

## The Effects of Different Lengths of Planning Time on Integrated Speaking Test Scores: The Mediating Roles of the Rhetorical Structure of Reading Passages and Test Wiseness Levels

Koo, Jinseon

Ewha Womans University

### ABSTRACT

*The Journal of Studies in Language* 36.1, 041-058. This paper aims to investigate the difference when different planning time lengths, different rhetorical structures used in reading passages, and different test-takers' test wiseness levels are the variables to be considered. Twenty-six college students took the integrated speaking tests with three different planning time lengths of 30, 60 or 120 seconds and completed survey questions on perceptions regarding the different planning time lengths given. The results indicated that there was no significant difference in test performance. The participants' test-wiseness level did affect the test-takers' performance significantly. In contrast, the speaking test scores for both experienced and novice participants were not affected by the differences in planning time length. The questionnaire revealed that 30-seconds of planning time was generally considered too short by most students, and that novice students preferred more planning time than experienced students. (Ewha Womans University)

**Keywords:** integrated speaking test, planning time length, rhetorical structure of reading passage, test wiseness, EFL learners



<https://doi.org/10.18627/jslg.36.1.202005.041>

pISSN : 1225-4770

eISSN : 2671-6151

**Received:** April 01, 2020

**Revised:** May 12, 2020

**Accepted:** May 18, 2020

This is an Open-Access article distributed under the terms of the Creative Commons Attribution NonCommercial License which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright©2020 the Modern Linguistic Society of Korea

본인이 투고한 논문은 다른 학술지에 게재된 적이 없으며 타인의 논문을 표절하지 않았음을 서약합니다. 추후 중복게재 혹은 표절된 것으로 밝혀질 시에는 논문게재 취소와 일정 기간 논문제출의 제한 조치를 받게 됨을 인지하고 있습니다.

### 1. Introduction

Speaking performance evaluations should solely reflect test-takers' speaking abilities. If a test-taker's speaking performance is shown to fluctuate because of factors other than an examinee's own speaking abilities, the examination would not be truly valid, and the resulting scores would not be an accurate measure of their speaking skills. Therefore, outside factors and influences, defined here as factors other than test-takers' speaking abilities, should be controlled as thoroughly as possible in order to enhance the validity of any speaking tests. The purpose of this study was to investigate the effects that different planning time lengths, the rhetorical structure of reading passages, and test wiseness had on

integrated speaking test performance, and how test-takers planning behaviors changed depending on the planning time length they were provided with.

It is widely understood that a certain amount of planning time is required for a test-taker to formulate an answer in a foreign language exam for test authenticity reasons. In real-life situations, people often don't plan what they will say before they speak. They do, however, often draw from shared information and the circumstances that they find themselves involved in. In a test context, a test-taker must answer questions instantly on a random topic, often without any background or context. This can cause difficulty with the production of accurate and intelligible speech for many reasons. To help overcome these limitations, a certain amount of planning time is allowed in speaking tests. However, there are no specific guidelines regarding how much planning time should be given to a test-taker for a certain task. Thus, how much planning time should be allocated for a task-taker is a critical issue which should be studied thoroughly.

There has been much research conducted exploring the impact of planning on language production in the area of second and foreign language acquisition (Foster, 1996; Foster and Skehan, 1996; Mehrang and Rahimpour, 2010; Mochizuki and Ortega, 2008). A principal conclusion of previous studies was that planning time is beneficial to learners. Therefore, it has been an important issue in language learning. Within the testing field, there has been some limited research conducted investigating the effects of planning time (Elder and Iwashita, 2005; Frost et al., 2011; Lee, 2018; Li et al., 2015; Wigglesworth, 1997, 2000, 2001; Wigglesworth and Elder, 2010). However, these studies have largely failed to show consistent results. Some studies have found that the provision of planning time positively affected learners' performance (Li et al., 2015), while other studies found planning time had no impact on subjects' test scores (Elder and Iwashita, 2005; Lee, 2018; Wigglesworth, 1997, 2001, 2010).

It should be noted that there has been few research, which has solely investigated planning time length effects with integrated tests (Lee, 2018). Integrated tests are more complex methods of testing, as compared to traditional independent testing methods, where language skills are tested separately, so integrated test performance may respond more sensitively to different planning time lengths. Therefore, studying the effects of planning time length in integrated tests is a subject which remains open to new investigation.

In this study, a secondary component to the pre-task planning time variable is to look at the effects of rhetorical structures of reading passages on performance. An integrative test is one that requires a test-taker to use several language skills at the same time (Richards and Schmidt, 2013). An example of this is the TOEFL iBT speaking test, which requires a test-taker to use reading skills, listening skills, and speaking skills simultaneously. The rhetorical structure of the reading passage is another critical component of integrated speaking tests, as performance also can be influenced by the rhetorical structure of the reading passage presented. There have been many studies, which concluded that rhetorical structure affects reading comprehension (Carrell, 1984, 1985; Kobayashi, 2002; Kintsch and Yarbrough, 1982; Luoma, 2004; Mann and Thomspon; 1987). Therefore, if rhetorical structure affects test performance seriously, it's a critical issue and should be examined thoroughly to determine if test-taker's performance is affected by the rhetorical structures of reading passages in integrated speaking tests, when different amounts planning time are given.

This paper also aims to investigate whether the effects of planning time are related to test-takers' test wiseness levels. Test wiseness refers to the amount of overall experience in taking tests a test-taker may have - if they have previously

taken many tests - or in contrast are novice in test taking (Richards and Schmidt, 2013). Previous studies have established that experienced test-takers who are familiar with a certain test method have advantages over novice test-takers, due to their testing experience (English and English, 1970; Gibb, 1964; Oakland, 1972). As a result, it can be hypothesized that when a group of test-takers with an equal proficiency level are given the same exact test, with the same amount of planning time, more experienced test-takers might be affected less than novice test-takers by the amount of planning time given. More study in this matter is needed to be conducted specifically in integrated speaking tests, as integrated tests require the use of more complex language skills.

