Korean Journal of English Language and Linguistics, Vol 22, September 2022, pp. 938-956
DOI: 10.15738/kjell.22..202210.938

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

Effects of Pre-task and On-line Planning on Complexity, Fluency, and Accuracy in Computer-based English Speaking and Writing Tests*

Mijin Joo (Kangwon National University)

Received: July 30, 2022
Revised: September 01, 2022
Accepted: September 30, 2022

Mijin Joo
Professor, Department of English, Division of Global Human Resources, Kangwon National University
Tel: (033)-570-6654
Email: ing1115@hanmail.net

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5A2A01060567). / This study has been worked with the support of a research grant of Kangwon National University in 2021.

ABSTRACT

Joo, Mijin. 2022. Effects of pre-task and on-line planning on complexity, fluency, and accuracy in computer-based English speaking and writing tests. *Korean Journal of English Language and Linguistics* 22, 938-956.

This study examined the effects of pre-task and on-line planning on discourse and scores in terms of complexity, fluency, and accuracy in CBT speaking and writing tests. Fifty-six Korean university students took both the CBT speaking and writing tests under three different planning time conditions (pre-task, on-line, and no planning time). All test performance was scored by two raters, and their discourse was transcribed and analyzed. The primary findings are as follows. First, while there were no significant differences in discourse measures of the CBT speaking test performance, those of the writing test were affected by different planning conditions. The test-takers produced more fluent and accurate language with planning time than without planning time. Second, the planning time did not influence the scores of both CBT speaking and writing tests. Lastly, in discourse analysis, complexity and fluency are negatively correlated with accuracy under pre-task and on-line planning conditions.

KEYWORDS

pre-task planning, on-line planning, CBT speaking and writing tests, complexity, fluency, accuracy

1. Research Rationale and Purpose

In second language acquisition (SLA) research, planning has been considered as one of the factors that might account for the variability in second language speaking or writing production. It seems generally known that planning time is critical in improving learners' output when performing a task. According to information processing theory, learners have limited ability to process information, so it is difficult to allocate attention to meaning and form of language simultaneously. When the learners are put under pressure in time, they focus on meaning rather than form (Skehan and Foster 2005). In other words, they attach weight to meaning and overlook language forms. This phenomenon is more pronounced in learners with low proficiency (Anderson 1995, Skehan 1996, VanPatten 1990). Therefore, it is believed that providing planning time can improve the content and the quality of the learners' output by supplementing this restriction of limited attention (Skehan 1996).

Until recently, numerous studies have been conducted on planning time in the field of SLA, and a significant number of studies found that planning time improved learners' output in fluency and/or complexity (e.g., Ahmadian et al. 2015, Crookes 1989, Foster and Skehan 1996, Mehnert 1998, Ortega 1999, Wendel 1997). Although the results on accuracy have been unclear and inconsistent, it seems evident that at least planning time has positive effects on learners' language output.

The effects of planning time in an assessment context are also recognized as theoretically necessary, but research on planning in a testing context is relatively scarce. Only a few studied the effects of planning in the assessment situation (e.g., Elder and Iwashita 2005, Tavakoli and Skehan 2005, Wigglesworth 1997), but none of them was in a computer-based test (CBT) context even though the computer has been used more than ever as the primary vehicle for teaching and assessing learners.

The results in a testing situation may markedly differ from those in a classroom setting. The two possible differences can be as follows. First, the most obvious difference is that the test is a high-stakes situation in which learner's ability is assessed, so performance with planning time can be different from those in other situations. (Ellis 2005, Li, Chen and Sun 2015). In the classroom situation, the learners may focus more on completing the task and pay more attention to complexity related to content delivery. On the other hand, the learners would increase attention to accuracy in the testing context because they try not to make a mistake that may affect their scores (Ellis 2005). This would result in decreased attention to complexity and fluency. The other is that the CBT test context can affect performance quality (Skehan 1998). It differs from the classroom or the test context with a live interlocutor. The test is self-administered and completed by the test-takers without human interruptions. For example, there is no interaction between the interlocutor and the test-taker, and one-way communication is carried out. Fluency and complexity would not matter much because no human is waiting for test-takers' responses. It is assumed that they would focus more on accuracy than those in general language testing contexts. In other words, they would try to make simple, slow, but accurate sentences.

In real-life situations, speaking and writing can occur in both planning and unplanning conditions (Skehan 1998). Still, it is expected that learners are more likely to encounter speaking and writing situations with planning time. For example, the learners in virtual academic settings spend 30 seconds to several hours preparing to answer a question, discussing on SNS, oral presentation, doing an assignment, writing a report, etc. In addition, it is more common to have sufficient planning time in various writing situations, including writing an e-mail.

In a speaking case, it may be argued that the planning time allotment violates authenticity (Wigglesworth 1997). Still, even in various speaking situations, the speaker has opportunities to plan what to say in advance. For example, in most worldwide customer service centers, people can chat on SNS with a human agent or a chat robot allowing sufficient plan/think time. Even in a telephone or one-to-one conversation, when communication is not smooth or

challenging, they can ask for some time to think, repeat changes a word, or use communication strategies to gain time to plan.

The ways of communication beyond time and space and in the virtual reality world are expected to be more diverse, allowing more planning time. In this regard, providing planning time does not harm authenticity. If the test-taker performance can be improved with planning time, the provision of planning time should be included in a test because language tests should be designed to maximize the test-taker abilities (Bachman and Palmer 1996).

