The Effects of Working Memory Span on Listening Tests without Preview Questions*

Bora Kim
(Seoul National University)

Kim, Bora. (2015). The Effects of Working Memory Span on Listening Tests without Preview Questions. *Language Research*, 51.2, 403-420.

This study aims to investigate the effects of working memory span on listening tests that do not include preview questions. In the first test session, question stems and answer options were only provided through oral input. However, in the second session, the same 31 test takers were allowed to read the question stems and answer options in a multiple-choice listening test. The gap score between the two sessions was then analyzed in relation to each student's working memory span, which was, in turn, measured through a conceptual working memory span task. The results show a clear correlation between working memory span and gap score. This study suggests that removing written cues can lead to significantly lower achievement scores for test takers with a low working memory span.

Keywords: working memory, multiple-choice listening test, large-scale listening test, TEPS listening, test validity

1. Introduction

The tasks included in listening comprehension tests can be either written or oral. Large-scale tests have multiple uses; in the listening comprehension parts of the TOEFL Internet-based Test (TOEFL iBT) developed by Educational Testing Service (ETS), for example, questions are displayed on the screen only after test takers finish listening to the entire passage. The listening section in the Test of English for International Communication (TOEIC) developed by ETS has four subsections that use both written and oral inputs: in the first two subsections, there are no preview questions, while in the latter two parts, the question stems

* This research was supported by the 2014 Seoul National University research grant.

and answer options are given in the answer sheet. The listening comprehension section in Test of English Proficiency developed by Seoul National University (TEPS) consists of four parts that do not contain any written form of input; all the question and answer options are delivered through oral input alone. One possible question can be raised here: if test takers are only able to view the questions after listening to the entire passage, that is, if they are unable to see the questions in advance, would their working memory affect their performance? This study aims to investigate the relationship between individual working memory span and task performance using two different tests: one with question previews and one without.

2. Literature Review

In this section, two concepts, listening preview and working memory span, will be reviewed.

2.1. Listening preview

Being able to view questions in advance can be an important listening support tool for test takers. Berne (1995) was one of the first studies to analyze the effects of previewing questions in listening activities. She divided the participants in her study (Spanish language learners) into three groups, each of which was assigned a different pre-activity. The first group studied the questions in advance, the second group studied the vocabulary lists and the control group had to write numbers from one to fifty in Spanish. The results showed that the first group scored significantly higher than the other two groups, whereas there was no difference between the second group and the control group. The results showed that viewing the questions in advance did help learner engagement in the listening activity, whereby they were able to form a meaningful context.

Subsequent studies examined different formats of preview questions. With 24,000 volunteers, Yanagawa and Green (2008) attempted to compare different multiple-choice formats using the questions in Part 3 of the TOEIC－lis-

tening comprehension. In their study, participants were divided into three groups: full question group, question stem only group, and answer options only group. The results revealed that the answer options only group scored significantly lower than the other two groups. What was noteworthy is that there was no significant difference between the full question group and the question stem only group. It seemed, then, that participants' access to the questions, without the answer options, was an ample resource for the test takers to activate focused listening. In other words, their access to the questions allowed them to gain meaningful context for the top-down approach, which facilitates comprehension. When test takers were provided with the answer options only, they used a less effective strategy: they matched the lexical items in the answer options with the relevant text. In addition to question types, Yanagawa and Green also tried to examine the underlying factors affecting test item difficulty. According to them, the test takers' reliance on lexical strategies affected their choice of answers. Chang and Read (2013) compared written and oral comprehension tests in which the test takers could preview the questions. In their study, which involved 160 Taiwanese undergraduate students, they offered preview questions both in oral and written forms. A two-way ANOVA revealed that there was no significant difference between the oral and written input modes. In terms of the variable of learner proficiency, though, the low-proficiency group scored higher in the written mode, whereas the high-proficiency group scored higher in the oral mode. Both proficiency groups felt that the written mode was easier than the oral mode. Further, test takers adopted more strategies when they were assigned the written questions. Chang and Read later commented on their study, saying that oral input can be more advantageous than written input because it excludes the additional cognitive process of reading, and it only concentrates on a single mode, that is, listening.

