English Teaching, Vol. 71, No. 4, Winter 2016
DOI: 10.15858/engtea.71.4.201612.27

The Effects of Presentation Mode and Item Type on L2 Learners' Listening Test Performance and Perception

Soohye Yeom
(Seoul National University)

Yeom, Soohye. (2016). The effects of presentation mode and item type on L2 learners' listening test performance and perception. *English Teaching, 71*(4), 27-54.

This study compared aural and written modes of presentation for the two item types, to explore the effects of question/option presentation mode and item type on EFL learners' listening comprehension performance and their perception. One hundred and fifteen Korean college students who were divided into three different proficiency groups participated in the study. The participants took a listening test which consisted of dialogue-completion and Q&A multiple-choice items in the aural and written modes, followed by a survey on their perceptions, and a stimulated recall interview. The results showed that the least proficient group was more critically affected by the mode than the other two groups. The least proficient group performed significantly better in the written mode than in the aural mode, while they received similar scores on the two item types. The major factors that caused the discrepancy among the groups were memory capacity in the aural mode and reading ability in the written mode. The implications and suggestions on listening test development are discussed.

Key words: second language listening, multiple-choice questions, test characteristics, test methods, presentation mode, item type

1. INTRODUCTION

The extent to which testing methods affect the performance of test-takers has been an important issue in developing language tests and interpreting the scores obtained from the tests (Bachman, 1990). This is mainly due to the strong influence of the test method on the test validity. Since tests seek to measure specific constructs, the degree to which a test can be considered valid depends on how aptly it can measure the constructs. However, different measurement characteristics can invite different test-taking processes, and this can change

the constructs that are measured through the test, which thus can yield varying outcomes.

Moreover, testing second language (L2) is even more complicated, inviting various factors that affect the test-takers' comprehension and performance. For example, processing a second language requires better memory of listeners than processing their first language (Cook, 2013; Ohata, 2006). In this case, an L2 listening test may measure test-takers' memory capacity in addition to their listening competence.

For assessing listening comprehension, one of the most frequently used testing methods is multiple-choice question (MCQ) items. MCQ tests can take various forms depending on the elements of the items, such as the type of listening stimuli and the way to present questions or options. Although several previous studies revealed what factors may affect the difficulty of listening MCQ tests (Brindley & Slatyer, 2002; Buck, 1990, 2001; Freedle & Kostin, 1996; Nissan, DeVincenzi, & Tang, 1995), not many have focused on each characteristic and the extent to which it impacts the test-takers performance.

Among various factors that affect test-takers' listening test performance, the present study focuses on the effects of two factors: the question/option presentation modes and item types. Making decisions on whether to use aural or written mode of question/option presentation is an important issue when developing a listening test, in that it is directly related to construct validity. Bachman (1990) emphasized that "in examining the effects of test method facets on language test scores, we are also testing hypotheses that are relevant to construct validity" (p. 258). The question/option presentation mode is one facet of the test method, and depending on the mode, other factors which are not directly relevant to the target constructs can influence the results. Therefore, to measure what is intended to be measured in the listening test, the presentation mode needs to be considered in development process, and construct-irrelevant method variance (Messick, 1996) should be minimized.

Construct-irrelevant factors of the written mode include reading ability, lexical attractiveness (Freedle & Fellbaum, 1987; Freedle & Kostin, 1996, 1999) and uninformed guessing (Wu, 1998). Chang and Read (2013) and Yanagawa and Green (2008) raised the possibility of the written mode negatively affecting the validity of the listening test, based on their results that test-takers with lower listening proficiency performed significantly better with the written form. They concluded that if lower level test-takers did well because of the written questions and options, a construct-irrelevant factor, the reading ability, might have been measured in the written mode. This could make the written mode less valid for a listening test. Aural mode, on the other hand, requires short-term memory capacity, which is known to be particularly limited for L2 learners (Cook, 2013; Ohata, 2006). It is also found that the aural mode increases test-takers' anxiety which could affect their performance (Buck, 1991; Chang & Read, 2006).

To investigate the influence of item type, the two most widely-used MCQ item types

were chosen to be examined for the present study: dialogue-completion and question-and-answer (Q&A) item type. The item type is one of the test method characteristics, and it has been considered to have effect on the test-takers' test performance (Berne, 1992; Wolf, 1993). Although both the dialogue-completion items and the Q&A items are MCQ items and both aim to measure the test-takers' listening comprehension, the two item types assess their listening ability in different ways. The two types involve different listening skills and this can differently affect the test-takers' performance. Moreover, there is a possibility of dissimilar effects when these two types are combined with the two question/option presentation modes, aural or written.

In addition, proficiency levels of test-takers are taken into considerations to see if different proficiency groups respond to the mode and item type differentially. Previous studies indicated that the test performance was differentially affected by the test method characteristics depending on the test-takers' proficiency levels, and mixed results were found particularly for the lower-level test-takers (e.g. Chang & Read, 2013; Yanagawa & Green, 2008). Therefore, in this study, how the lower-level test-takers react to different presentation modes and item types are closely examined compared to the performance of higher-level test-takers.

2. RESEARCH BACKGROUND

2.1. Modes of Question and Option Presentation

Although previous research generally agreed that providing questions or options in written form has a great effect on test-takers' performance (e.g. Chang, 2008; Iimura, 2010; Wu, 1998; Yanagawa & Green, 2008), they have yielded mixed results on the direction of its influence. Studies that showed positive effects of the written mode explained that previewing questions in the written mode, which was not possible in the aural mode, was beneficial for test-takers, since it motivated them by providing contextual information and relevant clues of the listening input and letting them employ metacognitive strategies such as goal setting and planning (Berne, 1995; Buck, 1991; Iimura, 2010; Yanagawa & Green, 2008).

Ur (1984) and Weir (1993), however, claimed that written questions and options distract test-takers' attention on listening input, since providing item stems and options in written form requires test-takers' reading ability. It was also discovered that written answer options invites uninformed guessing (Wu, 1998) and increases lexical attractiveness (Freedle & Kostin, 1996, 1999; Yanagawa & Green, 2008), letting listeners fall back on a lexical matching strategy. In other words, when the test-takers were provided with answer options

in written mode, they tended to just match words from the listening stimuli to the written options, without understanding their meaning.

In addition to the lexical matching strategy, Yanagawa and Green (2008) and Hemmati and Ghaderi (2014) also pointed out that the written answer options "provides contradictory cues and complicating planning strategies" (Yanagawa & Green, 2008, p. 110). This point is closely related to the authenticity of the test, which implies that the test-takers' cognitive processes should also be comparable to the cognitive processes that listeners employ in non-assessment situations. Since test-takers' cognitive processes in listening tests is influenced by the characteristics of test methods (Bachman & Palmer, 1996; Weir, 2005), providing the written answer options can undermine the validity of a listening test, as it may not reflect an authentic listening process.

