*-test for the MC/OE Tests in Each Proficiency Level

	T	Df	Sig. (2-tailed)	Mean difference	Std. error difference
High	-2.352	35	.024	-1.876	.797
Middle	-7.494	45	.000	-6.603	.881
Low	-7.803	43	.000	-6.458	.827

4.3. The Posttest Survey

The third research question was about the students' perception toward MC and OE. The questions were how the students perceived the validity (question #3), difficulty (question #4) of the two test formats and which format they prefer (question #5). It was also investigated whether one of the test formats was favorable to specific item types as discussed earlier in Literature Review part. For the purpose, the students were asked to write the easiest or the most difficult test items

with reasons, if any (questions #1 and #2).

The analysis of the answers to question #1 and #2 revealed that there were no specific types of test items that favored a test format. The determinant factors affecting the item difficulty were the readability of the text (54.60%) and the degree of implicitness of test items (41.06%) rather than the test formats. That is, the students responded that they had harder time in answering the items related to more difficult text and in dealing with questions asking for implicit information. In terms of test formats, only 7 participants (0.03%) reported that some questions in MC were more difficult because of the plausible distractors.

On the other hand, the answers to question #3, #4, and #5 revealed the overall perceptions of the students towards MC or OE and they are arranged with the responding rates in Table 6. The answers to question #3 showed that OE was perceived to be a more valid test than MC in that OE prevented test-takers from guessing randomly. In other words, there would be no way to get correct answers unless test-takers comprehend the text in OE. In question #4, the students reported that OE was more difficult than MC. The major reason was that they felt unarmed when they could not use test-taking techniques available in MC. A few students also mentioned that OE was demanding in that test-takers needed to comprehend, organized their thought and produced the answer in accurate words. The responses in question #5 revealed the students' clear preference toward MC and their eagerness to improve scores even by using the test-taking technique of guessing. A few students responded that they believed in the objective scoring of MC or that the options in MC helped them to comprehend the text and eventually motivated them to read.

On the other hand, some interesting contrasts were found when the students' proficiency level was considered. In terms of the item difficulty, the students in HP were more concerned about the degree of implicitness of test items(63.41%) rather than the readability of the text (24.39%). In case of LP, the length of the text was another factor to explain the difficulty of the test item (6.9%). In questions #4 and #5, the students in IP and LP mentioned they could not get any clues or help in comprehending the text in OE while they could do in MC, which were not reported in HP. Lastly, relatively higher rate of HP favored OE (31.25%) because they wanted to get accurate information about their reading ability and at the same time to improve their skill

Table 6. Results of Posttest Survey (question 3, 4 & 5)

Content of the question	Response rates and accounts
Perceived validity of the two test formats	OE: 82.83% (82 out of 99) - It prevents from guessing randomly (71.95%). - It requires deep process of thinking. - It requires thorough reading.
Perceived difficulty of the two test formats	OE: 86.18% (106 out of 123) - It prevents from using a guessing technique (66.98%). - It requires logical thinking and accurate production.
Preference for the two test formats	MC: 83.19% (99 out of 119) - They want to increase scores by using a guessing technique (72.73%). - They trust the objective scoring. - Options give help to comprehend the text, which motivates to read the text.

afterwards. They also liked OE in that they could express their thought and opinions freely.

5. Discussion

The first research question was whether there existed the effect of test formats in high school exams. The findings of the current study revealed that high school students performed differently in MC and OE and MC was easier than OE for them. Therefore, it can be said that the students would get higher scores when the questions are asked in MC format. There are three possible explanations for the results. First of all, the type of response itself can make the difference. In contrast with MC, when constructing OE answers, students need to consider the other language factors like vocabulary, grammar and length (phrase or sentence). This is true even when they are allowed to answer in L1, which was shown in some students' reports saying they had hard time articulating the answers in OE. Producing answers in OE means not only demonstrating the comprehension of the text but also thinking logically, organizing the thought, choosing appropriate vocabulary and writing in grammatically fitting forms. In other words,

producing in OE is more difficult than selecting in MC in terms of the processing time and the mental effort. This interpretation is aligned with the discussions of previous studies and many researchers mentioned that MC and OE were different tasks in terms of the demanding degree of mental processing (Shohamy 1984, Wolf 1993, Go 2010).