For listening comprehension tasks, learner proficiency appears to have an effect on the outcome when students were allowed to preview questions (Chang 2005; Wu 1998). With regard to the immediate recall process, Wu (1998) investigated the cognitive process of students when responding to multiple-choice questions after listening to an authentic material. Wu revealed that reading questions ahead provided listeners with contextual cues so they could focus on the relevant content more easily. However, low-proficiency learners compensated for their inferior linguistic skills with

their general knowledge, which often led to misunderstandings of the text. In other words, previewing the questions misled low-proficiency learners, making them fall back on their general background knowledge and choose the wrong answer options. Chang (2005) confirmed the finding that low-proficiency learners seem to benefit less from previewing questions because of their insufficient language skills.

As seen from the above studies, test administrators face a difficult and controversial decision on whether or not to allow test takers to preview test questions. Various factors such as option types and learner proficiency have been explored in terms of previewing questions. However, thus far, the influence of working memory span – a variable initially discussed only in psychology, but now applied in various fields – has not been studied in the context of listening tests. Before discussing the research design, the concept of working memory and its applications in language learning will be introduced.

2.2. Working memory

Working memory is defined as "the temporary storage of information in connection with the performance of other cognitive tasks such as reading, problem solving or learning" (Baddeley 1983). In the 1960s, cognitive psychologists studied patients with brain damage and concluded that memory is not a single, unitary concept, but instead, it consists of sub-parts. Patients who suffered from amnesic syndrome seemed to experience gross loss in their lasting memories but performed fairly well with tasks requiring short-term memory use. In contrast, there were patients who could retrieve information stored shortly before but faced problems when accessing their long-term memories. In this way, researchers began to acknowledge the workings of two different memory systems.

By the early 1970s, Atkinson and Shiffrin (1968) suggested that short-term memory actually acted as working memory, which is necessary for learning, retrieving, and performing cognitive tasks. This fact was confirmed by other researchers, and consequently, the single, unitary concept of short-term memory began to be replaced by the working memory system, which is a tripartite system (Figure 1) composed of the central executive system and two subsidiary slave systems: the phonological loop and the

visuospatial sketch pad. The central executive controls the two slave systems. It is believed that Alzheimer patients have difficulties executing central controls, which combine the two slave systems. Therefore, for instance, if they had to perform two tasks simultaneously – one visual and one verbal – their performance will be significantly poorer than if they had to complete tasks separately. Studies report that as the disease progresses, patients' performance of combined tasks markedly decreases. The concept of the phonological loop is the closest to traditional short-term memory and is therefore the most extensively researched using the memory-span procedure. It consists of two subparts – phonological store and articulation control. The phonological store retains acoustic or speech-based input for several seconds, and articulation control is similar to inner speech, an active sub-vocal rehearsal. The visuospatial sketch pad allows for the mental representation of input that is given visually or spatially.

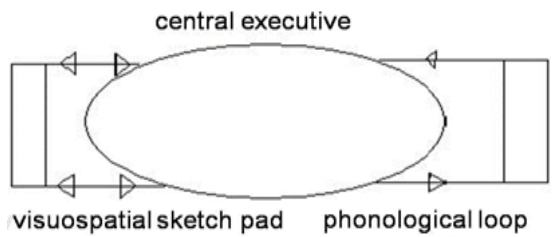

Figure 1. Simplified Representation of the Baddeley and Hitch Working Memory Model (Baddeley & Hitch, 1974).

Daneman and Carpenter (1980) argued that the differences between individuals' working memory spans are mostly due to the central executive system, which combines processing and storage, and is a powerful predictor of success in online language processing. Working memory was understood as a system necessary for storage and processing; now, the focus has shifted to how working memory can predict individual differences in cognitive tasks. The reading span task is one classical method developed by Daneman and Carpenter (1980). In this task, they examined how participants processed short sentences by either reading aloud or listening and at the same time memorizing the last words or certain cued

letters in each sentence. For example, in the listening span task, participants were made to listen to the sentence "Everybody needs to sing food and drink water." This was followed by a true–false question. When they check up the sentence, one letter, for example, "k" appears. Then, the second sentence, "This is a good paper full of insights," appears with a true–false option, followed by the letter "p." Participants then need to verbally say the two letters, "k" and "p," in sequence. This type of task ensures that participants are involved in processing the given information (i.e., evaluating the truthfulness of the sentence) and memorizing other cues (i.e., the letters "k" and "p"). Typically, the result of the task will be correlated with the results of a standard reading comprehension test with a coefficient of about 0.5 or 0.6. Daneman & Carpenter (1983) also revealed that students with high working memory spans cope better with garden path sentences and drawing inferences from texts.