The associations between the presentation modes and the test-takers' proficiency levels have been found in several previous studies (Buck, 1991; Chang, 2005; Chang & Read, 2013; Underwood, 1989; Wu, 1998; Yanagawa & Green, 2008). Although higher level L2 listeners' performance did not show a significant difference between the aural and written modes (Yanagawa & Green, 2008) or was even better in the aural mode (Chang & Read, 2013), the majority of test-takers' perceived that items in written form were easier than those in the aural form (Chang & Read, 2006). Chang and Read (2013) reported that 53% of their higher proficiency level students responded that they did not have difficulties in reading and listening at the same time. They also discovered that in the written mode, as soon as the test-takers heard the answer that they thought correct, they would stop listening, move on to the next item, and read it beforehand. In the aural mode, however, it was impossible for them to move on to the next item, so they sometimes lost attention.

It is notable that the effect of presentation modes on the performance of lower level test-takers has been controversial in previous studies (Chang, 2005; Chang & Read, 2013; Underwood, 1989; Wu, 1998; Yanagawa & Green, 2008). Compared to higher level listeners, it was clear that they were not good at forming anticipations of the input and performed more uninformed guessing, which means that they did not benefit as much as higher level test-takers from the written mode (Chang, 2005; Wu, 1998). Still, Chang and Read (2013) reported that the lower level test-takers performed significantly better with the written mode. Yanagawa and Green (2008) also stated that less proficient L2 learners were more disadvantaged when they could not access the questions in advance. They were less able to build a meaningful representation of a situation from the input without the support of the question. These findings were in line with that of Underwood (1989), who suggested that the written questions were helpful for lower level students.

The relatively limited short-term memory capacity of lower level test-takers might also account for their higher scores in the written mode than in the aural mode. Chang and Read (2013) explained that students from the lower level group were less able to hold the

question and options in short-term memory when they were presented in the aural mode. Some of the students reported in their post-test discussion that they sometimes just guessed randomly, because they forgot the answer options which were given aurally.

Buck (1991) and Chang and Read (2006), on the other hand, revealed that lower level test-takers did not benefit by the written mode. Unlike higher level test-takers who successfully used written questions to get the idea of what to listen for so that they could focus on the key words, lower level test-takers often failed to recognize the topic and key terms in the item stem and the answer options that were provided in written form. Lower level students in Chang and Read (2013) also reported in their interview that long options in written form caused them to give up reading, or if they could not finish reading, they would just guess. They also tended to depend much on word recognition – being readily diverted by distracters with words that match the recording, or confused by the use of negatives. They often considered the written questions and options as a source of background knowledge about the recording and made wrong assumptions. Regardless of their poorer performance in the written mode, however, they preferred the written form to the aural one.

In summary, previous studies had mixed results in the direction of the effects of the question/option presentation mode, particularly regarding the written answer options. Although the written mode provided test-takers with contextual information about the listening stimuli before listening, it allowed them to just match some words from the listening to choose the answer and the process did not reflect the real-life listening very well. The effect of the presentation mode was also affected by the proficiency level of the test-takers. While higher level test-takers were less influenced by mode difference, its influence on the lower level test-takers' test-taking process and performance was more critical.

Most aforementioned studies compared the aural and the written mode mainly focusing on the written mode's previewing effect, because in their studies, the test-takers could not access the question until they listen to the passage in the aural mode items. However, previewing questions is not a distinct characteristic of the written mode and can also be applied to the aural mode by inserting the question before the listening stimuli.

Therefore, in the current study, questions are presented before the listening stimuli in the aural mode to balance the two modes regarding the effect of previewing questions. The present study aims to compare and contrast two modes concentrating on the fundamental difference between the two modes, reading vs. listening, and discuss its implication on developing a valid listening test. In addition, the results will be discussed separately by test-takers' proficiency levels in order to add more explanation to the previous studies' relatively contradictory results on lower level test-takers' performance.

2.2. Item Types

Item type, as a characteristic of the test method, is also considered to have influence on the performance of test-takers (Berne, 1992; Wolf, 1993). Previous studies on listening comprehension test item types mostly focused on comparing different response formats (Berne, 1992; Cheng, 2004; Hansen & Jensen, 1994). The item types that were frequently examined were multiple-choice, cloze or open-ended questions. For example, as a part of her research on the role of different factors in L2 listening comprehension assessment, Berne (1992) compared multiple-choice, open-ended, and cloze test scores of university students who are native speakers of English studying Spanish as a foreign language. She found that the test item type significantly affected test-takers' L2 listening comprehension performance. Participants who received the multiple-choice items scored higher than those who received either the open-ended or cloze items. She attributed this result to the different skills that each item type is requiring, which was also suggested by Shohamy (1984) and Wolf (1993). For the multiple-choice items, test-takers only had to recognize the correct response, whereas the open-ended and cloze items required them to retrieve and produce the correct response.

In short, research on the effect of item types has been mainly on the comparison between different response formats, such as multiple-choice and open-ended. The discrepancy in the test results among the formats were reported to be largely due to the different skills that they are requiring. Although all the formats intended to measure the same construct, the listening ability, different formats could cause other factors to affect the performance. This suggests the need for further research on different item types within one response format, because even the same response format has several different item types that measure the same construct but require different skills. The current study specifically focuses on two different listening comprehension multiple-choice item types, dialogue-completion and Q&A, which are the two most frequently used item types for the multiple-choice listening comprehension questions. These two item types are developed to measure the same construct, the listening ability, but require test-takers of different skills. The dialogue-completion type asks the test-takers to complete a short conversation by choosing the most appropriate response to the last turn of the provided listening stimuli, whereas the Q&A type requires them to get the main idea or make inferences based on the conversation. Therefore, the two item types of MCQ items might differently affect the test-takers' process and performance on the listening test.

Also, the earlier studies mentioned above had limitations in that they only employed listening comprehension items in written mode, which means that all the questions and options of the items used for their study demanded test-takers' reading ability, because they were written on the test paper. For example, Cheng (2004) clearly stated as one of the

limitations of her study that all of her questions, regardless of item types, required students' reading skills as well as their listening skills. This gave rise to necessity of future studies on the varying effects of aural and written modes of question presentation when comparing different item types. The potential interaction between the presentation mode and the item type is discussed in the following section.

2.3. Potential Interaction Between the Presentation Mode and the Item Type

Among the different studies on the effect of presentation mode discussed in Section 2.1, no studies on the effect of aural and written mode of question/option presentation made any distinction between different item types (Chang, 2005; Chang & Read, 2013; Wu, 1998; Yanagawa & Green, 2008). Chang and Read (2013), for example, did include different item types for their listening test, but did not mention which item types they used and did not report their effect on the results.

Considering item types when examining the effect of presentation mode in multiple-choice items is worth investigating, because combined characteristics of test method affect test-takers' cognitive processing and test scores in a dissimilar way (Bachman & Palmer, 1996). In other words, since different characteristics of test methods could interact with other characteristics, the impact of any single item characteristic on test-takers' performance needs a detailed analysis in terms of its interaction with other factors (Brindley & Slatyer, 2002).

In fact, Cheng (2004) specifically called for a further study that considers both mode and item type, since most research on item types, including hers, used items presented only in the written mode. Regarding the effect of presentation mode that was discussed in the previous section, therefore, examining how the mode difference works in relation to different item types is needed to acquire a more comprehensive picture of these two factors' effect.