Despite the abundance of literature on the CBT (e.g., Brown 1993, Kenyon and Malabonga 2001, O'Loughlin 2001, Shohamy 1994), no studies have been found on planning time in CBT speaking and writing tests. Therefore, this study aims to investigate the effects of pre-task and on-line planning on the test-taker discourse and scores in the CBT English speaking and writing tests. Specific research questions are as follows.

(1) Do pre-task and on-line planning impact test-taker 'discourse' in the CBT English speaking and writing tests?
(2) Do pre-task and on-line planning make a difference to the 'scores' in the CBT speaking and writing tests?
(3) Is there any trade-off relationship between CFA in the CBT speaking and writing tests under different planning conditions?

2. Literature Review

2.1 Pre-task and On-line Planning

One of the task qualities affecting test-taker performance is planning time (Weigle 2002). According to Bachman (1990), one aspect of the task is the 'expected response' related to speededness. Speededness refers to the degree to which the test-taker plans and executes the response. In this respect, optimal time allotment is essential so that the test-taker can fully demonstrate his or her abilities. Pre-task and on-line planning were defined by Ellis (2005). These two types of planning are distinguished by when the planning is carried out. Pre-task planning occurs 'before' the task is performed, while on-line planning is 'during' the task.

Pre-task planning includes rehearsals and strategic plans. Rehearsal is to perform the same task once before actually performing the task. On the other hand, strategic planning is to plan content, words, and sentences related to the task before performing the task. It was believed that planning time before performing the task helps learners overcome limited language processing capabilities and pay attention to both meaning and form, thereby maximizing language ability and improving task performance. Although it was expected that planning time would allow focusing on the form and message delivery, it was not easy to find a significant improvement in accuracy in previous studies. Thus, it is not certain that the pre-task planning helps facilitate not only fluency and complexity but also accuracy.

On-line planning, which is also called as within-task planning, could be pressured or unpressured. Pressured planning is used when the task is completed within a limited amount of time, referred to as unplanning language use by Ochs (1979). On the other hand, unpressured planning takes place when the speaker or writer carries out the task with sufficient time, referred to as planning language use (Ochs 1979). When planning is pressured, learners (especially with low proficiency) are likely to have difficulty accessing and encoding their language knowledge (Ellis 2005). On the other hand, under unpressured on-line planning conditions, test-takers are likely to have sufficient time to (re)conceptualize, (re)formulate, and monitor their internal speech (to use Levelt's terminology) prior to articulation (Ellis 2005) resulting in the improvement of their speaking or writing performance.

Only a few studies have, however, dealt with on-line planning. They found that on-line planning improved accuracy (Ellis and Yuan 2004, Ghavamnia, Tavakoli and Esteki 2012, Yuan and Ellis 2003) and/or complexity (Kim 2017, Yuan and Ellis 2003). The studies suggest that on-line planning helps learners focus on the formulation process and facilitate access to grammar, leading to higher accuracy.

It appears that pre-task planning does not help much in shaping grammatical morphology. Even if the learners plan to form in detail under pre-task planning conditions, it is unlikely that the language form previously planned is fulfilled during a performance. The formulation may be attempted on the spot. In other words, the learners tend to focus on what to speak or write rather than how to speak or write under pre-task planning conditions. On the other hand, on-line planning aids the learner in search for long-term memories of grammatically correct expressions leading to overall accuracy improvement (Ellis and Yuan 2005, Yuan and Ellis 2003). It is more expected to contribute to improving accuracy than pre-task planning. As a result, it seems that on-line planning can help pay attention to formulation while pre-task planning gives more attention to the conceptualization stage.

2.2. Effects of Planning on Speaking and Writing Production in L2 Testing Context

Although the importance of planning and time allocation has been emphasized in plenty of studies (e.g., Iwashita et al. 2001, Mehnert 1998, Wigglesworth 2001), only a few (standardized) English speaking or writing tests allow test-takers to have a certain amount of planning time (over 60 seconds) before or during performing tasks. For example, the IELTS speaking test provides 60 seconds of preparation before 2 minutes of response time. TEPS CBT speaking test allows test-takers to have 60, 120, and 60 seconds of preparation time and 60, 90, and 90 seconds of response time respectively. OPIc iBT speaking test doesn't supply any particular preparation time but provides unlimited response time. Most of the other (CBT) speaking tests give less than 60 seconds of preparation time depending on the types of tasks: TOEIC CBT speaking test (3, 30, and 45 seconds), TOEFL CBT speaking test (15-30 seconds), and G-TELP iBT speaking test (30 seconds). On the other hand, most CBT writing tasks such as G-TELP, TEPS, TOEIC, TOEFL, and IELTS do not give any separate time for planning.