Finally, a number of studies have solely focused on the effects of planning time length on independent test performance (Elder and Iwashita, 2005; Frost et al., 2011; Lee, 2018; Li et al., 2015; Wigglesworth, 1997, 2000, 2001; Wigglesworth and Elder, 2010). In contrast, other than simply studying test-takers' performance fluctuations as it related to different amounts of planning time, in the present study, data were also analyzed according to the different planning time lengths, to look at how test-takers used their planning time, and to determine if any conclusions could be drawn from their study.

## 2. Literature Review

### 2.1 Research on the Effects of Different Planning Time Length on Speaking Test Performance

Much research has been conducted exploring the impact of planning on language production in the area of second language acquisition. A number of researchers have studied the differences that planning time has on the three central aspects of language performance, such as, fluency, accuracy and complexity (Ellis, 1987, 2005; Crookes, 1989; Foster, 1996; Foster and Skehan, 1996; Gilabert, 2007; Iwashita et al., 2001; Kawauchi, 2005; Mehrang and Rahimpour, 2010; Mochizuki and Ortega, 2008; Rutherford, 2001; Sangarun, 2005; Skehan and Frost, 2005; Tajima, 2003; Yuan and Ellis, 2003). One of the classroom studies conducted by Foster and Skehan (1996) concluded that planning time benefited complexity and fluency rather than accuracy. In another classroom study, Skehan and Foster (1999) also found out that there are clear effects on providing greater planning opportunities. In fact, offering planning time had positive effects on almost all measures. On the other hand, Rutherford (2001) found no effects of planning on accuracy. Though, Gilabert (2007) concluded planning conditions showed greater fluency and lexical richness than conditions that provided no planning time, but there were no effects on complexity. Furthermore, Tavakoli and Skehan (2005) reported that there were significant differences in fluency and complexity as related to amount of planning time. Guided planning groups given five minutes of planning time produced more accurate language than both a group given no planning time and an unguided group given five minutes of planning time. The studies have widely shown that planning played a positive role in improving the fluency and complexity of learners' output (Foster and Skehan, 1996; Ortega, 1999; Tavakoli and Skehan, 2005). However, the prior studies haven't provided consistent or conclusive results regarding the effects of planning tasks.

In a testing context, as compared to a regular classroom situation, an additional factor to consider is that test-takers might react differently in a test situation versus a classroom setting, depending on their given planning time, as students regularly express that they may not have done as well on a test as they could have, due to the simple pressure of being

tested. Only a few studies have explored the effects of planning in speaking assessment contexts (Elder and Iwashita, 2005; Frost et al., 2011; Lee, 2018; Li et al., 2015; Wigglesworth, 1997, 2000, 2001; Wigglesworth and Elder, 2010). For example, Wigglesworth (1997) investigated the effects of planning in relationship to speaking test discourse. She controlled the amount of planning time, proficiency level and task difficulty, and discovered that one minute of pre-task planning time helped students improve their production of proper grammar. Wigglesworth's study results showed that more planning time resulted in greater speech fluency. More recently, Wigglesworth and Elder (2010) investigated the effectiveness of planning time on performance on the IELTS oral interview. There were no significant differences found in either the amount of strategic planning time or the quality of discourse.

Elder and Iwashita (2005) found that planning time has no significant effect on complexity and accuracy in a testing situation. Further work by Li, Chen and Sun (2015) sought to analyze the effects of different lengths of pre-task planning time on performance, and also looked into measurements of complexity, fluency and accuracy. Results from this study showed that the degree of improvement between variable planning time differences of the 1-minute, 2-minute, 3-minute and 5-minute conditions were less significant. They showed that 1 to 3 minutes of planning time is optimal to improve accuracy and fluency. Significant differences were also detected in all three aspects of complexity, fluency and accuracy when the additional planning times were provided. They also found that too much planning time elicited diminishing effects. On the other hand, Mehnert (1998) found diminishing improvements with additional planning time. Studies in language testing have produced mixed and inconsistent results (Elder and Iwashita, 2005; Frost et al., 2011; Lee, 2018, Menert, 1998; Li et al., 2015; Wigglesworth, 1995, 1997, 2000, 2001; Wigglesworth and Elder, 2010). It is certain that more study in this area is warranted.

## 2.2 Research on the Effects of Rhetorical Structure on Speaking Test Performance

A further variable of this study is the rhetorical structure of the reading passages presented in a reading passage. Many research studies have proved that the rhetorical structure of text materials affects the reading comprehension of readers (Carrell, 1984, 1985; Kobayashi, 2002; Kintsch and Yarbrough, 1982; Luoma, 2004; Mann and Thomspen, 1987). Carrell (1984) investigated the effects of rhetorical organization on ESL students and concluded that certain more highly-structured English rhetorical patterns are generally more facilitative of recall for non-native readers. Kintsh and Yarbrough (1982) also studied the role that rhetorical form plays in text comprehension. They reported that rhetorical structure and structural complexity affected the participants' performance to answer topic questions and main idea questions.

More specifically, there is some research, which investigated main idea placement effects. Finding out the central idea in a reading passage has been regarded as essential to reading comprehension (Afflerbach, 1990). When the main idea of a text is not stated explicitly, the reader cannot simply select the main idea statement from surrounding sentences. Construction of the main idea is considered crucial to the comprehension of text (Van Dijk and Kintsch, 1983).