Thus, the present study investigated the effect of modes of question/option presentation together with that of different item types of multiple-choice questions on test-takers' listening comprehension test performance and their perception. Since previous research found mixed results regarding the effect of aural and written modes of question presentation, particularly for lower-level test-takers, different proficiency levels are taken into account. Test-takers' perceptions are also investigated using a survey and an in-depth verbal report. The following two research questions are addressed in this study:

1. To what extent do modes of question/option presentation and item type affect three proficiency groups' L2 listening performance?

2. What is the three proficiency groups' perception of aural and written modes of presentation on two item types?

3. METHODOLOGY

3.1. Participants

One hundred and fifteen Korean college students who have been learning English as a foreign language for 10 years on average participated in the study. The participants were recruited through online bulletin boards of two colleges, one of which is located in Seoul and the other in Cheongju, Korea. They came from a variety of majors/departments, including architecture, chemistry, computer information, English literature, medicine, and public administration.

The 115 students were divided into three groups of 38, 43, and 34 students by their proficiency levels: the advanced level[1] group with 700 points or above on TEPS[2], the mid to high-intermediate level group with 500 to 700 points on TEPS, and the low-intermediate level group with 500 points or lower on TEPS, respectively.

3.2. Instruments

3.2.1. The listening comprehension test

The two item types selected for the present study are the dialogue-completion type and the question-and-answer (Q&A) type. These two item types require different listening skills in that the test-takers have to complete a short conversation by choosing the most appropriate and spontaneous response for the dialogue-completion items, while they have to get the main idea or make inferences based on the conversation for the Q&A items. Since the stimulus material for the dialogue-completion tasks is short conversations between two people, only dialogues, not monologues, were used for the Q&A items in this study to keep the stimulus of the two item types the same. The overview of the listening comprehension test used for the study and the samples of each item type are summarized in Table 1 and 2, respectively. All items were reviewed by a researcher at the TEPS Council to balance the topics and difficulty levels between the sections.

---

[1] The level of each group is given according to the TEPS Council (2009).
[2] Test of English Proficiency developed by Seoul National University

TABLE 1

Overview of the Listening Comprehension Test Used for the Study

Section	Number of Questions	Item Type	Presentation Mode
1	4	Dialogue-completion	Aural
2	4	Dialogue-completion	Written
3	4	Q & A	Aural
4	4	Q & A	Written

3.2.2. Post-test survey

A survey was developed to explore the participants' perceptions toward each section of the listening test, looking into their impression of each format's difficulty and validity, their preference, and some possible factors that might have affected them. See Appendix for the complete form of the survey.

TABLE 2

Sample Items Used for Each Format

Item Type	Listening Stimuli (Presented aurally)	Question/option Presentation Mode (How it is presented on test paper)	
		Aural Mode	Written Mode
Dialogue-completion	1. M: How do I get to the library? W: Follow this road to the big brick building. M: And that's the library? W: _____	1. (a) (b) (c) (d)	1. Choose the most appropriate response to complete the conversation. (a) It's my book. (b)No, but I'll try. (c) Yes, that's it. (d) Study them first.
Q&A	2. M: Are you transferring to a different university? W: Yeah, to one that's closer to home. M: Oh? Which one? W: Scarborough University. It's located near my family and has my major. M: That sounds good. W: Yes, I'm happy about it.	2. (a) (b) (c) (d)	2. What is the main topic of the conversation? (a) A plan to change schools (b) The high cost of education. (c) A comparison of two universities (d) Programs at Scarborough University

3.3. Procedure

For the listening comprehension test papers, two forms, Form A and Form B, were developed to counterbalance the order of the sections. For dialogue-completion items, for instance, around one half of the test-takers from each proficiency group took Form A, receiving question 1~4 in the aural mode and 5~8 in the written mode. The other half who took Form B received question 1~4 in the written mode and 5~8 in the aural mode. The

two forms of the test and the form assignment for the three proficiency groups are shown in Table 3 and Table 4, respectively.

TABLE 3

Two Forms of the Test

	Form A	Form B
Dialogue-completion	Aural (1~4) – Written (5~8)	Written (1~4) – Aural (5~8)
Q&A	Written (9~12) – Aural (13~16)	Aural (9~12) – Written (13~16)
Total Items	16	16

After taking their assigned listening comprehension test, participants completed the survey. Twelve participants (4 from each group) were selected and asked to engage in stimulated recall interviews. Listening to the recording again, they reported how they chose the answer, including decisions they made in each step and the reasons for them. Also, some follow-up interview questions were given to ask participants to elaborate on their survey responses and to clarify their reasons. These verbal reports were conducted to qualitatively investigate their cognitive processes and factors that affected their judgements in each step of test-taking processes. All the processes of interviews were recorded and transcribed by the researcher.

TABLE 4

Form Assignment

	Low-Intermediate (Group L)	Mid/High-Intermediate (Group M)	Advanced (Group H)
Form A	21	15	18
Form B	22	19	20
Total	43	34	38

3.4. Data Analysis

SPSS 22.0 for Windows was employed for the statistical analysis. For each proficiency group, a repeated measure two-way ANOVA was used for analysis to examine the effect of presentation mode and item type on test-takers' L2 listening performance and perception. The two-way ANOVA was used because there were two independent variables (presentation mode and item type), and since all participants took the listening test in all different formats (both the aural and the written modes, and both the discourse-completion and the Q&A item types), the repeated measure was employed. A one-way ANOVA was also used to compare means of the three proficiency groups' responses for some survey questions. Qualitative analyses for some post-test survey questions and the stimulated

recall interviews were also conducted for an in-depth investigation of the meaning of the results drawn from quantitative analyses. For the verbal reports data, the researcher went through the transcriptions to group similar responses and to develop coding categories. The transcriptions were then reviewed thoroughly again and were coded for the analysis.

4. RESULTS AND DISCUSSION

4.1. The Effects of Mode and Item Type on the Test-Takers' Performance

For the first research question, a repeated measure two-way analysis of variance (ANOVA) was employed to examine the effect of mode and item type on the participants' scores on the listening comprehension test. The results varied according to different proficiency groups.

The three proficiency groups' mean scores and standard deviations for four different sets of listening comprehension tests are presented in Table 5. One point was given to each item and the four formats, DC-Aural, DC-Written, Q&A-Aural, and Q&A-Written, had 4 items each, so the highest score and the lowest score one could get was 4 and 0, respectively.

TABLE 5

Descriptive Statistics for Listening Comprehension Test Performance in Two Different Question Types and Two Different Modes

	DC-Aural		DC-Written		Q&A-Aural		Q&A-Written	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Group L	1.65	1.193	2.09	1.065	1.51	.935	1.70	1.103
Group M	2.97	.870	3.06	.814	2.53	.825	2.71	1.142
Group H	3.84	.370	3.82	.393	3.47	.687	3.58	.500

Note. DC: Dialogue-completion type, Q&A: Question-and-answer type

Table 6 provides a summary of the repeated measure two-way ANOVA. Significant main effects were shown for Item Type for Group M [F (1, 33) = 8.468, p = .006] and Group H [F (1, 37) = 11.426, p = .002], and their effect sizes were relatively large (partial η^2 = .204 for Group M and partial η^2 = .236 for Group H). This shows that Group M and Group H received much better scores on discourse-completion type. However, no main effect was found for Mode for the two groups, which means that Group M and Group H performed similarly on items presented in aural and written mode. For Group L, on the other hand, a significant main effect were detected for Mode [F (1, 42) = 4.394, p = .042, partial η^2 = .095], but no significant effect for Item Type, indicating that the low-level students performed better with the written mode and were not influenced by the item type.