Learners often insist that they could not perform better than they did in the classroom due to the pressure and nervousness on the test. Thus, in a testing context, they can react differently to the task according to the types and amount of planning time from those in the classroom situation. However, only a few studies have been conducted on the effects of planning in speaking test contexts (Elder and Iwashita 2005, Iwashita et al. 2001, Tavakoli and Skehan 2005, Wigglesworth 1997, Wigglesworth and Elder 2010). For example, Wigglesworth (1997) investigated the effects of planning time on the oral output across high proficiency and low proficiency groups and compared them under the planned and unplanned discourse on several tasks that differed in their difficulty levels. No differences in the analytic scores assigned by raters emerged between the planned and unplanned discourse. However, high proficiency test-takers showed higher complexity, better fluency, and higher accuracy under the planning condition on cognitively demanding tasks. These marked effects did not occur in the low proficiency test-takers although they also showed some evidence of increased fluency and accuracy. Therefore, it was suggested to provide one minute of pre-task planning time for the more challenging tasks.

Iwashita, et al. (2001) investigated the relationship between task characteristics and task performance under semi-direct oral test conditions. As a result, it was found that there was no significant impact on either the quality of oral test discourse or test scores under two planning conditions (with and without planning time). It was assumed that more planning time on more complex tasks would have increased fluency and accuracy. The effect of planning may not have been manifested in fluency and accuracy because the test-takers focused more on message delivery.

Elder and Iwashita (2005) investigated the effects of planning time on oral production under a testing context.

The test-takers took two tasks. One is a story told after 3.75 minutes of planning time, and the other with only 0.75 minutes. The test-taker oral productions were scored using analytical rating scales for complexity, fluency, and accuracy (CFA) and transcribed and analyzed with discourse measures of CFA. As a result, it was found that there were no significant effects on both scores and discourse measures of CFA.

Tavakoli and Skehan (2005) examined the effects of pre-task planning and task structure on speaking performance according to different proficiency levels in a testing context. They found that different aspects of performance were affected differently by task structure and pre-task strategic planning. Strategic planning significantly influenced test-taker performance by improving complexity, fluency, and (especially) accuracy. It was claimed that learners tended to focus on accuracy more in an assessment situation.

Unlike oral test performance research, studies on planning effects in a writing test context could not be found. In addition, there has been little consideration of the interaction between types of planning and test context on test-taker performance. Test conditions of task implementation can have a significant influence on performance. Thus, this study is expected to demonstrate what role planning plays in speaking and writing test conditions.

3. Methodology

3.1 Participants and Procedure

The participants were 56 university students taking a liberal arts English class at a Korean university. The participants were informed of the research purpose, procedures, risks, benefits, and ways to withdraw participation and were required to sign the written consent form. They were 46.4% male and 53.6% female and their age ranged from 19 to 24. They had been learning English for more than ten years. They all had taken the TOEIC test, and based on the results they could generally be considered as low proficiency overall (Mean score = 437).

The participants were randomly assigned to three groups. Counterbalancing was performed by presenting planning time conditions and tasks to each group in a different order (Table 1). The counterbalance was to minimize the impact of the order of the speaking and writing tests for each task and task plan type on performance. Each group was required to carry out three tasks under each different planning condition. The design of the study meant that the same group performing the oral and written tasks was compared under three planning time conditions.

TABLE 1. Counterbalanced Sequence of CBT English Speaking/Writing Tasks

Group	N	Task	Planning	Task	Planning	Task	Planning
G1	18	task 1	pre-task	task 3	on-line	task 2	no
G2	19	task 2	on-line	task 1	no	task 3	pre-task
G3	19	task 3	no	task 2	pre-task	task 1	on-line

3.2 Tasks and Planning Conditions

In consideration of participants' English proficiency, the tasks were constructed as simple and familiar as possible not to cause any cognitive burden and not to affect discourse and scores due to variables other than English ability and planning time. They were required to freely express their opinions or thoughts on familiar topics such as robots, pets, jobs, or love.

Sample of writing task

Instruction (given in Korean): Based on the following topic with questions, you should write at least more than 7 sentences in English. You will be given 10 minutes to write English in 30 seconds. Make sure to use up all 10 minutes.

Topic: Robots

Questions: Robots can do many different jobs. What jobs do you think robots can or cannot do? or what are some of the advantages and disadvantages of having robots work in factories and other places such as hospitals and homes for people?

The instructions indicated that the test-takers should be given one to two minutes to prepare. Given previous research which has indicated that as little as one minute can affect performance on some measures (see Mehnert 1998, Wigglesworth 1997), this study set out to investigate if there were any differences according to three different planning time conditions: pre-task, on-line, and no planning.

Under no planning time condition, the test-takers had to complete the task immediately after reading the instructions and the topic with questions (30 seconds) and within a limited time (2 minutes). Therefore, the test takers had little time to plan the task. In the pre-task planning condition, three minutes of planning time and two minutes of performance time were given before performing the task. In the on-line planning condition, 30 seconds of planning time for reading the instructions and the topic and five minutes of performance time were provided (see Tables 2-3).

TABLE 2. CBT Speaking Test Planning Conditions

Condition	Planning time	Performance time
No planning	30 secs	2 mins
Pre-task planning	3 mins	2 mins
On-line-planning	30 secs	5 mins

TABLE 3. CBT Writing Test Planning Conditions

Condition	Planning time	Performance time
No planning	30 secs	10 mins
Pre-task planning	5 mins	10 mins
On-line-planning	30 secs	15 mins

3.3 CBT Speaking and Writing Tests

For the current study, various CBT language tests were searched and chose OWL test because it was manageable and flexible enough to meet the research needs. The OWL Test is a web-based test allowing users to create, administer, and manage their tests. It can incorporate multimedia allowing to use of text, sound, pictures, graphics, video, or a combination to create items that assess all four language skills (speaking, reading, listening, and writing). The test is delivered through the Microsoft Azure Global Network, including Transparent Data Encryption at rest and secure data transmission using Hypertext Transfer Protocol Secure (HTTPS) (see https://owlts.com/).