The mechanism behind the working memory has been applied in studies on second language (L2) acquisition as well. In Baddeley, Papagno, and Vallar (1988) patients suffering from a very specific short-term phonological memory deficit were taught eight Russian words. They failed to learn the words, which were foreign to them, even with visual presentations. This suggests that short-term phonological storage is one of the crucial factors in learning foreign vocabulary. The most widely studied subcategory involves L2 reading comprehension. In her meta-analysis of previous studies on working memory and L2 reading, Oh (2011) concluded that the central executive system is a powerful predictor of L2 reading success. Oh did not explore in as much detail the facet of listening in relation to working memory. Jeong and Kim (2010) investigated the effects of working memory on English listening comprehension using two types of tasks: verbal information and texts of rules. They found that working memory appeared to positively influence listening to verbal information, but there was no significant influence in participants' performance when they listened to the texts of rules.

3. Research Question

If test takers are not provided with questions in advance, they need to retain the necessary information until the questions become available. At the same time, they need to comprehend L2 inputs. Working memory would then be activated when choosing the correct option from the multiple-choice answers. However, thus far, no study has provided empirical evidence on how this difference in working memory is related to individual task performance, especially when test takers are not allowed to preview questions.

For the testing material in this study, part 4 of the TEPS is used; three question types used in TEPS are separately examined, namely, looking for the main idea, finding details, and inference. No study has yet investigated the effects that different question types have on performance when test takers are allowed to preview questions. Therefore, it would be meaningful to see the effects of question types on task performance as well as their relationship to working memory span. This study, then, attempts to investigate the following questions.

1. To what extent does a test taker's performance in each listening test differ in two varying circumstances – one with questions written on paper and the other with only oral input? Further, does the question type (e.g., selecting topics, searching for detailed information, or inference) affect the result?
2. Does the difference in the scores of each individual between the question preview and no preview contexts correlate with that individual's working memory span? Does the question type (e.g., selecting topics, searching for detailed information, or inference) affect the result?

4 Method

4.1. Participants

Thirty-one undergraduate students – 25 female and 6 male – majoring in English education at Seoul National University participated in this study. All the participants had an advanced level of English proficiency. The test was implemented as a partial requirement in the English phonology class for sophomores.

4.2. Design

4.2.1. Listening Task

Two sets of questions were taken from part 4 of the TEPS listening tests. Each set contained 24 questions with three subcategories designed to elicit information on finding the main ideas, reporting details, and drawing inferences. Table 1 illustrates the test format. The length of the listening material was about 40 to 50 seconds, delivered in monologues, and the topics varied from economics to politics and education. In the first session, the test takers could not preview the question and answer options, similar to the actual TEPS test. In the second session, on the other hand, students were allowed to read the question and answer options in advance, as all this information was provided in the answer sheet. All the questions were randomly chosen from *It's TEPS Actual Test Practice*, published by *Educhosun* (2008), in order to avoid differences in the level of difficulty.

Table 1. Number of Test Items

	SESSION 1 (no preview)	SESSION 2 (preview)
MAIN IDEA	n = 8	n = 8
DETAIL	n = 8	n = 8
INFERENCE	n = 8	n = 8

4.2.2. Working Memory Task

After the first session, test takers undertook the conceptual span task, wherein they answered 20 questions; this task lasted approximately 15 minutes. Traditionally, working memory span has been measured by reading span tasks or listening span tasks. Using these methods, participants memorize interpolated letters while performing secondary tasks like reading or listening to a sentence. Since these traditional methods require participants to work on memorizing semantically unrelated stimuli in order and verbatim, an alternative method of measurement is used, wherein test takers memorize selections that are more meaningful. The conceptual span task, proposed by Haarmann, Davelaar, and Usher (2003), presents, for example, nine words, each belonging to three different categories, in random order; learners are asked to immediately recall the words belonging to one cued category, such as furniture, family, or fruit. In this study, the conceptual span task is employed. Similar to the case in Haarmann, Davelaar, and Usher (2003), a similar task is designed with four categories: furniture, animals, fruits, and jobs. For example, students were shown the words *chair*, *apple*, *nurse*, *cat* in sequence; each word appeared on a screen for one second and then disappeared. After all the words were displayed, the cued category *furniture* appeared on the screen. Then, the students had to remember and write down *chair* on the answer sheet. The entire task took about 15 minutes, including instructions and sample demonstration time.