Another possible account for the better performance in MC is the utilization of guessing. As Cohen's (1984) study showed, there is a high possibility that test-takers in MC would get some points from guessing based on the given options, which are unavailable in OE. Likewise, one student of the present study marked option #2 in all questions of MC and got some points out of them. In addition, many students acknowledged in the survey that they selected an option randomly when they did not know the answer in MC. The analysis of total scores of MC, however, could not capture whether the student really knew the answer or just guessed out of luck. In fact, this problem is an inevitable and inherent characteristic of MC (Hughes 2009). However, two types of guessing need to be differentiated here; guessing as a random selection and as a strategy. The former type of guessing is the one discussed earlier and is one of the factors to deteriorate the validity of MC format because the test-takers can get some points without reading the text or only by seeing the questions and options. The latter type of guess, on the other hand, is controversial just to be seen as a deteriorating factor. In the process of guessing, test-takers would recognize the given options and compare and contrast with each other by frequently referring to the text. This process may reinforce the reading process and require a kind of mental effort. In the study of Currie and Chirammanee (2010), this issue was addressed and some scholars mentioned "*cued recall* and *guessing* might also be interpreted as indicating the use of a degree of knowledge by the test taker" (p. 484). Nevertheless, one sure thing is that test-takers surely have more chances to get some points by selecting guessed answers than by writing uncertain answers and whether guessing is a test-taking strategy or one of processes representing reading comprehension remains to be discussed more.

The last possible explanation is related with the different amount of information between MC and OE. In MC, test-takers may understand the text better or at least get the gist of the given text by looking into the options. Cohen's (1984) study showed the participants' strategy to

read the options first before actually reading the text in an MC test. In the present study, the students at lower level also reported that they got some cues in options of MC to comprehend the text. In MC of this study, six to seven questions were created per a text and each question included five options. The amount of information in MC was sharply contrasted to none given in OE, and the available information might contribute to better comprehension. Comprehension is a complex process including inferences and the inferential process can be based on internal factor like the text itself or on external factors like background knowledge. In this regard, the options in MC can help to activate test-takers' content schema and enable them to predict the main idea of the text.

The second research question was to investigate the effect of proficiency levels in relation to MC or OE. The findings showed the students' performance was affected by test formats regardless of their proficiency. Even though the proficiency levels in the current study were defined carefully to reflect the real high school situation, the results were congruous with previous studies (Wolf 1993, Go 2010). These consistent findings on the effect of proficiency may lead to a persuasive conclusion that high school students may get higher scores when being tested by MC than by OE whether they are at an advanced or beginning level. On the other hand, Figure 2 showed that the mean difference between MC and OE was smaller in HP than that in IP and LP and Table 3 indicated that the difference of standard deviation between MC and OE was greater in HP than in the other groups. These results imply that both MC and OE are appropriate for the high proficiency group, but OE can subdivide them better. Thus, it can be said that for high proficiency group, OE works better than MC in terms of its assessment function and its requisite for a deeper mental processing.

The third research question was addressed to see the students' perceptions and preference. Apart from comparing the two test formats, two questions were asked in the survey to investigate whether the participants took advantages or disadvantage of a test format in terms of individual test items, if any. The results showed that easy or difficult items were mostly irrelevant to test formats. That is, the students reported that the more difficult the text was and the more implicit the question was, the harder they got to answer. In fact, the readability

index of the Flesch-Kincaid Grade Level initially predicted the difficulty and the survey results just checked it. The readability formulas are calculated based on two or more directly measurable characteristics of the text such as the number of letters per word and the number of words per sentence. Even though the Flesch-Kincaid Grade Level was developed for native English readers, the current study proved that it was also valid for EFL learners (Greenfield 2004). Though this study did not examine whether the students actually got correct in the items they perceived to be easier or vice versa, the survey results indicated that readability or the degree of implicitness of items rather than test formats affected the perceived difficulty of individual items.

When directly comparing MC and OE in terms of their validity, difficulty and preference, the students chose MC rather than OE to be assessed by because the former was easier. However, they perceived OE to be a more valid test because they thought a reading test should elicit a deep mental processing of thinking or a thorough reading while blocking the usage of guessing strategy. Nonetheless, the urgent issue for them was not to measure their ability more accurately, but to increase their scores. This result is consistent with that of Go's (2010) study. In contrast, many students in HP showed aspiration to be assessed accurately and to improve their skills based on the test result. They also valued the freedom to demonstrate their comprehension in OE. These findings may imply that most of high school students take a reading test driven by extrinsic motivation but some highly-achieved students tend to be oriented to intrinsic motivation.