To take the test, test-takers should log in to the OWL testing software with their usernames and passwords. Then they can take the assigned speaking and writing tests. While taking the tests, test-takers hear test directions with

accompanying text and questions. It is also designed to set up preparation and test time for each task. All the oral and written responses are automatically recorded on the test software. Once the test-takers are ready to start the test, the first question appears.

In this study, the test-takers were given instructions in Korean and tasks in English. The test-takers were allowed to write on paper during the planning time. All test-taker verbal or written responses were automatically recorded right after planning time. The test-taker responses were rated using rating scales with descriptions, and oral responses were transcribed as soon as the test-takers completed each task.

3.4 Discourse Analysis

For discourse measurement of complexity, accuracy, and fluency, the number of words, clauses, and t-units were calculated by referring to Bygate, Skehan, and Swain (2001: 34) and Skehan and Foster (1999: 107). Each method of discourse measurement is as follows. For the reliability of the data, approximately 10% of the data was coded again by another independent transcriber and resulted in 75% level of agreement (Cohen's $k = 0.333$).

Fluency
: The total number of words is divided by the number of T-units. Thus, the higher the fluency, the higher the number is measured.

Accuracy
: The error-free clauses are divided into the total number of sentences and measured as a percentage. All syntactic, morphological, and lexical errors are taken into account.

Complexity
: The total number of clauses is divided by the total number of T-units. That is, the number of clauses per each T-unit is indicated. Here, T-unit includes the main clause and all kinds of subordinate clauses nested or linked to the main clause. Therefore, the more complex sentences are used, the higher the number is.

3.5 Scoring

Test-taker performance was rated using analytical rating scales for CFA. The speaking rating scales were adopted from Elder and Iwashita (2005), and on the basis of the speaking rating scales, the writing rating scales were developed and used for the study.

Two raters were employed for scoring. Both raters had experience in scoring speaking tests as well as teaching English at Korean universities. Before actual scoring the speaking and writing tests, the raters were required to participate in a one-day intensive rater training. Inter-rater reliability coefficients were obtained on all scores of CFA for analysis by two raters working independently. Inter-rater reliability was above 83.54% on all scores (Cohen's $k = 0.451$).

4. Results

4.1 Test-taker Discourse under Pre-task and On-line Planning Conditions

The first research question addressed the issue of whether planning time makes a difference to the discourse of

test-taker output in the CBT speaking and writing tests. First, the means of words, sentences, t-units, and error-free clauses in the speaking test are presented in Table 4. It reveals significant differences according to the different planning time. The mean of each item increased in the order of no, pre-task, and on-line planning (Cohen's *f* effect size word = 2.027; sentence = 2.117; *t*-unit = 1.037; error-free clause = 1.650). It can be seen that the mean difference between the on-line and no planning conditions was the largest (Table 5).

TABLE 4. Descriptive and ANOVA results on the CBT Speaking Test

Measures	Planning	N	Mean	S.D.	ANOVA		Effect size
					F	*p*	Cohen's *f*
Word	no	48	72.52	41.795	17.43**	.000	2.027
	pre-task	44	84.57	42.284			
	on-line	43	143.79	88.609			
Sentence	no	48	9.85	5.161	18.93**	.000	2.117
	pre-task	44	12.48	5.258			
	on-line	43	19.28	10.833			
T-unit	no	48	1.52	1.167	5.30**	.006	1.037
	pre-task	44	1.93	1.704			
	on-line	43	2.63	1.964			
Error-free clause	no	48	5.92	4.094	11.90**	.000	1.650
	pre-task	44	8.00	4.549			
	on-line	43	11.33	6.968			

*$p < .1$, **$p < .05$, ***$p < .001$

TABLE 5. Post-hoc Pairwise Comparisons (the CBT Speaking Test)

Measures	Planning	Mean difference	S.E.	*p*
Word	no planning vs. pre-task	-12.05	12.70	1.000
	pre-task vs. on-line	-59.22***	13.05	.000
	on-line vs. no planning	71.27***	12.78	.000
Sentence	no planning vs. pre-task	-2.62	1.56	.285
	pre-task vs. on-line	-6.80***	1.60	.000
	on-line vs. no planning	9.43***	1.57	.000
T-unit	no planning vs. pre-task	-.41	.34	.69
	pre-task vs. on-line	-.70	.35	.15
	on-line vs. no planning	1.11**	.34	.005
Error-free clause	no planning vs. pre-task	-2.08	1.11	.19
	pre-task vs. on-line	-3.33**	1.14	.012
	on-line vs. no planning	5.41***	1.12	.000

*$p < .1$, **$p < .05$, ***$p < .001$

As shown in Table 6, which shows the results of the CBT writing test, the items present an increase in a similar pattern to those in Table 4 revealing the highest means in the on-line planning condition. However, the difference was not big enough to reach statistical significance, unlike those of the CBT speaking test.