Many studies have shown that the rhetorical organization of text materials affects second language reading comprehension in the area of reading contexts. Not many studies, though, have focused on this subject in integrated speaking tests. In some integrated speaking tests, test-takers are required to answer a question according to what they

have read and listened to. In the tests, being able to find the main idea of a written passage in a short period of time will be helpful to the test-taker. If test-takers encounter an easily-structured passage, they can comprehend it quickly. This is particularly important to consider in integrated testing, where students are required to use and understand various communication methods at the same time. Subsequently, more research is needed especially in the area of integrated tests, since test-takers' performance can be influenced differently by rhetorical structure, when they are given varying amounts of planning time.

### 2.3 Research on the Effects of Test Wiseness on Speaking Test Performance

Test wiseness is defined as “a test-taking skill, which enables a person to do well on certain kinds of tests by using their familiarity with the characteristics and formats of exams to help them guess the correct answer (Richards and Schmidt, 2013).” Generally higher-proficiency level students should outperform lower-level students in any language tests, but the reason why they do so should be because of their higher-level of English proficiency, and not because they are simply more familiar with the test structure. Experienced test-takers who are used to the testing format can more efficiently utilize their learned strategies effectively within the given time. On the other hand, novice test-takers may feel additional anxiety and pressure, as they encounter a test format that is not familiar to them.

From reviewing test wiseness research literature, it has been previously shown that test wiseness is a factor creating differential effects in testing contexts (Allan, 1992; Houston 2005; Millman et al., 1965). Many studies have been conducted on test wiseness in the reading area. Nonetheless, only a few research studies related to test wiseness have been done in the field of English speaking tests. Allan (1992) investigated the validation of a scale to measure test wiseness in reading skills, with multiple-choice questions. He found that the reading results of some learners were influenced by test wiseness skills, which were not the focus of the test, and thus, the true measurement of their speaking skills were invalidated. Ford (1973) and Houston (2005) suggested that making a good test wiseness program and implementing it in the regular classroom curriculum would be highly beneficial in assessing overall language ability.

Many researchers have suggested test wiseness as a component to be considered when creating language tests (Allan, 1992; English and English, 1970; Gibb, 1964; Oakland, 1972). But, it is not clear whether performance remains the same with different amounts of planning time, regardless of test-takers' familiarity with the integrated speaking tasks presented to them. For these reasons, the present study also focuses on the issue of test wiseness and its effects on language testing.

Considerable prior studies have been done, which have focused on studying each independent variable separately. In contrast, the focus of this research project was aimed at examining the interaction of all three variables altogether, different lengths of planning time, the rhetorical structure of reading passages and test wiseness. It was designed to examine the effects that the rhetorical structure of reading passages and test wiseness have on speaking test performance, when participants were given varying amounts of pre-task planning time. In addition to the quantitative research on the three dependent variables in this study, a qualitative analysis regarding a survey of test-takers' opinions on ideal planning time length were both examined with survey responses. The research questions are as follows:

- (1) Does the different amount of planning time provided affect test-takers' performance in integrated speaking tests?
- (2) Do the effects of the amount of planning time on test performance differ by the rhetorical structure of reading passages?
- (3) Does test performance differ between novice and experienced test-takers when given different amounts of pre-task planning time?
- (4) What are test-takers' perceptions of the different amount of planning time given?

### 3. Methodology

#### 3.1 Participants

##### 3.1.1 Test-Takers

Research data was collected from 26 Korean test-takers, who had all previously taken official English speaking tests, such as TOEFL, TOEIC speaking tests or the OPIc exam. The data is included only for those who had official standardized English-speaking test scores prior to this testing, in order to help study the effects of participants' proficiency level on their performance. The 26 participants were an average of 22-year-old. They were all female, had all studied in an EFL setting at one time, and were living in Korea at the time they took the test.

##### 3.1.2 Raters and a Computer Programmer

One native English rater and one Korean rater were recruited, both have extensive teaching and rating experience with adult English learners in an EFL context. A computer programmer was hired to develop a new on-line speaking test for this research, and a speaking test computer program was developed solely for the purpose of conducting this research.

#### 3.2 Instruments

##### 3.2.1 The Speaking Tests

The questions were taken from official TOEFL iBT Tests guidebooks (2016; 2017) published by ETS. While test-takers took the exam, they were allowed to take notes on a given piece of paper to help them prepare their answers. The participants were given six integrated speaking questions.

The test format for this research was identical to existing iBT TOEFL speaking tests, with the only difference being that this current test varied the length of planning time given for each item. The test-takers were required to answer questions using the information given to them in the reading and spoken passages.

In the iBT TOEFL integrated test, test-takers were required to read a short, written passage, for 45 or 50 seconds, and then listen to a class lecture for approximately 90 seconds. According to the structure of the integrated tests, after reading and listening to the passage, they were presented with a question, and were given 30 seconds of preparation time

to answer that question. One varying factor was that test-takers were given a different amount of planning time, either 30 seconds, 60 seconds or 120 seconds.

Additionally, in order to investigate the effects that the rhetorical structure of the written passages had on test performance, test-takers answered six questions. The six integrated questions were presented in two different written structures. Specifically, three out of six reading passages (Questions 1, 3 and 5) had key topic sentence at the beginning of the paragraph, while the other three texts (Questions 2, 4 and 6) had key topic sentences placed later in each passage. There are examples of two types of reading passages, as shown in the Appendix 1.

### 3.2.2 Questionnaire

A survey was administered after the testing was finished. to study their views on planning time length. The following three questions were asked in the questionnaire, and these questions aimed at examining participants' perceptions regarding the different planning time lengths. The first open-ended question was: 'Was the planning time beneficial?' What is the reason you answered 'yes' or 'no'? The second question was: 'To bring out your best abilities, how much planning time do you think is needed?' The last question was: 'Please state the reason, or reasons, that you chose the specific amount of time for the second question?'. These questions are adapted from the questionnaire developed by Wigglesworth and Elder (2010).