TABLE 6

Results of the ANOVA for the Effects of Item Type and Mode on Test-Takers' Listening Comprehension

	Source	*SS*	*df*	*MS*	*F*	*p*	partial η^2
Group L	Item Type	3.076	1	3.076	2.891	.096	.064
	Mode	4.238	1	4.238	4.394	.042	.095
	IT * Mode	.703	1	.703	.738	.395	.017
	Error	40.047	42	.953			
Group M	Item Type	5.360	1	5.360	8.468	.006	.204
	Mode	.596	1	.596	.766	.388	.023
	IT * Mode	.066	1	.066	.070	.793	.002
	Error	31.184	33	.945			
Group H	Item Type	3.480	1	3.480	11.426	.002	.236
	Mode	.059	1	.059	.285	.597	.008
	IT * Mode	.164	1	.164	.635	.431	.017
	Error	9.586	37	.259			

This indicates that the higher proficiency groups performed equally well whether the questions and answers were recorded or written. This result is consistent with earlier studies which compared the scores of aural and written mode between participants with higher and lower proficiency levels (Chang & Read, 2013; Yanagawa & Green, 2008). The two higher proficiency groups' relatively lower scores on the Q&A type items were somewhat expected due to its longer input and options that could have made it more difficult. While the input listening stimuli for the dialogue-completion items consisted of 3 turns, that for the Q&A items consisted of 6 turns. The answer options for the Q&A items were also relatively longer. The average number of words in an option for a Q&A item was 8.1, whereas that for a dialogue-completion item was 6.2.

The lowest proficiency group performed differently. Although they did feel that the Q&A items were much more difficult than the dialogue-completion items (see 4.2.1), the result shows that they obtained similar scores on both item types. However, they were significantly influenced by the presentation mode, scoring much lower when the questions and options were given aurally. This could mean that the difficulty they had in listening to all four options and processing them far outweighs that in reading the options and processing them in time. Therefore, choosing between the aural and written mode in the development process of a listening test may have much more critical influence on the test-takers with the low-intermediate proficiency level than on any other groups with higher proficiency, whether the items are dialogue-completion or Q&A.

4.2. Test-Takers' Perception of Aural and Written Modes of Presentation on Two Item Types

To answer the second research question, perceived difficulty and validity of items in each format and test-takers' preference for the two presentation modes (aural and written) on the two item types (dialogue-completion and Q&A) were analyzed.

4.2.1. Perceptions on the difficulty and validity of each format

A repeated measure two-way ANOVA was used to investigate the impact of Mode and Item Type on the perceived difficulty of the items. A five-point Likert response scale was used for each question, "how easy or difficult was items in each format?". Means and standard deviations for three proficiency groups are shown in Table 7 below.

Relatively consistent results for the perceived difficulty were obtained throughout the three different proficiency groups. The results of the ANOVA on the perceived difficulty are summarized in Table 8. No interaction effect was found between Item Type and Mode. There was a statistically significant main effect for Mode in all three groups' perceived difficulty, which was significantly higher for the aural mode than the written mode of question and option presentation [F (1, 41) = 11.978, p = .001 for Group L; F (1, 31) = 15.583, p = .000 for Group M; and F (1, 37) = 8.388, p = .006 for Group H]. In other words, all three groups felt that the aural mode was more difficult than the written mode. Chang and Read's (2006) study also had the same result, in which test-takers felt the written mode easier than the aural mode. A significant main effect for Item Type was also revealed in all three groups' perceived difficulty, suggesting that the Q&A item type was perceived to be much more difficult than the dialogue-completion type in all three groups [F (1, 41) = 15.626, p = .000, partial η^2 = .276 for Group L; F (1, 31) = 15.127, p = .000, partial η^2 = .314 for Group M; and F (1, 37) = 23.347, p = .000, partial η^2 = .387 for Group H].

TABLE 7

Descriptive Statistics for Perceived Difficulty of the Items in Two Different Question Types and Two Different Modes

	DC-Aural		DC-Written		Q&A-Aural		Q&A-Written	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Group L	2.62	1.188	2.24	.850	3.31	1.047	2.86	.899
Group M	2.88	1.094	2.41	.988	3.59	.892	3.06	.919
Group H	2.47	1.059	2.24	1.149	3.32	.842	2.89	.981

TABLE 8

Results of the ANOVA for the Effects of Item Type and Mode on Perceived Difficulty of the Listening Comprehension Items

	Source	*SS*	*df*	*MS*	*F*	*p*	partial η^2
Group L	Item Type	18.006	1	18.006	15.626	.000	.276
	Mode	7.292	1	7.292	11.978	.001	.226
	IT * Mode	.054	1	.054	.166	.685	.004
	Error	13.196	41	.322			
Group M	Item Type	15.559	1	15.559	15.127	.000	.314
	Mode	8.500	1	8.500	15.583	.000	.321
	IT * Mode	.029	1	.029	.063	.804	.002
	Error	15.471	33	.469			
Group H	Item Type	21.375	1	21.375	23.347	.000	.387
	Mode	4.112	1	4.112	8.388	.006	.185
	IT * Mode	.322	1	.322	1.505	.228	.039
	Error	7.928	37	.214			

To analyze the effect of Mode and Item Type on the participants' perception on the validity of the items in each format, a repeated measure two-way ANOVA is used. A five-point Likert response scale was used for each question, "how well do you think does each item format assess your listening competence?". Table 9 presents means and standard deviations for the three groups' responses for each format.

TABLE 9

Descriptive Statistics for Perceived Validity of the Items in Two Different Question Types and Two Different Modes

	DC-Aural		DC-Written		Q&A-Aural		Q&A-Written	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Group L	3.47	.882	3.33	.865	3.65	.870	3.86	.833
Group M	3.62	.739	3.32	.768	3.91	.621	3.79	.770
Group H	3.39	.974	3.26	.891	3.84	.973	3.76	.820

According to the result of ANOVA, as summarized in Table 10, a marginally significant interaction effect between Mode and Item Type was detected in Group L [F (1, 41) = 3.941, p = .054], indicating that the participants with lower proficiency felt that the aural mode was more appropriate for the dialogue-completion items, while the written mode was more valid for the Q&A items. For the two other more proficient groups, Group M and Group H, no significant interaction effect was detected, and the two groups thought that the Q&A items can assess their listening ability much better than the dialogue-completion items [F (1, 31) = 5.555, p = .025 for Group M; and F (1, 37) = 9.715, p = .004 for Group H]. In terms of the presentation mode, the means of perceived validity were higher for the aural mode in both dialogue-completion and Q&A, but there was no significant main effect.