TABLE 6. Descriptive and ANOVA results on the CBT Writing Test

Measures	Planning	N	Mean	S.D.	ANOVA		Cohen's *f*
					F	*p*	
Word	no	51	94.90	55.954	.587	.557	0
	pre-task	50	96.58	55.474			
	on-line	51	105.86	53.504			
Sentence	no	51	13.35	6.731	.232	.232	0
	pre-task	50	13.46	6.831			
	on-line	51	14.16	5.787			
T-unit	no	51	2.67	6.731	.204	.816	0
	pre-task	50	2.60	6.831			
	on-line	51	2.84	5.787			
Error-free clause	no	51	11.31	6.731	.360	.699	0
	pre-task	50	10.48	6.831			
	on-line	51	11.49	5.787			

*$p < .1$, **$p < .05$, ***$p < .001$

The CFA discourse measures in the speaking test are presented in Table 7. The means of complexity and accuracy revealed a gradual increasing pattern when they had planning time in the order of no, pre-task, and on-line planning conditions. However, the mean differences were not statistically significant even though those of word, sentence, t-unit, and error free clause were significant in Table 4. The results reveal that the different planning time conditions had no effect on the speaking performance.

TABLE 7. Descriptive and ANOVA results on CFA of the CBT Speaking Test

Measures	Planning	N	Mean	S.D.	ANOVA		Effect size Cohen's *f*
					F	*p*	
Complexity	no	38	6.81	3.57	1.328	.269	0.286
	pre-task	37	7.830	5.11			
	on-line	38	8.71	6.24			
Fluency	no	38	58.29	21.42	.466	.629	0
	pre-task	37	61.74	18.81			
	on-line	38	58.08	19.43			
Accuracy	no	38	49.80	26.99	1.272	.284	0.261
	pre-task	37	52.21	35.93			
	on-line	38	61.72	39.20			

*$p < .1$, **$p < .05$, ***$p < .001$

In contrast to speaking test performance, the discourse of the writing test was affected by the different planning conditions (see Table 8). Fluency and accuracy differed depending on the planning time. Table 8 presented that the test-takers performed better when they had pre-task and on-line planning time in aspects of fluency and accuracy (Fluency $F = 4.521$, $p = .013$, Cohen's $f = 1.898$; Accuracy $F = 15.403$, $p = .000$, Cohen's $f = 0.938$).

In Table 9, it can be seen that the test-takers improved accuracy when given planning time (no planning vs. pre-task planning M.D. = 26.06, $p = .000$; no planning vs. on-line planning M.D. = 29.076, $p = .000$). On the other hand, there was no significant improvement in accuracy between performances with pre-task and on-line planning conditions. When it comes to fluency, the test-takers improved fluency only under on-line planning condition (no planning vs. on-line planning M.D. = 20.073, $p = .013$) (Table 9).

There was also a slightly higher means of complexity under on-line planning condition than those of no planning and pre-task planning, but the differences were not statistically significant. It can be seen that the test-taker performance was positively influenced in the order of no, pre-task, and on-line planning conditions. In other words, the on-line planning condition was especially more effective in improving writing performance than pre-task or no planning condition.

TABLE 8. Descriptive and ANOVA results on CFA of the CBT Writing Test

Measures	Planning	N	Mean	S.D.	ANOVA		Effect size
					F	*p*	Cohen's *f*
Complexity	no	47	5.925	3.499	.147	.863	0
	pre-task	45	6.1698	3.049			
	on-line	48	6.303	3.732			
Fluency	no	47	51.014	40.960	4.521**	.013	1.898
	pre-task	45	77.077	21.608			
	on-line	48	80.090	19.528			
Accuracy	no	47	28.845	34.692	15.403**	.000	0.938
	pre-task	45	43.552	26.675			
	on-line	48	48.919	38.566			

*$p < .1$, **$p < .05$, ***$p < .001$

TABLE 9. Post-hoc Pairwise Comparisons

Measures	Planning	Mean difference	S.E.	*p*
Complexity	no planning vs. pre-task	.500	.697	1.000
	pre-task vs. on-line	-.133	.697	1.000
	on-line vs. no planning	-.366	.685	1.000
Fluency	no planning vs. pre-task	-14.71	7.02	.114
	pre-task vs. on-line	-5.37	7.02	1.000
	on-line vs. no planning	20.07**	6.90	.013
Accuracy	no planning vs. pre-task	-26.06***	5.78	.000
	pre-task vs. on-line	-3.01	5.78	1.000
	on-line vs. no planning	29.08***	5.75	.000

*$p < .1$, **$p < .05$, ***$p < .001$

To summarize, the means of words, sentences, t-units, and error-free clauses in the CBT speaking test increased in the order of no, pre-task, and on-line planning conditions. However, there were no significant differences in CFA discourse measures of the CBT speaking test according to the different planning conditions. In contrast to the discourse of the speaking performance, those of the writing performance were affected by the different planning conditions. The test-takers showed better performance when they had pre-task and on-line planning time in aspects of fluency and accuracy. The on-line planning especially made a difference in writing performance improving fluency and accuracy.

4.2 Test Scores under Pre-task and On-line Planning Conditions

The test scores were analyzed to examine the impact of provision of the different planning time. Table 10 reveals that there were no significant effects of planning time on speaking test scores. It seems that the provision of planning time didn't make a difference to the scores achieved by the test-takers.