## 3.3 Procedures

### 3.3.1 Grouping Test-Takers and Counterbalancing Planning Time Order

Using the speaking test scores that the participants submitted, the participants were divided into two levels: advanced and intermediate. These groupings were based on either the test-takers' TOEFL speaking test or TOEIC speaking test scores, using the score-band cut-off range used by Hyun (2017). Test-takers who submitted OPIc test scores were grouped according to the ACTFL classification guidelines as either intermediate or high-level proficiency.

To answer research question one, the order of the planning times was counterbalanced. Test-takers were given different planning time lengths for each question item. Some test-takers were given decreasing planning times and others will take increasing planning time lengths. As it is possible that performance may also be affected by the order in which the different planning time lengths were given at the beginning, three sets of tests were created to help randomize the order in which the test-takers received their different amounts of planning time. Set one has increasing planning time (e.g., 30, 60 to 120 seconds.). Set two has decreasing planning time (e.g., 120, 60 to 30 seconds.) and set three has random planning time (e.g., 120, 30 to 60 seconds). Each set had identical reading, listening and speaking questions. The only difference between test-takers was the planning time given for each participant's item.

As mentioned earlier, the 26 participants were first divided into two groups, advanced and intermediate, according to their official standardized speaking test scores, and were then divided into three groups again by sets, as shown in Table 1.

**Table 1.** Test-taker groupings by sets

Numbers of Sets	Proficiency Groups	<i>N</i>
Set 1	Advanced	5
	Intermediate	4
Set 2	Advanced	4
	Intermediate	5
Set 3	Advanced	4
	Intermediate	4
Total		26

Additionally, for research question three, which looked into whether the effects of planning time are related to test-takers' test wiseness level, test takers were divided further into two groups; novice test-takers and experienced test-takers, as Table 2 presents. If test-takers have ever taken the TOEFL iBT test before, they were assigned to the experienced group. If not, they belonged to the novice group. There were seven TOEFL iBT experience test-takers who had taken a TOEFL iBT test in the last two years.

**Table 2.** Test-taker groupings by test wiseness and proficiency level

Proficiency Level	Test Wiseness Level	<i>N</i>
Advanced	Experienced	7
	Novice	6
Intermediate	Experienced	6
	Novice	7
Total		26

### 3.3.2 The Speaking Tests and Questionnaire

In this study, 26 students answered six integrated questions during the speaking tests. After being given time to prepare their response, they had to record themselves speaking in a given time. Their answers were automatically recorded on the computer during the exams. After the speaking tests were completed, students filled out the questionnaire in Korean, so that all students had a reasonably equal opportunity to express their full thoughts. It took approximately 20 minutes for the test takers to complete all of the questions in the questionnaire.

### 3.3.3 Data Analysis

Based on TOEFL iBT rubrics, there were four rating areas used: general description, delivery, language use, and topic development. The one Korean and one English native rater analyzed the responses in four categories, which were given equal importance in this study. Inter-rater reliability was computed using Pearson's correlation coefficient, to compare the ratings produced by the two raters, and there was a statistically significant correlation (.74) between the two raters. Thus, the two raters' rating results were shown to be reliable and useful for this study.

To answer research question one, which focused on if the amount of planning time provided made a difference to test-takers' performance in the integrated speaking tasks, a mixed repeated-measures ANOVA was performed to

examine. In addition, a Paired Sample T-test was performed to examine if test performance differed by the rhetorical structure of the reading passages. A Paired Sample T-test was performed to examine if test performance differed by the rhetorical structure of the reading passages. Rather than using a two-way ANOVA a Paired Sample T-test was chosen to use to compare the results of two mean scores of the test-takers, where the two kinds of topic placement passages were placed either at the beginning, or anywhere else later in the paragraph. To answer research question three about whether test wiseness and proficiency level affect test performance statistically when given different amounts of planning time, a mixed repeated-measures ANOVA was run.

## 4. Results

### 4.1 Effect of Amount of Planning Time on Speaking Test Performance

#### 4.1.1 Descriptive Statistics for Integrated Test Scores

In Table 3, descriptive statistics demonstrate the test scores obtained by the participants when they took the integrated speaking tests, with three different lengths of planning time. The mean, standard deviation, and the number of subjects in each group are shown in this table.

**Table 3.** Descriptive statistics of integrated speaking test scores

Scoring Categories	Proficiency Level	Planning Time (sec.)	<i>M</i>	<i>SD</i>	<i>N</i>
General Description	Advanced	30	3.53	.477	13
		60	3.53	.406	13
		120	3.67	.328	13
	Intermediate	30	2.63	.617	13
		60	2.75	.735	13
		120	2.88	.960	13
Delivery	Advanced	30	3.48	.504	13
		60	3.57	.328	13
		120	3.73	.259	13
	Intermediate	30	2.71	.660	13
		60	2.63	.642	13
		120	2.75	.883	13
Language Use	Advanced	30	3.51	.572	13
		60	3.59	.415	13
		120	3.73	.296	13
	Intermediate	30	2.78	.894	13
		60	2.78	.675	13
		120	2.75	.797	13
Topic Development	Advanced	30	3.51	.590	13
		60	3.48	.450	13
		120	3.53	.431	13
	Intermediate	30	2.71	.652	13
		60	2.76	.672	13
		120	2.88	.904	13

As presented in Table 3, for all four scoring categories, the mean scores mostly indicated a gradual increasing pattern when they had more planning time. However, neither the intermediate groups' delivery and language use, nor the advanced group's topic development demonstrated a gradual increase as the planning time increased.