TABLE 10

Results of the ANOVA for the Effects of Item Type and Mode on Perceived Validity

	Source	*SS*	*df*	*MS*	*F*	*p*	partial η^2
Group L	Item Type	5.587	1	5.587	9.515	.004	.185
	Mode	.052	1	.052	.109	.743	.003
	IT * Mode	1.308	1	1.308	3.941	.054	.086
	Error	13.942	42	.332			
Group M	Item Type	4.971	1	4.971	5.555	.025	.144
	Mode	1.441	1	1.441	2.788	.104	.078
	IT * Mode	.265	1	.265	.946	.338	.028
	Error	9.235	33	.280			
Group H	Item Type	8.526	1	8.526	9.715	.004	.208
	Mode	.421	1	.421	1.238	.273	.032
	IT * Mode	.026	1	.026	.054	.817	.001
	Error	17.974	37	.486			

The result that the Group L was the only group who felt that the written mode can assess their listening ability better than the aural mode in a listening test for at least one item type might have a close relation to the Group L's performance discussed in the previous section. Group L was the only group whose listening comprehension scores were significantly lower in the aural mode than in the written mode. This trend existed in the earlier studies such as Chang and Read (2013) and Yanagawa and Green (2008), which revealed that the lower level test-takers were more disadvantaged without written items. A possible explanation for these results is that among the L2 listeners who were reported to have limited short-term memory when listening in the L2 (Cook, 2013; Ohata, 2006), participants with lower proficiency were even more affected by the memory problem associated with the aural mode. This possibility was also suggested in Chang and Read (2013). Also, without the written options, they could not use the word-matching strategy that they frequently resorted to when they could not fully understand the listening stimuli. A more detailed and comprehensive examination on the factors that affected the participants' performance and perception is discussed in Section 4.3.

4.2.2. Mode preference for each item type

Participants were asked to choose one presentation mode they prefer for each item type. The reasons for their preference were elicited by open-ended questions following each preference question, and by seven five-point Likert response scale questions. The number and the percentage of the three groups' preference are shown in the Table 11.

TABLE 11

Mode Preference for Dialogue-Completion and Q&A Item Types

Item Type	Mode	Group L	Group M	Group H	Total
DC	Aural	12 (27.9%)	10 (29.4%)	6 (15.8%)	28 (24.3%)
	Written	31 (72.1%)	24 (70.6%)	32 (84.2%)	87 (75.7%)
	Total	43(100.0%)	34(100.0%)	38(100.0%)	115(100.0%)
Q&A	Aural	9 (22.0%)	6 (18.2%)	5 (13.2%)	20 (17.9%)
	Written	32 (78.0%)	27 (81.8%)	33 (86.8%)	92 (82.1%)
	Total	41(100.0%)	33(100.0%)	38(100.0%)	112(100.0%)

The majority of all three groups answered that they preferred the written mode to the aural mode for both of the question types. Also, more participants preferred the written mode of question/option presentation for the Q&A type than for the dialogue-completion type. The positive attitude toward the written mode was comparable to those reported by Buck (1991) and Chang and Read (2006). It was interesting to note that even though Group H did equally well on both modes of tests and was not affected much by the mode, they strongly preferred the written mode for both item types, and their preference was stronger than any other group.

Responses they gave for the open-ended questions each of which immediately followed the two preference questions were categorized according to the participants' different reasons for the preference. The reasons for preferring the written mode were similar in all three groups: no memory burden, possibility of predicting the stimuli by reading the options, no need to worry about not hearing and missing options. The difference between Group H and the other two groups lies in the reasons for choosing the aural mode. That reading the options in time was difficult was the major reason for the Group L and Group M's participants who preferred the aural mode. The participants from Group H who preferred the aural mode, however, did not have this reading problem. Not even a single person from Group H said that they preferred the aural mode because they could not read the questions and options fast enough. Their reasons for choosing the aural mode were that they thought the aural mode was more valid for assessing listening and that it was somewhat uncomfortable and unnatural to switch modes.

TABLE 12

Descriptive Statistics for Seven Questions on Reasons for Mode Difficulty

		Q5-1	Q5-2	Q5-3	Q5-4	Q5-5	Q5-6	Q5-7
Group L	*M*	2.52	3.60	3.33	2.81	2.48	2.57	2.86
(N = 42)	*SD*	1.042	.939	1.162	1.174	.994	1.272	1.299
Group M	*M*	2.58	4.12	3.06	2.36	2.73	2.42	2.85
(N = 33)	*SD*	1.324	.696	1.144	1.084	1.098	1.200	1.176
Group H	*M*	2.79	3.97	3.17	2.05	2.79	2.16	2.89
(N = 38)	*SD*	1.212	.822	1.152	1.089	1.069	1.079	1.226

TABLE 13

Results of One-Way ANOVA for the Seven Questions

		SS	*df*	*MS*	*F*	*p*
Q5-1	Between Groups	1.537	2	.768	.546	.581
	Within Groups	154.853	110	1.408		
	Total	156.389	112			
Q5-2	Between Groups	5.658	2	2.829	4.062*	.020
	Within Groups	76.608	110	.696		
	Total	82.265	112			
Q5-3	Between Groups	1.420	2	.710	.534	.588
	Within Groups	146.350	110	1.330		
	Total	147.770	112			
Q5-4	Between Groups	11.603	2	5.802	4.624	.012
	Within Groups	138.007	110	1.255		
	Total	149.611	112			
Q5-5	Between Groups	2.202	2	1.101	.998	.372
	Within Groups	121.337	110	1.103		
	Total	123.540	112			
Q5-6	Between Groups	3.468	2	1.734	1.228	.297
	Within Groups	155.399	110	1.413		
	Total	158.867	112			
Q5-7	Between Groups	.045	2	.022	.015	.986
	Within Groups	168.964	110	1.536		
	Total	169.009	112			

This tendency was also supported by the results of some additional survey questions on why each mode was easy or difficult, which were measured with seven questions (Q5-1 to Q5-7) using a five-point Likert response scale (see Appendix), following Chang and Read (2013). Means and standard deviations for the seven questions are shown in Table 12 and the results of the one-way ANOVA of the three groups for each question are summarized in Table 13. According to the table, the difference between the three groups was confirmed to be significant in Q5-2 ("I felt written MCQ was easy because I did not have to worry that I might not aurally comprehend the aural questions and answer options") [F (2, 110) = 4.062, p = .020] and Q5-4 ("I felt written MCQ was difficult because I could not finish reading the questions and options") [F (2, 110) = 4.624, p = .012]. The post hoc comparison (Tukey HSD) showed significant differences between Group L and Group M for Q5-2 (p = .021) and between Group L and Group H for Q5-4 (p = .009).

In other words, Group L was much more anxious about not being able to comprehend questions and options that are delivered aurally than Group M. Also, Group L had much more difficulty in reading the options in allocated time with the written mode of question/option presentation, compared to Group H. It was noteworthy that Group H was more confident reading the questions and options than listening to them. When they were asked if they felt the written mode was difficult because they could not finish reading the questions and options, their response on the Likert scale was significantly lower than

Group L. On the contrary, there was no significant difference between the two groups when they were asked if they did not have to worry that they might not aurally comprehend the aural questions and options.