6. Conclusion

As many experts in testing area mentioned, the "perfect" reading testing format does not exist. The major reason is due to the complexity of the construct of reading comprehension itself. As Wolf (1993) mentioned, comprehension cannot be defined as "one true comprehension" but as a "range" of comprehension (p. 473). Therefore, no single test can measure the whole range of comprehension. Nevertheless, the test developers and users should not give up looking for a better test format in terms of reliability and validity. The effort to identify and estimate various factors affecting test performance should be made

constantly. Hopefully, the findings of the present study would contribute some to the ongoing research on this area.

What the present study concerned was the effect of different test formats and proficiency on high school students' performance. The findings showed MC was significantly easier than OE and this effect was consistent regardless of the students' proficiency. These results are in keeping with those of the previous studies and it can generally be insisted that there exists a test format effect when students are tested in MC or OE. On the other hand, high school students preferred MC to OE because they wanted to get higher scores while some students of high proficiency wanted to be assessed accurately by OE.

Several implications can be discussed based on the findings. First, when developing a reading test in high school, teachers can use a test format according to their intent. If teachers want to discriminate students better or to motivate highly-achieved students, they are advised to write questions in OE format because it subdivides the students and may require deeper levels of mental processing including reading, thinking and producing. This is so true especially for high proficiency students. On the other hand, if teachers are not interested in the difference of individual student' performance but want to make a decision on the achievement of the whole students or to encourage the poorly-achieved students by giving them confidence, MC format is more appropriate in that it is practical in administering and scoring and that it generates higher scores. Second, when interpreting the scores of a reading test, teachers and students need to take the effect of the test formats into consideration. That is, the high scores in MC may not necessarily mean the success of the teachers' instruction or the progress of the students' achievement or the lower scores in OE may not mean the failure of achieving the objectives of the course or lacking of students' effort. Especially in a test combining MC and OE formats which is common in high school exams, the scores should be interpreted even more carefully in this regard. Lastly, in terms of the backwash effect, the students should be given more opportunities to get used to an OE format in a class through various activities. One of the reasons that the students prefer MC is that they got used to it and felt comfortable with it (Go 2010). However, the students, ultimately, should be encouraged to demonstrate what they understood in appropriate words and to express their opinions based on what they read.

Therefore, the students should have plenty of exercises to get familiar with OE format before actually get tested by OE.

For further research, it would be worthy to investigate the effect of MC and OE in individual items. In the current study, MC and OE were assumed to measure the same trait in that they were stem-equivalent as suggested in the review of literature. To testify the assumption, the participants' actual performance in each test item of MC and OE should be directly compared as in Currie and Chirammanee (2010) did. In addition, the strategies the participants use in OE need to be introspected in the future study since there are fewer studies on OE strategies compared to those on MC. If test-taking strategies are found different between MC and OE, the effect of MC and OE can be discussed qualitatively. As these kind of empirical studies are accumulated, we can get a clearer sight and a power to measure the true ability we want to focus on.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Shanghai: Oxford university.

Bachman, L. F. and Palmer, A. S. (1996). *Language testing in practice : designing and developing useful language tests*. New York: Oxford University Press.

Campbell, D. T. and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56.2, 81-105.

Cheng, H.-F. (2004). A Comparison of Multiple-Choice and Open-Ended Response Formats for the Assessment of Listening Proficiency in English. *Foreign Language Annals* 37.4, 544-555.

Chon, Y.-F. and Shin, T. (2010). Item Difficulty Predictors of a Multiple-choice Reading Test. *English Teaching* 65.4, 257-282.

Cohen, A. (1984). On taking language tests: what the students report. *Language Testing* 1.1, 70-81.

Currie, M. and Chirammanee, T. (2010). The effect of the mutiple-choice item format on the measurement of knowledge of language structure. *Language Testing* 27.4, 471-491.

Davey, B. and LaSasso, C. (1984). The interaction of reader and task factors in the assessment of reading comprehension. *Journal of Experimental Education* 52, 199-206.