#### 4.1.2 Analysis of Planning Time Length Difference on Integrated Test Performance

A mixed repeated-measures ANOVA was calculated to investigate how the amount of planning time provided made a difference on test-takers' performance in the integrated speaking tasks, setting the test-taker's planning time length as a within-subjects factor and proficiency level as between-subjects factors. Prior to conducting this, Mauchly's sphericity test was performed. The result showed that the assumption was not violated, indicating that the variances of the three scores were equal. Table 4 provides the results of the mixed repeated-measures ANOVA. It was conducted to determine whether there were any statistically significant differences in the average test scores obtained by the participants with the three different planning times.

**Table 4.** Multivariate statistics for integrated test scores with different planning time lengths

Scoring Categories	Source	Type III Sum of Square	<i>df</i>	<i>Mean Square</i>	<i>F</i>	<i>P</i>
General Description	Proficiency Level	13.334	1	13.334	13.827	.001
	Planning Time	.506	2	.253	2.421	.100
	Planning Time x Proficiency Level	.058	2	.029	.276	.760
Delivery	Proficiency Level	15.705	1	15.705	18.485	.000
	Planning Time	.338	2	.169	1.874	.165
	Planning Time x Proficiency Level	.165	2	.083	.915	.407
Language Use	Proficiency Level	13.542	1	13.542	13.682	.001
	Planning Time	.079	2	.039	.299	.743
	Planning Time x Proficiency Level	.242	2	.121	.921	.405
Topic Development	Proficiency Level	10.232	1	10.232	10.233	.004
	Planning Time	.146	2	.073	.674	.514
	Planning Time x Proficiency Level	.079	2	.039	.363	.698

As Table 4 presents, the results of test of between-subjects effects showed significant difference among the test scores for general description, delivery, and language use categories. However, there was no statistically significant interaction between planning time and integrated test scores. Conclusively varying planning time did not influence test scores when participants were given integrated speaking tests with different amounts of planning time.

## 4.2 The Effect of the Rhetorical Structure of Reading Passages on Test Performance

Integrated tasks have a reading passage element, so integrated test performances were used to test the rhetorical structures of the passages used with different amounts of planning time. To examine if test performance differed by the

rhetorical structure of the reading passages, a Paired Sample T-test was performed. The results are shown in Table 5. The two different rhetorical structures were not found to show a statistically significant difference in participants' performance. To sum up, the effects of amount of planning time on the test performance of four scoring categories do not differ by the different types of the rhetorical structure of reading passages.

**Table 5.** A paired sample t-test summary table for integrated speaking test performance with different rhetorical structures for reading passages and planning time lengths

Planning Time (sec.)	Scoring Categories	Proficiency	Beginning Topic Sentence Passage		Other Placements Topic Sentence Passage		<i>t</i>	<i>p</i>
			<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
30	General Description	High	3.60	.893	3.80	.44	-1.00	.37
		Intermediate	2.75	.523	2.91	.97	-.79	.46
30	Delivery	High	3.50	.503	3.50	.50	.00	1.00
		Intermediate	2.68	.373	2.62	.58	.35	.73
30	Language Use	High	3.50	.703	3.60	.54	-1.00	.37
		Intermediate	2.81	.53	.262	.64	1.15	.28
30	Topic Development	High	3.50	.86	3.90	.22	-1.37	.24
		Intermediate	2.93	.77	.28	.84	1.00	.35
60	General Description	High	3.90	.22	3.40	1.08	1.29	.26
		Intermediate	3.00	1.03	3.12	.60	-.32	.75
60	Delivery	High	3.60	.41	3.30	.83	1.17	.30
		Intermediate	2.68	.59	2.87	.44	-1.15	.28
60	Language Use	High	3.50	.50	3.40	.89	.40	.70
		Intermediate	2.81	.25	2.87	.44	-.42	.68
60	Topic Development	High	3.90	.22	3.30	1.03	1.23	.28
		Intermediate	2.87	.99	3.00	.59	-.38	.71
120	General Description	High	3.80	.27	3.90	.22	-1.00	.37
		Intermediate	3.25	1.00	2.93	1.05	1.66	.14
120	Delivery	High	3.50	.50	3.90	.22	-2.13	.09
		Intermediate	2.93	.82	2.62	.79	1.25	.25
120	Language Use	High	3.30	.83	3.60	.41	-1.50	.20
		Intermediate	2.81	.53	2.87	.69	-.35	.73
120	Topic Development	High	3.60	.41	3.90	.22	-1.50	.20
		Intermediate	3.18	.96	2.97	1.09	2.37	.05

### 4.3 The Effect of Test Wiseness on Integrated Test Scores

To find out whether the effects of planning time were related to test-takers' test wiseness level, participants were divided into two groups, novice test-takers and experienced test-takers. The effects of test wiseness on integrated task were analyzed respectively. Table 6 presents descriptive statistics of integrated speaking test scores differing by test wiseness and proficiency levels. All novice test-takers' mean scores were much lower than the more experienced participants' test scores across all proficiency levels, which shows that experienced test-takers outperformed those who have no experience in identical English test formats.