4.3. Factors Found in the Different Test Formats and Their Effects on Test-Takers' Performance and Perception

The stimulated recall interviews showed the test-takers' process of arriving at the answers in the listening comprehension test. The analysis of the data suggested some possible factors related to the two presentation modes (aural and written) on each item type (dialogue-completion and Q&A) that were found to affect the process and the performance of the participants in the listening comprehension test. Some of them were not directly relevant to the listening comprehension, which might have a negative influence on the validity of the test. The interview processes were done in Korean, and all Korean utterances in Examples were translated into English.

4.3.1. Issues found in the aural mode

For the aural mode, the need of a high level of concentration and good memory seemed to be the most problematic for test-takers. Test-takers needed to sustain a high level of concentration all the time not to miss any option. Two participants from Group L and three participants from Group M said they missed options while answering some questions in the aural mode. This phenomenon appeared in both dialogue-completion and Q&A types.

Example 1: Aural, Dialogue-completion, Participant L1
I knew (a) wasn't the answer and missed (b), so I thought (c) was the answer. (wrong)

Example 2: Aural, Q&A, Participant L4
I heard (a) but didn't hear (b), (c), and (d) well. So I just chose (a). (wrong)

This transient nature of the aural mode also required good short-term memory, which was reported to be more limited for L2 learners (Cook, 2013; Ohata, 2006) and to be even more constrained for L2 learners with lower proficiency (Chang & Read, 2013). If participants wanted to compare options, they had to do it in their memory, recalling the options and the listening stimuli and comparing them at the same time to choose the best answer. Most of the test-takers found it difficult to do this. Some participants mentioned that this problem was more serious in Q&A items, as can be seen in the Example 4,

because Q&A items usually required more logical thinking than dialogue-completion questions.

Example 3: Aural, Dialogue-completion, Participant L4
By the time (c) and (d) came out, I forgot what (a) and (b) were. I just had to choose the answer randomly.

Example 4: Aural, Q&A, Participant M3
I didn't have any memory problem in dialogue-completion questions, but for Q&A questions, I often missed (c) and (d). I like to read options before listening but for these questions I couldn't.

Some test-takers with high levels of proficiency managed to overcome this problem by taking notes of all the questions and options, visualizing them like those in the written mode, as shown in Example 5.

Example 5: Aural, Q&A, Participant H3
I didn't have any problem with remembering the options, because I took notes of all the options as much as possible. Then I could easily compare the options.

However, the lower level participants did not employ the same strategy. Listening and writing at the same time might have been another high-level ability that the lower level participants did not possess.

This also led to another problem found in the aural mode, the random guessing of the answer. Wu (1998) noted uninformed guessing as one of the construct-irrelevant factors found in the multiple-choice questions in general, and those presented aurally seemed to exacerbate this problem. For the written mode, the participants rarely made complete random guesses. They at least provided some pieces of evidence for their choices, because they could cross out the options that were most unlikely to be the answer and then had a guess at the answer among remaining options. For the aural mode, however, when they were not sure about the answer immediately, they often failed to recall all the options and ended up making random guesses. In fact, several random guesses were noted in the aural mode. Two participants from Group L and three participants from Group M showed random guessing for the answer in the aural mode, without having any clue from the listening. Example 6 demonstrates this.

Example 6: Aural, Dialogue-completion, Participant M4
I understood the conversation, but I couldn't find the answer among (a), (b), (c),

and (d). So I just randomly picked one.

In short, participants reported that they often missed options and found it hard to remember them when the options were presented aurally. Radom guessing was more frequently reported for the items in the aural mode. The participants from Group L and Group M, who have lower proficiency levels, had more of these problems compared to those with higher proficiency level, Group H.

4.3.2. Issues found in the written mode

For the written mode, the most salient one of the construct irrelevant factors was that it required certain level of reading ability, and the better a participant was in the reading comprehension, the more beneficial it was. This unfavorably affected the participants with lower proficiency, because their reading skills and speed were often not as good as those that are required for understanding the written options in time. The participants from Group L frequently reported that they had difficulty reading the questions and options and comprehending them quickly, while Group H did not have any difficulty with it. Example 7 shows difficulty arising from having to read the options in the allotted time in the written mode. Example 8 illustrates a different tendency showed in a case of a participant from Group H who had relatively high reading ability.

Example 7: Written, Dialogue-completion, Participant M4
I had to choose the answer without fully understanding them because there was not enough time.

Example 8: Written, Q&A, Participant H3
I could read all the answer options before listening to the passage. There was enough time. Actually, the pauses between questions were too long.

Mode switching was another difficult or annoying factor for some test-takers on the written mode, as can be seen in the accounts below. The test-takers had to listen to the passage and read the questions and options alternately or simultaneously. This was also more disadvantageous for the lower level participants. They were sometimes confused by reading and listening at the same time, or missed the first part of the listening stimuli, because they were reading the options. The confusion that the test-takers experienced was also reported by some previous studies (Ur, 1984; Weir, 1993), which found that test-takers' attention to listening was distracted by the written options.

Example 9: Written, Q&A, Participant L3
The conversation was long, and I also had to read the options at the same time. This made me more confused, and I got lost.

Example 10: Written, Q&A, Participant M3
The listening started while I was still reading the options, so I missed the first part of the listening.

Example 11: Written, Dialogue-completion, Participant H2
I suddenly had to read the lines instead of listening and it wasn't that comfortable.

On the other hand, some factors were helpful to the test-takers in terms of finding the answer, but not directly relevant to listening comprehension. The written mode allowed the test-takers to read the options before listening to the passage. In this way, they were able to predict the general ideas of the listening stimuli in advance, a behavior which was also discovered in earlier studies (Buck, 1991; Yanagawa & Green, 2008). This was especially helpful for those with high proficiency, as shown in Example 12, because their reading speed was high enough for reading all the options before listening to the passage.

Example 12: Written, Q&A, Participant H4
I read through the options before the listening came out, and then I could know that the listening would be about "leather jacket" and particularly "fake one." And then I heard "moral" and "cruelty" in the listening, so I immediately knew that (d) was the answer.

Employing this strategy on Q&A type items was reported to benefit participants more than doing so on dialogue-completion type items. Participant M1 explained that since the options in Q&A items were statements about the listening stimuli, she could easily predict the topic.

Example 13: Written, Q&A, Participant M1
I read all the answer options before listening to the passage. I didn't do this for the dialogue-completion items because it was not really helpful to predict the listening stimuli, but it helped me a lot for the Q&A items.

However, it did not help participants from Group L much, because it was not possible for them to read options before listening to the passage due to their low

reading speed, as shown in Example 14.

Example 14: Written, Q&A, Participant L3
I wanted to read the options before the listening, but I couldn't. I didn't have time to do so.