4.3. Procedure

The first session, in which students could not preview the 24 questions, was followed by the working memory span test. One week later, the second session was given, with 24 preview questions. The gap time was inevitable because students reported high cognitive fatigue after the first test session. Students were allowed to take notes, similar to the case in the actual TEPS test. The average time duration for one question item was 60 seconds, and the total length of the test was about 25 minutes, including instructions.

4.4. Analysis

The scores for each part of the test as well as for the subcategories (i.e., main idea, details, and inference) were measured. First, a paired sample t-test was conducted to compare the mean difference between session 1 (without preview questions) and session 2 (with preview questions). The differences in the three subcategories – main idea, details, and inference – were also investigated. Second, the gap score (GS) was measured by subtracting the score for session 1 (without preview questions) from the score for session 2 (with preview questions). This was repeated with the three subcategories. Then, a correlation analysis was applied to compare the working memory span scores (WMSSs) with the GS. Third, a simple regression was used based on a standardized beta coefficient to present how working memory span affects GS. In this analysis, the WMSS was defined as the independent variable, and the GSs of the main idea, details, inference, and the total score were considered as dependent variables. All statistic procedures were conducted with SPSS 20.0 (for Windows, IBM Corporation, Armonk, NY, USA).

4. Results

1. To what extent does a test taker's performance in each listening test differ in two varying circumstances – one with questions written on paper and the other with only oral input? Further, does the question type (e.g., selecting topics, searching for detailed information, or inference) affect the result?

Table 2 shows the result of the paired sample t-test. It can be seen that the session 2 scores were significantly higher than those for session 1. In the first subcategory – searching for the main idea – the session 2 score (M = 7.17, SD = 0.90) was two points higher than the session 1 score (M = 5.13, SD = 1.44) on average, and it was statistically significant (t = -11.976, $p < 0.01$). In the second subcategory – looking for details – the average score for session 2 (M = 6.73, SD = 1.13) was 1.8 points higher

than that for session 1 (M = 4.93, SD = 1.99), showing a statistical significance with a p value of 0.01 (t = -8.836, p < 0.01). With regard to inference, the score for session 2 (M = 5.77, SD = 1.64) was 1.67 points higher than that for session 1 (M = 4.10, SD = 2.00). In terms of the total score, the score for session 2 (M = 19.67, SD = 2.93) was 5.5 points higher, on average, than that for session 1 (M = 14.17, SD = 4.47), and its p value was below 0.01 (t = -14.078, p < 0.01). Figure 2 illustrates the mean comparison using bar graphs.

Table 2. Mean Comparison: Session 1 Versus Session 2

Variable	Session 1 (no preview)		Session 2 (preview)		t	p
	Mean	SD	Mean	SD		
Main Idea	5.13	1.44	7.17	0.90	-11.976**	0.000
Detail	4.93	1.99	6.73	1.13	-8.836**	0.000
Inference	4.10	2.00	5.77	1.64	-7.477**	0.000
Total Score	14.17	4.47	19.67	2.93	-14.078**	0.000

**: p < 0.01, *: p < 0.05

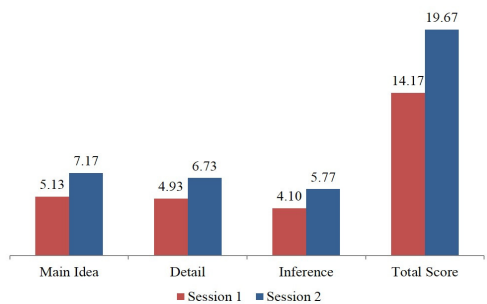

Figure 2. Mean Comparison: Session 1 Versus Session 2.

2. Does the difference in the scores of each individual between the question preview and no preview contexts correlate with that individual's working memory span? Does the question type (e.g., selecting topics, searching for detailed information, or inference) affect the result?

The GS was measured by subtracting the score for session 1 from that for session 2 (GS = session 2 score − session 1 score). The GS for the subcategory of finding the main idea had a positive correlation with the gap for total score (r = .517, $p < 0.01$), and it was statistically significant with a p level of 0.01. The GS for this subcategory had a negative correlation with working memory span (r = -.251, $p < 0.05$), and it was statistically significant with a p level of 0.05. The WMSSs also had a significant negative correlation with the GS for the subcategories of looking for details (r = -.270, $p < 0.05$) and inference (r = -.288, $p < 0.01$). The GS for the total score was also negatively correlated with working memory span (r = -.222, $p < 0.05$). Working memory span had the strongest correlation with the GS for inference. Table 3 shows the correlation analysis between GSs and working memory.