**Table 6.** Descriptive statistics of integrated test scores differing by test wiseness and proficiency levels

Scoring Categories	Proficiency Level	Planning Time (sec.)	Test Wiseness	<i>M</i>	<i>SD</i>	<i>N</i>
General Description	Advanced	30	Experienced	3.64	.45	7
			Novice	3.41	.51	6
		60	Experienced	3.67	.34	7
			Novice	3.37	.44	6
		120	Experienced	3.46	.39	7
			Novice	2.82	.81	6
	Intermediate	30	Experienced	3.08	.51	6
			Novice	2.25	.40	7
		60	Experienced	3.20	.29	6
			Novice	2.35	.78	7
		120	Experienced	3.75	.32	6
			Novice	3.58	.34	7
Delivery	Advanced	30	Experienced	3.78	.39	7
			Novice	3.12	.37	6
		60	Experienced	3.71	.30	7
			Novice	3.41	.30	6
		120	Experienced	3.75	.25	7
			Novice	3.70	.29	6
	Intermediate	30	Experienced	3.20	.45	6
			Novice	2.28	.48	7
		60	Experienced	3.00	.35	6
			Novice	2.32	.68	7
		120	Experienced	3.33	.58	6
			Novice	2.25	.80	7
Language Use	Advanced	30	Experienced	3.78	.36	7
			Novice	3.20	.64	6
		60	Experienced	3.78	.26	7
			Novice	3.37	.46	6
		120	Experienced	3.78	.30	7
			Novice	3.66	.30	6
	Intermediate	30	Experienced	3.50	.83	6
			Novice	2.21	.41	7
		60	Experienced	3.16	.30	6
			Novice	2.46	.75	7
		120	Experienced	3.37	.54	6
			Novice	2.21	.54	7
Topic Development	Advanced	30	Experienced	3.64	.47	7
			Novice	3.37	.72	6
		60	Experienced	3.64	.49	7
			Novice	3.29	.33	6
		120	Experienced	3.67	.34	7
			Novice	3.37	.49	6
	Intermediate	30	Experienced	3.16	.54	6
			Novice	2.32	.47	7
		60	Experienced	3.08	.20	6
			Novice	2.50	.82	7
		120	Experienced	3.29	.60	6
			Novice	2.53	1.01	7

It was hypothesized that novice test-takers' performance would be more likely to be affected positively when they were given different amounts of planning time. To investigate if test wiseness and proficiency level affected test performance, a mixed repeated-measures ANOVA was run. As you can see in Table 7, there was a significant difference found in test wiseness levels in this study, which indicates that a test-takers' test wiseness level affected their test performance in a statistically significant way. On the contrary, in this study participations' test scores were generally not influenced by different planning time lengths. There were, however, statistically differences shown in the delivery category among the four scoring criteria, in the novice test wiseness group. This result was consistent in both intermediate and advanced participants. This finding revealed that both advanced and intermediate novice test-takers' delivery performance improved gradually when participants were given longer planning time.

**Table 7.** Multivariate statistics for planning time length, test wiseness and proficiency level on integrated test scores

Scoring Categories	Source	Type III Sum of Square	df	Mean Square	F	p
General Description	PL	11.93	1	11.93	17.08	.00
	TW	5.83	1	5.83	8.35	.00
	PL x TW	1.94	1	1.94	2.77	.11
	PT	.52	2	.26	2.29	.11
	PT x PL	.06	2	.03	.26	.77
	PT x TW	.00	2	.00	.03	.96
	PT x PL x TW	.03	2	.01	.14	.86
Delivery	PL	14.01	1	14.01	26.68	.00
	TW	7.31	1	7.31	13.92	.00
	PL x TW	1.52	1	1.52	2.91	.10
	PT	.40	2	.20	2.60	.08
	PT x PL	.19	2	.09	1.27	.29
	PT x TW	.32	2	.16	2.07	.13
	PT x PL x TW	.57	2	.28	3.65	.03
Language Use	PL	11.75	1	11.75	22.00	.00
	TW	9.75	1	9.75	18.25	.00
	PL x TW	2.24	1	2.24	4.20	.05
	PT	.09	2	.04	.40	.67
	PT x PL	.27	2	.13	1.13	.33
	PT x TW	.50	2	.25	2.06	.13
	PT x PL x TW	.45	2	.22	1.87	.16
Topic Development	PL	9.08	1	9.08	11.12	.00
	TW	5.19	1	5.19	6.36	.01
	PL x TW	.85	1	.857	1.05	.31
	PT	.14	2	.073	.63	.53
	PT x PL	.075	2	.038	.32	.72
	PT x TW	.027	2	.014	.11	.88
	PT x PL x TW	.099	2	.049	.42	.65

\* Planning Time (PT) Proficiency Level (PL) Test Wiseness (TW)

#### 4.4 Amount of Planning Time Effects on Perceptions of the Examinees

Based on the students' questionnaire responses, examinees' perceptions on the planning time length was analysed. The open-ended question was: 'To help you answer to the best of your ability, how much planning time was appropriate for the integrated speaking test?' The answers were recorded and the mean score was calculated from them. The mean time preferred by the experienced test-takers was 55 seconds, while the novice group mean's preferred planning time was much longer, at 82 seconds.

Most of the participants stated that 120 seconds of planning time was too long and contributed to them feeling too relaxed. 30 seconds was generally thought to be too short to prepare their answers fully. Two advanced participants mentioned that when given planning time that was too long, they just read the script more closely to try to find the answers, so it was more like a reading test than a speaking test for them. An advanced-level participant mentioned that the allotted planning time was too long, so she didn't do anything for the last 30 seconds and she felt too nervous to wait to start recording an answer. On the other hand, one novice test-taker mentioned that if she could control her planning time length for her answers, she could have performed much better.

### 5. Discussion and Conclusions

The purpose of this research was to investigate the effects that different planning time lengths, the rhetorical structure of reading passages, and test wiseness had on integrated speaking test performance. The results of the mixed repeated-measures ANOVA indicated that there was no significant statistical difference found between planning time and integrated speaking test scores. The results revealed that the differences in planning time for both advanced and intermediate participants did not prove to be statistically significant in this particular study for all scoring categories: general description, delivery, language use and topic development. This finding is consistent with studies done by Iwashita, McNamara and Elder (2001), Wigglesworth (2000), and Wigglesworth and Elder (2010), in contrast to other studies, which found that giving more planning time results in better performance (Li et al., 2015). Therefore, it is suggested that test-takers were not able to use their planning time to their best advantage with the limited materials available to them, since the reading and listening part of the test are transitory and participants were primarily obliged to rely on their notes and their memory.