This result was in line with Wu's (1998) finding in his study on 10 Chinese ESL students' listening test-taking processes using retrospective verbal report. He concluded that viewing the questions and options helped advanced students by allowing them to predict the listening, but it was not beneficial to students with lower proficiency. The unsuccessful use of the written questions and options was also discovered by the findings of Buck (1991) and Chang and Read (2006).

Instead, the lower-level participants frequently used the word-matching strategy. Without understanding the listening stimuli, they guessed the answer by matching some words they heard from the listening stimuli to those in the written options, similar to what Participant L4 explained in Example 15. This trend occurred more frequently in the Q&A section. The strategy was not always successful, but was often used when they could not understand the listening.

Example 15: Written, Q&A, Participant L4
I didn't understand the conversation, but I heard the phrase "tried out" so I chose (d) for the answer. (correct)

This lexical attractiveness has been reported by several previous studies (Freedle & Fellbaum, 1987; Freedle & Kostin, 1996, 1999). Freedle and Fellbaum (1987) found overlapping words between single-sentence listening passage and the answer options played a significant role in determining the item difficulty, and Freedle and Kostin (1999) yielded a similar result using longer listening stimuli, TOEFL mini-talks. Yanagawa and Green (2008) also noted that previewing answer options encouraged students to use lexical matching strategy.

The participants' preference (see 4.2) showed that all three groups showed a stronger preference for the written mode in Q&A items than in the dialogue-completion items. This can also be partly explained by the participants' use of the two strategies: prediction and word-matching. The result that the prediction strategy was more frequently used for Q&A items than dialogue-completion items means that the strategy was more useful for Q&A items particularly for the high level test-takers. The word-matching strategy was also more suitable for Q&A items in the written mode, which made the written mode more attractive for Q&A items for the lower level test-takers as well.

In sum, written questions and options interfered with test-takers choosing the right answer in some situations, while they helped them in other situations. Mode switching in the written mode was confusing to a few participants regardless of their proficiency levels, and reading options was not easy for participants with lower proficiency. On the other hand, test-takers reported having taken advantage of written options by predicting the topic and using word-matching strategy particularly for Q&A items. The predicting strategy was exclusively used by the participants with relatively higher proficiency, while the word-matching strategy was often used by those with lower proficiency.

5. CONCLUSION

This study investigated the effect of question/option presentation mode and item type on Korean EFL learners' listening comprehension performance and their perception. The test-takers with mid/high-intermediate to advanced proficiency level (Group M and Group H) and those with low-intermediate level (Group L) performed and felt differently in the listening test that was given in the two modes and the two item types. Group L was significantly influenced by the mode, receiving much lower scores in the items in aural mode. They felt the aural mode was much more difficult than the written one, and preferred to have items with written questions and options. Unlike Group M and Group H, Group L thought the aural mode was better for the dialogue-completion items whereas the written mode was more appropriate for the Q&A items.

The reasons Group L found the aural mode much more difficult were that it required them with a high level of concentration and sufficient short-term memory capacity, in addition to their listening skills. Using the aural mode in this sense might undermine the validity of the listening test. To some extent, however, Group L's higher score in the written mode was partly due to their tendency to resort to the word-matching strategy. This can negatively affect the validity of the test, because the test-takers just matched some words from the listening to the written options, not understanding them. In addition, reading options in time was an obstacle for taking the listening test in the written mode for many Group L participants. In other words, they could not get the item right if they could not finish reading the options in time, even if they understood the listening stimuli.

Unlike Group L, Group M and Group H were not significantly affected by the presentation mode. Still, they felt the aural mode was much more difficult than the written mode, and Group H even expressed a stronger preference for the written mode than the other two groups did. Even though they could perform equally well on both modes, they were more confident about reading the questions and options than listening them. Additionally, they did not have any difficulty in reading in time. With their high reading

proficiency, they could read the options in advance and predict the listening stimuli before actually listening to it. Group L did not get this chance to predict the topic because their low reading proficiency prevent them from reading the options in advance. Thus, since the written mode inevitably requires a certain level of reading ability, the test-takers' reading ability, in addition to the listening ability, can critically affect the test-takers' scores.

Since construct validity is "central to the appropriate interpretation of test scores" (Bachman, 1990, p. 255), identifying constructs that a test is measuring is one of the key issues in the test development process. This study has drawn attention to several issues related to developing and choosing listening comprehension multiple-choice tests regarding different formats.

Both the aural and the written modes are found to have strengths and weaknesses with respect to construct validity. The influence of reading ability on test-takers' listening comprehension performance could be avoided with the aural question/option presentation mode by providing all questions and options aurally, but this imposed an additional memory burden on the test-takers. On the other hand, test-takers felt less anxious with written questions and options, because they did not have to remember them. However, the written mode required a certain level of reading proficiency and entailed other reading related construct-irrelevant factors, such as predicting without listening and using word-matching strategies. The impact of each mode's characteristics was intensified in Q&A items, since they had longer options compared to the dialogue-completion ones.

The proficiency level of the target population also needs to be taken into consideration when making decisions on choosing or developing a listening comprehension test. Participants with lower proficiency were more critically affected by mode difference. Thus, if the target population of a listening test includes participants with low-intermediate proficiency level or even lower levels, the test developers and teachers should keep in mind that the memory capacity and the reading ability, which are not the constructs that are intended to be measured, could significantly affect the test-takers' listening performance.

Based on participants' performance and perception on the different formats of listening tests and reports from the recall interviews, some suggestions can be made for test developers and teachers who develop or choose listening comprehension multiple-choice tests for EFL learners. It is most recommendable to give the questions and options aurally, because it does not require reading ability which could hinder the test-takers with low reading proficiency from fully demonstrating their listening ability and because the written mode allows word-matching strategy. The memory burden on the test-takers in the aural mode could be relieved by reducing the number of options and by making them short and clear. This is particularly advisable for the dialogue-completion items, since the test-takers have to choose a response that is a part of the whole aurally-given conversation. However, the written mode can be more appropriate in some cases, such as when the options are too

long for the aural mode and too difficult for the test-takers to process with their memory capacity only by listening. The written mode can also have a positive effect on the test-takers by reducing their test anxiety. Yet, if the questions and options are to be delivered in written form, they should be written in easy language not to require a high level of reading ability.

The current study has limitations that could be improved in the future study. First, the number of participants who took part in the stimulated recall interviews might not have been representative enough to generalize the findings. Still, the interviews did reveal some important issues regarding the presentation mode and the item type for the listening multiple-choice questions. Another limitation of the study is that participants' proficiency levels varied within each group. Participants from Group M were particularly diverse in their proficiency levels, therefore the performance of the group's higher end was more like the advanced level while that of the lower end was similar to the low-intermediate level.

The findings and suggestions of this study can be developed into future research regarding the following two research topics. Firstly, to relive the memory burden of the aural mode not by giving the written options but by reducing their number, the interaction effect between the number of options and the presentation mode needs to be investigated. Secondly, the appropriate length or complexity of the language for the options in the written mode can be examined in relation to the test-takers' proficiency levels. These further research would help provide a clearer picture about how to employ the aural and the written mode for the listening tests with multiple-choice questions.