Table 10. Descriptive and ANOVA Results on the Scores in the CBT Speaking Test

Dependent variable	Condition	N	Score Mean	S.D.	ANOVA		Effect size
					F	p	Cohen's *f*
Complexity	no	43	2.291	.638	.796	.454	0
	pre-task	44	2.398	.752			
	on-line	44	2.489	.796			
Fluency	no	43	2.593	.692	2.016	.137	0.503
	pre-task	44	2.500	.755			
	on-line	44	2.364	.838			
Accuracy	no plan	43	2.114	.908	.992	.374	0
	pre-task	44	2.455	.746			
	on-line	44	2.384	.858			

*$p < .1$, **$p < .05$, ***$p < .001$

On the other hand, in the CBT writing test, it can be seen that the test-takers had slightly better performance in CFA when given on-line planning time than any other planning time (Table 11). However, there were no statistically significant differences in the test scores with the three different planning time. This was a contrast to the results of discourse measures of the writing test in Table 8. Conclusively, the planning conditions did not influence the scores in the CBT writing test.

TABLE 11. Descriptive and ANOVA Results on the Scores in the CBT Writing Test

Items	Planning	N	Score Mean	S.D.	ANOVA		Effect size
					F	p	Cohen's *f*
Complexity	no	51	3.088	.563	.514	.599	0
	pre-task	50	3.080	.609			
	on-line	51	3.186	.591			
Fluency	no	51	2.804	.701	1.095	.337	0.154
	pre-task	50	2.820	.668			
	on-line	51	2.980	.624			
Accuracy	no	51	3.240	.803	.299	.742	0
	pre-task	50	3.284	.559			
	on-line	51	3.343	.636			

*p < .1, **p < .05, ***p < .001

While the writing scores were related to the CFA discourse measures, the speaking scores were not connected with those as presented in Tables 12-13. This may mean that the speaking performance was scored with an emphasis on accuracy. Further investigation seems to be in need on this issue.

TABLE 12. Results of Correlations between Discourse and Scores in the CBT Speaking Test

Discourse	Complexity	Accuracy	Fluency
Complexity	-.133(.160)	-.104(.271)	.006(.949)
Fluency	-.055(.562)	-.062(.514)	.117(.217)
Accuracy	.336**(.000)	.338**(.000)	.270**(.002)

*p < .1, **p < .05, ***p < .001

Table 13. Results of Correlations between Discourse and Scores in the CBT Writing Test

Discourse	Complexity	Accuracy	Fluency
Complexity	-.220**(.009)	-.032(.708)	.044(.605)
Fluency	-.008(.927)	.127(.134)	.305**(.000)
Accuracy	.189*(.020)	.231**(.004)	.270**(.001)

*p < .1, **p < .05, ***p < .001

In brief, the provision of pre-task and on-line planning time did not make any significant differences to the scores in both CBT speaking and writing tests. This indicates no effects of planning time on the scores. Finally, the CFA discourse measures had associations with the writing scores, but not with the speaking scores.

4.3. Trade-off Relationship between CFA

The third question was answered by conducting a correlation analysis derived from discourse measures and test scores. Complexity and fluency of the speaking test performance were closely related as presented in Table 14. Complexity had significant correlations with fluency under all three planning conditions. However, complexity and fluency had a negative correlation with accuracy under the pre-task planning condition. This indicates that the more complexity and fluency the test-takers had, the less accurate they were in their speaking output with pre-task planning time.

TABLE 14. Results of Correlation Analysis between CFA Speaking Discourse Measures

Planning		Complexity	Fluency	Accuracy
No	Complexity	1	.939**(.000)	.023(.891)
	Fluency	.939**(.000)	1	.007(.969)
	Accuracy	.023(.891)	.007(.969)	1
Pre-task	Complexity	1	.951**(.000)	-.361*(.028)
	Fluency	.951**(.000)	1	-.416*(.011)
	Accuracy	-.361(.028)	-.416*(.011)	1
On-line	Complexity	1	.966**(.000)	-.101(.545)
	Fluency	.966**(.000)	1	-.111(.508)
	Accuracy	-.101(.545)	-.111(.508)	1

*p < .1, **p < .05, ***p < .001

Like the speaking test, there was a significant correlation between complexity and fluency of the CBT writing test performance, as revealed in Table 15. This association became more robust in the order of no, pre-task, and on-line planning time. That is, complexity had the most substantial connection with fluency under on-line planning condition. Also, it is noteworthy that there tended to have negative correlations between complexity/fluency and accuracy in both pre-task and on-line planning even though the association was not strong enough to be significant.

TABLE 15. Results of Correlation Analysis between CFA Writing Discourse Measures

Planning		Complexity	Fluency	Accuracy
No	Complexity	1	.763**(.000)	.181(.223)
	Fluency	.763**(.000)	1	.604**(.000)
	Accuracy	.181(.223)	.604**(.000)	1
Pre-task	Complexity	1	.825**(.000)	-.099(.517)
	Fluency	.825**(.000)	1	-.278(.064)
	Accuracy	-.099(.517)	-.278(.064)	1
On-line	Complexity	1	.902**(.000)	-.139(.348)
	Fluency	.902**(.000)	1	-.144(.328)
	Accuracy	-.139(.348)	-.144(.328)	1

*p < .1, **p < .05, ***p < .001

Unlike discourse measures, CFA scores in both speaking and writing tests were associated with one another (Tables 16-17). In other words, the higher the complexity score was, the higher the accuracy and fluency scores were or vice versa.