Table 3. Correlation Analysis between Gap Scores and Working Memory

Variable	(1)	(2)	(3)	(4)	(5)
Gap for Main Idea (1)	1				
Gap for Detail (2)	0.100	1			
Gap for Inference (3)	0.053	0.239*	1		
Gap for Total Score (4)	0.517**	0.701**	0.718*	1	
Working Memory (5)	-0.251*	-0.270*	-0.288**	-0.222*	1
Mean	2.03	1.80	1.67	5.50	21.43
Standard Deviation	1.61	1.93	2.11	3.71	4.12

**: $p < 0.01$, *: $p < 0.05$

A regression analysis was also conducted by setting the WMSS as the independent variable and the GS as the dependent variable. First, the WMSS had a significant influence on the GS for the subcategory of finding the main idea, with a p level of 0.05 (beta = -0.251, $p < 0.05$). The higher the WMSS, the lower is the GS for the main idea. Table 4 illustrates the effects of working memory on the GS for main idea.

Table 4. Effect of Working Memory on Gap Score for Main Idea
Dependent variable: Gap score for main idea

Variable	B	SE	Beta	*t*	*p*
Constant	4.138	0.880		4.700**	0.000
Working Memory	-0.098	0.040	-0.251	-2.434*	0.017

**: $p < 0.01$, *: $p < 0.05$

Second, the regression analysis revealed that the WMSS had significant effects on the GS for finding the details, with a p level of 0.05 (Beta = -0.270, $p < 0.05$). As shown in Table 5, the higher the WMSS, the lower is the GS for details.

Table 5. Effect of Working Memory on Gap Score for Details
Dependent variable: Gap score for detail

Variable	B	SE	Beta	*t*	*p*
Constant	3.094	1.082		2.859**	0.005
Working Memory	-0.124	0.050	-0.270	-2.480*	0.013

**: $p < 0.01$, *: $p < 0.05$

Third, the GS for the subcategory of inference was also influenced by the working memory span, with a p level of 0.01 (Beta = -0.288, $p < 0.01$). The higher the working memory, the lower is the GS for inference. Table 6 shows the effect of working memory on the GS for inference.

Table 6. Effect of Working Memory on Gap Score for Inference
Dependent variable: Gap score for inference

Variable	B	SE	Beta	*t*	*P*
Constant	2.547	1.190		2.139*	0.035
Working Memory	-0.141	0.055	-0.288	-2.564**	0.009

**: $p < 0.01$, *: $p < 0.05$

By comparing the beta values for the three question types, the effect size can be compared. The main idea was the least affected subcategory from

among the three, with a beta level of -0.251, followed by finding details, with a beta level of -0.270. Inference was the most affected, with a beta level of -0.288.

Table 7 presents the effect of working memory on the gap for total score. The gap for total score was affected by the WMSS, with a p level of 0.05 (beta = -0.272, $p < 0.05$). WMSS was negatively correlated with the gap for total score.

Table 7. Effect of Working Memory on Gap for Total Score
Dependent variable: Gap for total score

Variable	B	SE	Beta	*t*	*p*
Constant	9.779	2.041		4.792**	0.000
Working Memory	-0.234	0.094	-0.272	-2.489*	0.012

**: $p < 0.01$, *: $p < 0.05$

5. Discussion and Conclusion

This study aimed to explore the effect of working memory span on the gap score of two types of tests. This study revealed that the test types were significantly correlated with working memory span. Students with relatively short working memory spans performed much better when the test items were written on paper. Those with longer memory spans performed better in the absence of written cues. In terms of question types, all three types – main idea, details, and inference – were affected by the working memory span. Among the three subcategories, inference was shown to be the most affected by test takers' working memory span, followed by details and main idea.

The rationale behind providing oral input only is that the listening test will only test students' listening ability, not reading. However, is it possible to measure the "listening faculty" alone? This study showed that the working memory span was affected when written questions were

removed. Therefore, test administrators should take into account that some test takers with poor working memory spans will be at a significant disadvantage if they are made to memorize what they heard. If the test is a large-scale one, in particular, test creators should be more careful to consider the possible washback effects. In the lectures focused on preparing for TEPS listening, what instructors focus on the most is not the listening skill itself but the note-taking strategies students use to compensate for their unreliable memory. In a sense, it is like putting the cart before the horse. Listening tests should not be a memory game, an exhibition of shorthand skills, or a guessing game.