The outcome for research question one revealed that different planning time lengths did not affect test performance. However, it cannot be concluded that test-takers had sufficient planning time to prepare their answer in the integrated speaking test for the face validity reasons. Despite the results, survey answers revealed that all participants expressed preference for longer amounts than 30 seconds of planning time. The questionnaire also revealed that the novice group preferred a much longer planning time (82 seconds) than the experienced group did (55 seconds).

The results of the study also showed that the rhetorical structure of the reading passage on speaking test results had no impact on test-takers' performance. This result was not consistent with previous research (Carrell, 1984, 1985; Kobayashi, 2002; Kintsch and Yarbrough, 1982; Luoma, 2004; Mann and Thompson; 1987). Some participants answered that the rhetorical structure of the reading passages did not affect their understanding, but that they spent more time focusing on listening and processing what was being heard. Some students did not even notice that the passages

had different structures. One of the reasons that rhetorical structures of reading passages did not affect test-takers' performance might be because the reading passages used in the integrated tests were not long enough to affect test-takers' understanding in a measurable way. Also, it can be concluded that whether the topic is placed at the beginning or not, does not impact test results.

As far as test-wiseness is concerned, this study found that there was statistically significant in test wiseness levels. The more experienced test-takers did outperform the novice test-takers in this study. This result is consistent with previous studies (Allan, 1992; Ford, 1973; Huston 2005; Millnam et al., 1965). The results also showed that the speaking test scores for both groups of experienced and novice participants were not affected by the differences in planning time length during the testing. This is to say that providing more planning time for both novice and experienced test-takers showed no benefit towards improving performance in integrated speaking tests.

On the other hand, if research shows that some test-takers have an advantage over others, merely from having greater previous experience in test-taking, then the actual test results would not be an accurate measurement of language ability as was hoped for, and the testing method itself may not be said to be fair to all test-takers. If simply giving additional planning time to prepare an answer is not the factor which minimizes the advantage that high test-wiseness test-takers have in a test situation, it can certainly be said that more research in this area is necessary.

Some limitations of this study should also be mentioned. Firstly, the main issue is that the sample size was not large enough so as to represent all test-takers, and because of that, some results might be not statistically significant. Adding low-proficiency level participants may lead to different research findings. Giving more planning time might be significantly more advantageous to lower proficiency level participants. Additionally, the participants' general English proficiency in this study might not vary greatly enough or cover a broad enough range of ability. With this small sample size, parametric statistics were chosen over nonparametric statistics to aid in analyzing the results.

For a future study, it might be interesting to analyze the quality or effectiveness of the test takers' note-taking, and possibly have their answers transcribed and studied, to see if there are noticeable differences between the three different time length groups, and to see what impact that note-taking has on performance. More research is needed to continue to find the most valid and reliable testing methods possible, which will most accurately reflect students' English-speaking skills.

## References

- Afflerbach, P. P. 1990. The Influence of Prior Knowledge on Expert Readers' Main Idea Construction Strategies. *Reading Research Quarterly* 1, 41-46.
- Allan, A. 1992. Development and Validation of a Scale to Measure Test-wiseness in EFL/ESL Reading Test Takers. *Language Testing* 9.2, 101-119.
- Carrell, P. L. 1984. The Effects of Rhetorical Organization on ESL Readers. *TESOL Quarterly* 18.3, 441-469.
- Carrell, P. L. 1985. Facilitating ESL Reading by Teaching Text Structure. *TESOL Quarterly* 19.4, 727-752.
- Educational Testing Service. 2016. *Official TOEFL iBT Tests Volume 2 with DVD-Rom*. New York: McGraw-Hill.
- Educational Testing Service. 2017. *Official Guide to the New TOEFL Test with DVD-Rom*. New York: McGraw-Hill.
- Elder, C. and N. Iwashita. 2005. Planning for Test Performance: Does It Make a Difference? in R. Ellis (Ed.), *Planning and Task Performance in a Second Language*, 219-238. Amsterdam: John Benjamins.