TABLE 16. Results of Correlation Analysis between CFA Speaking Test Scores

Planning		Complexity	Fluency	Accuracy
No	Complexity	1	.661**(.000)	.638**(.000)
	Fluency	.661**(.000)	1	.710**(.000)
	Accuracy	.638**(.000)	.710**(.000)	1
Pre-task	Complexity	1	.707**(.000)	.543**(.000)
	Fluency	.707**(.000)	1	.527**(.000)
	Accuracy	.543**(.000)	.527**(.000)	1
On-line	Complexity	1	.734**(.000)	.564**(.000)
	Fluency	.734**(.000)	1	.678**(.000)
	Accuracy	.564**(.000)	.678**(.000)	1

*p < .1, **p < .05, ***p < .001

TABLE 17. Results of Correlation Analysis between CFA Writing Test Scores

Planning		Complexity	Fluency	Accuracy
No	Complexity	1	.552**(.000)	.491**(.000)
	Fluency	.552**(.000)	1	.477**(.000)
	Accuracy	.491**(.000)	.477**(.000)	1
Pre-task	Complexity	1	.563**(.000)	.252(.078)
	Fluency	.563**(.000)	1	.595**(.000)
	Accuracy	.252(.078)	.595**(.000)	1
On-line	Complexity	1	687**(.000)	.385**(.005)
	Fluency	.687**(.000)	1	.408**(.003)
	Accuracy	.385**(.005)	.408**(.003)	1

*p < .1, **p < .05, ***p < .001

To sum up, there was no trade-off relationship between the CFA scores in both speaking and writing tests, whereas it was likely that a trade-off relationship was formed between complexity/fluency and accuracy in discourse. Complexity and fluency tended to have negative correlations with accuracy under pre-task and on-line planning conditions even though the associations were not significant except for those of writing test performance under the pre-task planning condition.

5. Discussion

The results of this study attempted to demonstrate the assumption that providing planning time in the context of the CBT language test makes a difference in the quality of test-taker performances. Some important findings are presented in light of research questions as follows. First, there were no significant discourse differences between CFA in the CBT speaking test. There are two plausible explanations for the result. First, even if planning time was sufficiently provided it might still have been difficult for the low proficiency test-takers to perform better because they should formulate rapidly with real-time processing (Ellis and Yuan 2005). There would have been little time

to monitor and thus no marked improvement in the speaking discourse.

Another possible reason may be that the test-takers might have been allowed too much time (2 minutes) for the task completion. The test-takers might have been able to engage in sufficient amount of on-line planning time even under no planning condition (Elder and Iwashita 2005). Thus, they may not have shown any differences in performance with and without planning time because they could not experience a high level of communication pressure.

Lastly, unfamiliarity with speaking or writing under planning conditions might also have affected the test-taker performance. Allowing three or five minutes for planning is rare, especially in speaking tests. There is a possibility that the test-takers did not know how to deal with the planning time and failed to improve the quality of their language performance. As proposed by Elder and Iwashita (2005), training the learners for effective use of planning time would contribute to making their speech more complex, accurate, and fluent.

Secondly, in contrast to the speaking test, those of the writing test were affected by the different planning conditions. The test-takers produced more fluent and accurate language with planning time than with no planning time. In other words, the pre-task and on-line planning led to the test-takers producing more fluent and accurate sentences. The on-line planning was especially more effective in improving fluency and accuracy. The results were similar to those seen in low proficiency test-takers in Wigglesworth (1997) study although her research was on a semi-oral test. Wigglesworth also presented that the low proficiency test-takers showed evidence of improvements in fluency and accuracy, except complexity.

In terms of fluency, the test-takers in this study may have had opportunities to add more words and make more sentences by monitoring before and/or after articulation. The explanation for the increased accuracy may lie in that the planning time allowed the test-takers to focus more on linguistic knowledge and to monitor more through controlled processing (Ellis and Yuan 2005). In particular, the on-line planning may have facilitated the test-takers to formulate and monitor with their explicit L2 knowledge, increasing fluency and accuracy.

Planning time in this study, however, did not help the test-takers make their formulations more complex. This may be because complexity was more related to the message convey. The tasks used in this study were simple monologic discussions about familiar topics. The tasks may not have inspired the test-takers to elicit complex performance because they could simply achieve the goal of the task by conveying their personal opinions or feelings about the topics freely. There would not have been great difficulty in delivering messages.

Thirdly, the planning time did not affect the test scores. It is, however, worth considering that the scores were awarded by the raters with subjective judgments. The judgments based on impression aided only by a rating scale may have made the scores somewhat less objective and accurate (Elder and Iwashita 2005). The possibility cannot be ruled out that the judgments based on impression, in part, attributed to the result of the scores. The evidence may be that more objective discourse measures of writing performance was, as discussed above, significantly different under the planning time conditions. However, if this is the case, it is associated with the issue of rater reliability. Another possibility is that the absence of a live interlocutor/examiner, as discussed earlier, can reduce the motivation for the test-takers to try their best to improve performances (Wigglesworth 1997). There was no human being listening and reacting to their responses. This may have acted as a factor that failed to improve test-takers' performance adequately to be realized by the scores.