It can be argued that listening to monologues without any written cues is analogous to real-life situations, such as lectures, announcements, or news broadcasts. However, they are not the same in many respects. First, test items are much shorter, and they deal with more diverse topics. The unpredictability of the subsequent topic inevitably causes serious test fatigue, which can be damaging to test reliability. In addition, kinetic cues such as facial expressions, gestures, or eye contact are eliminated in the test setting, making the comprehension process far more complicated. Furthermore, listening to monologues in a test setting is completely non-interactive. In a real life lecture, listeners can signal a comprehension breakdown through puzzled looks or shrugs, for example. They can even ask questions. In a listening test, however, listening is a receptive skill only.

Therefore, in a test setting, students are required to pay focused attention for a long time. When questions were given in a textual input, the influence of the working memory span was greatly lessened, as test takers were better prepared for the upcoming topics and could listen selectively. In the absence of text input, however, test takers needed to listen for every detail, not knowing which information would be tested later on. On one hand, their minds work to encode the foreign language. On the other hand, they have to work to store content knowledge.

In conclusion, when questions without previews were delivered orally after students listened to monologues, the students experienced a great cognitive burden. The implication of this study lies in the new perspective

it provides on the role of written input in listening comprehension tests. Written input not only provides contextual cues but also reduces the gap in individual differences in working memory span. The results of this study suggest that written cues can compensate for the excessive burden placed on test takers with weaker working memory spans.

6. Limitations

One of the limitations of this study lies in the design of the test. To ensure that the two sets of the test have the same level of difficulty and to reduce practice effects, the participants should be divided into two groups, in a counterbalanced design. This method was not possible in this study, because the test was conducted as part of the regular course work requirements, and all the students had to be given the same test items. Future research is needed to test how the GS is correlated with working memory span in low level of speakers or other types of input such as dialogues or longer passages. In addition, a qualitative approach will be needed to provide a deeper understanding of the different effects of working memory span on two different types of listening tests.

References

Atkinson, R. C. and Shiffirin, R. M. (1968). Human memory: a proposed system and its control processes. *The psychology of learning and motivation: advances in research and theory* 2, 89-195.

Baddeley, A. D. (1983). Working Memory. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 302.1110, 311-324.

Baddeley, A. D. and Hitch, G. J. (1974). Working memory. *The psychology of learning and motivation* 8, 47-90.

Baddeley, A. D., Papagno, C., and Vellar, G. (1988). When long-term learning depends on short-term storage. *Journal of Memory and language* 27, 586-595.

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

Call, M. (1985). Auditory short-term memory, listening comprehension, and the Input Hypothesis. *TESOL Quarterly* 19, 765-781.

Chang, A. C. S. (2005). The perceived effectiveness of question preview in listening comprehension tests by EFL learners. New Zealand Stud. *Applied Linguistics* 11, 75-99

Chang, A. C.-A. and Read, J. (2006). The effects of listening support on the listening performance of EFL learners. *TESOL Quarterly* 40.2, 375-397.

Chang, A. C. S. and 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.

Daneman, M. and Carpenter, P. A. (1980).Individual differences in WM and reading. *Journal of Verbal learning and Verbal Behavior* 19, 450-466.

Daneman, M. and Carpenter, P. A. (1983). Individual differences in integrating information between and within sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 9, 561-584.

Ericsson, K. A. and Simon, H. A. (1987). *Protocol analysis*. (2nd ed.) Cambridge, MA: MIT Press.

Haarmann, H. J., Davelaar, E. J., and Usher, M. (2003). Individual differences in semantic short-term memory capacity and reading comprehnsion. *Journal of Memory & Language* 48, 320-345.

Jeong, J. and Kim, H. (2010). The effects of types of acvance organizer and working memory for English listening comprehension, *Korean Journal of Educational Methodology Studies* 22.2, 187-206.

Mackey, A., Adams, R., Stafford, C., and Winke, P. (2010). Exploring the relationship between modified output and working memory capacity. *Language Learning* 60.3, 501-533.

Oh, E. J. (2011). A review on a construct of working memory and its role in L1 and L2 reading comprehension. *English Teaching* 66.1, 3-22.

Wu, Y. (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. and 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.

Bora Kim
Department of English Education
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul
E-mail: borahkim@gmail.com

Received: June 30, 2015
Revised version received: August 4, 2015
Accepted: August 20, 2015