Go, M.-H. (2010). A Comparison of Reading Comprehension Tests: Multiple-

Choice vs. Open-Ended. *English Teaching* 65.1, 137-159.
Greenfield, J. (2004). Readibility formulas for EFL. *JALT Journal* 26.1, 5-21.
Hughes, A. (2009). *Testing for language teachers*. Cambridge: Cambridge University Press.
In'nami, Y. and Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing* 26.2, 219-244.
Kim, S.-A. (2009). An Analysis of the Multiple-Choice Test Items Constructed by Middle School English Teachers. *Applied Liguistics* 25.2, 143-169.
Kim, S.-A. (2010). An Analysis of the Test Items on Text Structures in High School English Tests. *Korean Journal of Applied Linguistics* 26.4, 251-274.
Norris, J. M. and Ortega, L. (2000). Effectiveness of L2 Instruction: A Research Synthesis and Quantitative Meta-analysis. *Language Learning* 50.3, 417-528.
Rodriguez, M. C. (2003). Construct Equivalence of Multiple-Choice and Constructed-Response Items: A Random Effects Synthesis of Correlations. *Journal of Educational Measurement* 40.2, 163-184.
Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing* 1.2, 147-170.
Wolf, D. F. (1993). A Comparison of Assessment Tasks Used to Measure FL Reading Comprehension. *The Modern Language Journal* 77.4, 473-489.

Appendix A.

Multiple-Choice Test (Sample)

When Oakland High School in California was moved into a new building the students and teachers noticed a strong smell. Then almost half of the students began to have headaches and sore throats and suffer from fatigue. These three symptoms disappeared on weekends. The reason was ⓐa mystery. Experts came to investigate and find the cause of the sickness. Finally, they discovered that the air in the building was not safe to breathe. They were surprised to find that the cause was the shelves in the school library! These shelves were made of particle board; that is, an inexpensive kind of board made of very small pieces of wood held together with a chemical. This is just one example of a modern problem that is most common in cities; indoor air pollution.

A recent study reached a surprising conclusion; ⓑthe air inside building is almost always two to five times more polluted than the air outside the buildings. This is true even in buildings that are close to factories that produce chemicals. Better ⓒventilation – a system for moving fresh air- can cut indoor pollution to a safe level, but lack of ventilation is seldom the main cause of the problem.

* 위 글을 읽고, 각 질문에 대한 답을 고르시오.

1. 위 글에서 밑줄 친 ⓐa mystery의 구체적인 내용은 무엇인가?
   ① 학생들이 새 건물로 이사 후 아프기 시작했다.
   ② 아픈 증상이 세 가지나 나타났다.
   ③ 학생들의 증상이 나타났다 사라지기를 반복했다.
   ④ 새 건물에서 이상한 냄새가 났다.
   ⑤ 아팠던 학생들도 학교에 오면 증상이 완화되었다.

2. 전문가들에 의하면 Oakland 고등학교 학생들이 아팠던 이유는 무엇인가?
   ① 새 건물 내 나쁜 공기
   ② 오래된 가구에서 나오는 화학 물질
   ③ 노후된 건물 내 시설들
   ④ 학생들의 먹는 음식에서 발견된 독소

⑤ 운동 부족

3. Oakland 고등학교 도서관에 있었던 책장의 문제는 무엇인가?
① 학생들이 사용하기에는 값비싼 고가의 물건이다.
② 책장을 만드는데 많은 나무가 필요해서 자연을 훼손시킨다.
③ 책장에서 나오는 나무 부스러기들 때문에 주변이 더러워진다.
④ 튼튼하지 못해서 많은 책을 꼽아놓을 수 없다.
⑤ 값싼 재질로 만들어져서 실내 공기를 오염시킨다.

4. 위 글에서 밑줄 친 ⓑ의 의미는 무엇인가?
① 건물 내부는 건물 외부에 비해 더 지저분하다.
② 건물 내 공기 오염은 건물 외부에 비해 더 심각하다.
③ 건물 내 공기 양은 건물 외부 공기에 비해 2～5배 더 많다.
④ 건물 내 공기는 항상 건물 외부의 공기에 의해 오염된다.
⑤ 건물 내 공기는 항상 2～5회 정도 환기시켜야 한다.