The present study provides some implications on the importance of considering the testing method effects when developing the listening tests. Testing methods can change what the test is measuring, interacting with other factors. Therefore, the characteristics of the target test-takers and the item types, as well as the constructs of the test, are the first things to consider when deciding the presentation mode of questions and options in a listening test.

REFERENCES

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford, UK: Oxford University Press.

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

Berne, J. E. (1992, August). *The role of text type, assessment task, and target language experience in L2 listening comprehension assessment*. Paper presented at the the Annual Meetings of the American Association for Applied Linguistics and the

American Association of Teachers of Spanish and Portuguese, Cancun, Mexico.

Berne, J. E. (1995). How does varying pre-listening activities affect second language listening comprehension? *Hispania, 78*(2), 316-329.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing, 19*(4), 369-394.

Buck, G. (1990). *The testing of second language listening comprehension.* Lancaster, UK: University of Lancaster.

Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing, 8*(1), 67-91.

Buck, G. (2001). *Assessing listening*. Cambridge, UK: Cambridge University Press.

Chang, C.-S. (2005). The perceived effectiveness of question preview in EFL listening comprehension tests. *New Zealand Studies in Applied Linguistics, 11*(2), 75-96.

Chang, C.-S. (2008). Listening strategies of L2 learners with varied test tasks. *TESL Canada Journal, 26*(1), 1-26.

Chang, C.-S., & Read, J. (2006). The effects of listening support on the listening performance of EFL learners. *TESOL Quarterly, 40*(2), 375-397.

Chang, C.-S., & Read, J. (2013). Investigating the effects of multiple-choice listening test items in the oral versus written mode on L2 listeners' performance and perceptions. *System, 41*(3), 575-586.

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), 544-553.

Cook, V. (2013). *Second language learning and language teaching*. London, UK: Routledge.

Freedle, R., & Fellbaum, C. (1987). An exploratory study of the relative difficulty of TOEFL's listening comprehension items. In F. Feedle & R. Duran (Eds.), *Cognitive and linguistic analyses of test performance* (pp. 162-192). Norwood, NJ: Albex.

Freedle, R., & Kostin, I. (1996). *The prediction of TOEFL listening comprehension item difficulty for minitalk passages: Implications for construct validity*. ETS Research Report Series No. RR-96-29. Princeton, NJ: Educational Testing Service.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension?: The case for the construct validity of TOEFL's minitalks. *Language Testing, 16*(1), 2-32.

Hansen, C., & Jensen, C. (1994). Evaluating lecture comprehension. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 241-268). Cambridge, UK: Cambridge University Press.

Hemmati, F., & Ghaderi, E. (2014). The effect of four formats of multiple-choice questions

on the listening comprehension of EFL learners. *Procedia-Social and Behavioral Sciences, 98*, 637-644.

Iimura, H. (2010). Factors affecting listening performance on multiple-choice tests: The effects of stem/option preview and text characteristics. *Language Education & Technology, 47,* 17-36.

Messick, S. (1996). *Validity and washback in language testing.* ETS Research Report Series No. RR-96-17. Princeton, NJ: Educational Testing Service.

Nissan, S., DeVincenzi, F., & Tang, K. L. (1995). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension*. ETS Research Report Series No. RR-95-37. Princeton, NJ: Educational Testing Service.

Ohata, K. (2006). Auditory short-term memory in L2 listening comprehension processes. *Journal of Language and Learning, 5*(1), 21-28.

Shohamy, E. (1984). Does the testing method make a difference?: The case of reading comprehension. *Language Testing, 1*(2), 147-170.

The TEPS Council. (2009). *TEPS*. Retrieved from http://www.teps.or.kr.

Underwood, M. (1989). *Teaching listening.* Boston, MA: Addison-Wesley Longman Ltd.

Ur, P. (1984). *Teaching listening comprehension.* Cambridge, UK: Cambridge University Press.

Weir, C. J. (1993). *Understanding and developing language tests.* Upper Saddle River, NJ: Prentice Hall.

Weir, C. J. (2005). *Language testing and validation*. London, UK: Macmillan.

Wolf, D. F. (1993). A comparison of assessment tasks used to measure FL reading comprehension. *The Modern Language Journal, 77*(4), 473-489.

Wu, Y. A. (1998). What do tests of listening comprehension test?: A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing, 15*(1), 21-44.

Yanagawa, K., & Green, A. (2008). To show or not to show: The effects of item stems and answer options on performance on a multiple-choice listening comprehension test. *System, 36*(1), 107-122.

APPENDIX

Survey (translated into English)

1. How easy or difficult was items in each format?	
Part 1-1 (Dialogue Completion – Aural mode)	Very easy 1 2 3 4 5 Very difficult
Part 1-2 (Dialogue Completion – Written mode)	Very easy 1 2 3 4 5 Very difficult
Part 2-1 (Q&A – Aural mode)	Very easy 1 2 3 4 5 Very difficult
Part 2-2 (Q&A – Written mode)	Very easy 1 2 3 4 5 Very difficult

2. How well do you think does each item format assess your listening competence?	
Part 1-1 (Dialogue Completion – Aural mode)	Very Poorly 1 2 3 4 5 Very well
Part 1-2 (Dialogue Completion – Written mode)	Very Poorly 1 2 3 4 5 Very well
Part 2-1 (Q&A – Aural mode)	Very Poorly 1 2 3 4 5 Very well
Part 2-2 (Q&A – Written mode)	Very Poorly 1 2 3 4 5 Very well
3-1. Which mode do you prefer for the dialogue completion items?	Aural mode / Written mode
3-2. Give reason(s) for your answer above.	
4-1. Which mode do you prefer for the Q&A items?	Aural mode / Written mode
4-2. Give reason(s) for your answer above.	
To what extent do you agree with the following statements?	
5-1 I felt aural MCQ was easy because I did not have to worry about not understanding the questions and answer options	Strongly disagree 1 2 3 4 5 Strongly agree
5-2 I felt written MCQ was easy because I did not have to worry that I might not aurally comprehend the aural questions and answer options	Strongly disagree 1 2 3 4 5 Strongly agree
5-3 I felt aural MCQ was difficult because I could not read the questions and options before hearing the input	Strongly disagree 1 2 3 4 5 Strongly agree
5-4 I felt written MCQ was difficult because I could not finish reading the questions and options	Strongly disagree 1 2 3 4 5 Strongly agree
5-5 I had no difficulty remembering the questions and options while doing the aural MCQ questions	Strongly disagree 1 2 3 4 5 Strongly agree
5-6 Doing written MCQ was difficult because I had to read and listen at the same time	Strongly disagree 1 2 3 4 5 Strongly agree
5-7 When doing aural MCQ items, I did not have to wait until the speaker finished all the options. I chose the right one once I heard it	Strongly disagree 1 2 3 4 5 Strongly agree

Applicable levels: Tertiary

Soohye Yeom
Department of English Education
College of Education, Seoul National University
1, Gwanak-ro, Gwanak-gu
Seoul 151-748, Korea
Phone: 02-880-7670
Email: soohye90@snu.ac.kr

Received on September 1, 2016
Reviewed on October 15, 2016
Revised version received on November 15, 2016