- English, H. B. and A. English. 1970. *Comprehensive Dictionary of Psychological and Psychoanalytical Terms*. New York: David McKay Company, Inc.
- Ford, V. A. 1973. *Everything You Wanted to Know about Test-Wiseness*. Princeton, New Jersey: Educational Testing Service.
- Foster, P. 1996. Doing the Task Better: How Planning Time Influences Students' Performance. In J. Willis and D. Willis (Eds.), *Challenge and Change in Language Teaching*, 126-135. London: Heinemann.
- Foster, P. and P. Skehan. 1996. The Influence of Planning on Performance in Task-based Learning. *Studies in Second Language Acquisition* 18.3, 299-324.
- Frost, K., C. Elder, and G. Wigglesworth. 2011. Investigating the Validity of an Integrated Listening-speaking Task: A Discourse-based Analysis of Test Takers' Oral Performances. *Language Testing* 29.3, 345-369.
- Gibb, B. G. 1964. *Test-wiseness as a Secondary Cue Response*. Unpublished Doctoral Dissertation. Stanford University, California, U.S.A.
- Gilabert, R. 2007. The Simultaneous Manipulation of Task Complexity Along Planning Time and (+/- Here-and-Now): Effects on L2 Oral Production. In M. Garcia-Mayo (Ed.), *Investigating Tasks in Formal Language Learning*, 44-68. Bristol: Multilingual Matters.
- Houston, S. E. 2005. *Test-wiseness Training: An Investigation of the Impact of Test-wiseness in an Employment Setting*. Doctoral dissertation. University of Akron, Ohio, U.S.A.
- Hyun, I.-S. 2017. *Effects of Interlocutor Proficiency on Test-takers' Test Scores, Interaction Patterns, and Anxiety Levels in a Paired Speaking Test..* Doctoral dissertation. Ewha Womans University, Seoul, Korea.
- Iwashita, N., T. McNamara, and C. Elder. 2001. Can We Predict Task Difficulty in an Oral Proficiency Test? Exploring the Potential of Information Processing Approach to Task Design. *Language Learning* 51, 401-436.
- Kintsch, W. and J. C. Yarbrough. 1982. Role of Rhetorical Structure in Text Comprehension. *Journal of Educational Psychology* 74.6, 828.
- Kobayashi, M. 2002. Method Effects on Reading Comprehension Test Performance: Text Organization and Response Format. *Language Testing* 19.2, 193-220.
- Lee, S. 2018. *Effective Planning in Real-Time Speaking Test Tasks*. Doctoral dissertation. Michigan State University, Michigan, U.S.A.
- Li, L., J. Chen, and L. Sun. 2015. The Effects of Different Lengths of Pretask Planning Time on L2 Learners' Oral Test Performance. *TESOL Quarterly* 49.1, 38-66.
- Luoma, S. 2004. *Assessing Speaking*. Ernst Klett Sprachen.
- Mann, W. C. and S. A. Thompson. 1987. *Rhetorical Structure Theory: A Theory of Text Organization*, 87-190. University of Southern California, California, U.S.A.
- Mehrang, F. and M. Rahimpour. 2010. The Impact of Task Structure and Planning Conditions on Oral Performance of EFL Learners. *Procedia Social and Behavioral Sciences* 2.2, 3678-3686.
- Mehnert, U. 1998. The Effects of Different Lengths of Time for Planning on Second Language performance. *Studies in Second Language Acquisition* 20, 83108.
- Millman, J., H. I. Bishop, and R. Ebel. 1965. An Analysis of Test Wiseness. *Educational and Psychological Measurement* 25.1, 707-726.
- Mochizuki, N. and L. Ortega. 2008. Balancing Communication and Grammar in Beginning-Level Foreign Language Classrooms; A Study of Guided Planning and Relativization. *Language Teaching Research* 12.1, 11-37.
- Oakland, T. 1972. Effects of Test-Wiseness Materials on Standardized Test Performance of Pre-School Disadvantaged

- Children, *Journal of School Psychology* 10, 355-360.
- Ortega, L. 1999. Planning and Focus on Form in L2 Oral Performance. *Studies in Second Language Acquisition* 21, 109-148.
- Richards, J. C. and R. W. Schmidt. 2013. *Longman Dictionary of Language Teaching and Applied Linguistics*. Routledge.
- Rutherford, K. 2001. *An Investigation of the Effects of Planning on Oral Production in a Second Language*. Unpublished MA thesis. University of Auckland, Auckland, New Zealand.
- Skehan, P. and P. Foster. 1999. The Influence of Task Structure and Processing Conditions on Narrative Retellings. *Language Learning* 49.1, 93-120.
- Tavakoli, P. and P. Skehan. 2005. Strategic Planning, Task Structure, and Performance Testing. In R. Ellis (Ed.), *Planning and Task Performance in a Second Language*, 239-273. Amsterdam: John Benjamins.
- Van Dijk, T. A. and W. Kintsch. 1983. *Strategies for Discourse Comprehension*. New York.
- Wigglesworth, G. 1997. An Investigation of Planning Time and Proficiency Level on Oral Test Discourse. *Language Testing* 14, 21-44.
- Wigglesworth, G. 2000. Issues in the Development of Oral Tasks for Competency-based Assessments of Second Language Performance. *Studies in Immigrant English Language Assessment* 1, 81-124.
- Wigglesworth, G. 2001. Influences on Performance in Task-based Oral Assessments. In M. Bygate, P. Skehan, and M. Swain (Eds.), *Task Based Learning*, 186-209. London: Longman.
- Wigglesworth, G. and C. Elder. 2010. An Investigation of the Effectiveness and Validity of Planning Time in Speaking Test Tasks. *Language Assessment Quarterly* 7.1, 1-24.

## Appendix

### 1. Examples of Two Types of Reading Passages

#### 1.1 Key topic sentence at the beginning of the paragraph

##### Cognitive Dissonance

Individuals sometimes experience a contradiction between their actions and their beliefs—between what they are doing and what they believe—between what they are doing and what they believe they should be doing. These contradictions can cause a kind of mental discomfort known as cognitive dissonance. People experiencing cognitive dissonance often do not want to change the way they are acting, so they resolve the contradictory situation in another way: they change their interpretation of the situation in a way that minimizes the contradiction between what they are doing and they believe they should be doing.

#### 1.2 Key topic sentence at the end of the paragraph

##### Subliminal Perception

Humans are constantly perceiving visual and auditory stimuli. Sometimes our perception of these stimuli occurs consciously; we are aware of a stimulus and know that we are perceiving it. But our perception of a stimulus can also occur without our awareness; an image might appear and disappear before our eyes too quickly for us to notice that we saw it, or a sound might be too faint for us to realize that we heard it. This phenomenon – the perception of a stimulus just below the threshold of conscious awareness – is called subliminal perception. Experiments have shown that subliminally perceived stimuli can influence people’s thoughts and attitudes.

Koo, Jinseon, Doctoral Candidate  
Ewhayedae-gil 52, Seodaemun-gu, Seoul, Korea  
Dept. of English Education, Ewha Womans University  
E-mail: jinseonkoo@ewhain.net