5. 위의 밑줄 친 ⓒventilation의 의미는 무엇인가?
① 공기를 환기시키는 것
② 공기 청정기를 이용하는 것
③ 산소를 더 많이 공급하는 것
④ 건물 내부에 더 많은 식물을 심는 것
⑤ 냉난방 장치를 이용하는 것

6. 위 글의 주제를 가장 잘 나타낸 세 단어의 어구를 본문에서 찾은 것은?
① Oakland High School
② moving fresh air
③ a modern problem
④ indoor air pollution
⑤ lack of ventilation

Appendix B

Open-ended Test (Sample)

When Oakland High School in California was moved into a new building the students and teachers noticed a strong smell. Then almost half of the students began to have headaches and sore throats and suffer from fatigue. These three symptoms disappeared on weekends. The reason was ⓐa mystery. Experts came to investigate and find the cause of the sickness. Finally, they discovered that the air in the building was not safe to breathe. They were surprised to find that the cause was the shelves in the school lirary! These shelves were made of particle board; that is, an inexpensive kind of board made of very small pieces of wood held together with a chemical. This is just one example of a modern problem that is most common in cities; indoor air pollution.

A recent study reached a surprising conclusion; ⓑthe air inside building is almost always two to five times more polluted than the air outside the buildings. This is true even in buildings that are close to factories that produce chemicals. Better ⓒventilation – a system for moving fresh air- can cut indoor pollution to a safe level, but lack of ventilation is seldom the main cause of the problem.

* 위 글을 읽고, 각 질문에 대한 답을 한글이나 영어로 쓰시오.

1. 위 글에서 밑줄 친 ⓐa mystery의 구체적인 내용은 무엇인가?

2. 전문가들에 의하면 Oakland 고등학교 학생들이 아팠던 이유는 무엇인가?

3. Oakland 고등학교 도서관에 있었던 책장의 문제는 무엇인가?

4. 위 글에서 밑줄 친 ⓑ의 의미는 무엇인가?

5. 위의 밑줄 친 ⓒventilation의 의미는 무엇인가?

6. 위 글의 주제를 가장 잘 나타낸 세 단어의 어구를 본문에서 찾아 쓰시오.

Appendix C.

Questionnaire

1. 위의 문제 중에 몇 번 문제가 쉬웠나요? 왜 쉬웠는지 그 이유를 가능한 자세히 써 주세요. (쉬운 문제가 많은 경우 그 문제 번호를 모두 써 주세요.)

2. 위의 문제 중에 몇 번 문제가 어려웠나요? 왜 어려웠는지 그 이유를 가능한 자세히 써 주세요(어려운 문제가 많은 경우 그 문제 번호를 모두 써 주세요).

3. 다음과 같은 객관식/주관식 문제 유형 중 어느 것이 더 정확하게 학생들의 영어 읽기 능력을 평가할 수 있을 것이라 생각하나요? 그 이유는 무엇인가요?

① 객관식 문제의 예

다음 글을 읽고, 물음에 답하시오.

> In a Western-style conversation, emphasis is placed on interaction between speakers. For example,...(중략)

17. 일본식 대화에서 주의해야 할 점은 무엇인가?
    ① 상대방이 말하는 도중에 끼어들지 말아야 한다.
    ② 상대방의 말이 끝난 후 즉시 반응해야 한다.
    ③ 상대방의 말에 대해 자신의 의견을 제시해야 한다.
    ④ 상대방이 말할 때 시선을 맞추어야 한다.
    ⑤ 상대방의 말에 전적으로 동의해야 한다.

② 주관식 문제의 예

다음 글을 읽고, 물음에 한글이나 영어로 답하시오.

> In a Western-style conversation, emphasis is placed on interaction between speakers. For example,...(중략)

17. 일본식 대화에서 주의해야 할 점은 무엇인가?

4. 위 3번의 주관식이나 객관식 문제 유형 중 어느 것이 더 어렵다고 생각하나요? 그 이유는 무엇인가요?

5. 여러분이 중간고사나 기말고사의 영어 시험을 본다고 할 때, 위 3번의 주관식이나 객관식 문제 유형 중 무엇으로 평가를 받고 싶나요? 그 이유는 무엇인가요?

Lee Yunhee
Department of Foreign Language Education
Graduate School of Seoul National University
1 Gwanak-ro Gwanak-gu, Seoul 151-742, Korea
E-mail: daek7327@snu.ac.kr

Received : July 5, 2012
Revised version received: July 30, 2012
Accepted: August 10, 2012