Lastly, complexity and fluency tended to be negatively correlated with accuracy under pre-task and on-line planning time in discourse analysis. Accuracy was likely to have a trade-off relationship with fluency and complexity. When given planning time, the test-takers focused more on conceptual planning of what to speak or write rather than on detailed linguistic forms (Ellis 2005). In the case of speaking, this was even more likely. Speaking should spontaneously be produced in real-time even under the planning time conditions. Accordingly,

the test-takers could not afford the time to monitor. They had to choose which aspect of language production to concentrate on. Thus, as also claimed by Ellis (2005), focusing on fluency/complexity came at the expense of accuracy and vice-versa.

6. Conclusion

This study investigated the effects of pre-task and on-line planning on test-takers' output in the CBT speaking and writing tests. It demonstrated that the opportunity for planning did not influence test-takers' discourse and scores in the speaking test, but had positive effects on fluency and accuracy of the writing test output. In conclusion, the provision of planning time under the CBT testing context made a difference to the quality of writing performance even though it did not lead to achieving a higher score of writing.

Planning time would have been more effective for improving writing than speaking output. The test-takers, who had difficulties formulating messages due to their lack of L2 knowledge, might have more opportunities to conceptualize, formulate, and monitor when given planning time, resulting in meaningful discourse improvements. In contrast to the findings of previous studies on planning time in a classroom context, accuracy was significantly improved with both pre-task and (especially) on-line planning time. It can be seen that the test situation made the quality of performance different from that of the classroom context. As assumed earlier, the CBT writing test might have induced the test-takers to focus on accuracy more because they tend to prioritize correct language form than message convey.

Finally, the limitations of this study should be acknowledged. First, it is possible that the level of participants and task types might, in part, have attributed to the results of this study. In this study, the proficiency of test-takers was generally low, and simple and familiar personal tasks were used to elicit oral and written production. Thus, further studies on planning time with high proficiency test-takers and cognitively demanding types of tasks may attain different findings. Second, the purpose of language behavior in a test context is distinguished from that of the classroom or the real situation. Different language output can be produced depending on the purpose under each context. Therefore, the findings of this study may not be generalized to classroom or other contexts. Lastly, the results may differ in accordance with score analysis methods and inter- and intra-rater reliability. The raters in this study might have bias toward a particular rating item (i.e., accuracy). It is also possible that the raters gave similar scores throughout CFA due to lack of complete understanding of the scoring criteria. Thus, in future studies it is required to pay lose attention to the rater reliability and/or the score statistical analysis method taking into account rater bias and differences.

In view of the limitations above, follow-up studies are needed to explore the validity of using other types of tasks with different levels of proficiency test-takers. The effect of planning time appears to vary in accordance with the characteristics of test-takers and tasks. Also, it is necessary to investigate whether there is a performance difference between talking to a computer and a live interlocutor under planning conditions. It was assumed that there would be a difference especially in fluency, but this study could not examine it clearly.

Reference

Ahmadian, M. J., M. Tavakoli and H. Dastjerdi. 2015. The combined effects of online planning and task structure on complexity, accuracy, and fluency of L2 speech. *Language Learning Journal* 43(1), 41-56.

Anderson, J. R. 1995. *Learning and Memory: An Integrated Approach*. New York: Wiley.

Bachman, L. F. 1990. *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L. F. and A. Palmer. 1996. *Language Testing in Practice.* Oxford: Oxford University Press.

Brown, A. 1993. Test development process: Test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing* 10(3), 277-303.

Crookes, G. 1989. Planning and interlanguage variation. *Studies Language Acquisition* 11(4), 367-83.

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*, 219-239. Amsterdam: John Benjamins.

Ellis, R. 2005. Planning and task-based performance: Theory and research. In R. Ellis, ed., *Planning and Task Performance in a Second Language*, 3-34. Amsterdam: John Benjamins.

Ellis, R. and F. Yuan. 2005. The effects of careful within-task planning on oral and written task performance. In R. Ellis, ed., *Planning and Task Performance in a Second Language*, 167-192. Philadelphia, PA: John Benjamins.

Ellis, R. and F. Yuan. 2004. The effects of planning and fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition* 26(1), 59-84.

Foster, P. and P. Skehan. 1996. The influence of planning and task type on second language performance. *Studies in Second Language Acquisition* 18(3), 299-323.

Iwashita, N., T. McNamara. and C. Elder. 2001. Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design, *Language Learning* 21(3), 401-436.

Kenyon, D. M. and V. Malabonga. 2001. Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning and Technology* 5(2), 60-83.

Ochs, E. 1979. Planned and unplanned discourse. In T. Givon, ed., *Syntax and Semantics 12: Discourse and Syntax*, 51-80. New York: Academic Press.

O'Loughlin, K. 2001. *Studies in Language Testing 13: The Equivalence of Direct and Semi-direct Speaking Tests.* Cambridge: Cambridge University Press.

Ortega, L. 1999. Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition* 21(1), 108-48.

Shohamy, E. 1994. The validity of direct versus semi-direct oral tests. *Language Testing* 11(2), 99-123.

Skehan, P. 1996. Second-language acquisition research and task-based instruction. In J. Willis and D. Willis, eds., *Challenge and Change in Language Teaching,* 17-30. Oxford: Heinemann.

Skehan, P., and P. Foster. 2005. Strategic and on-line planning: The influence of surprise information and task time on second language performance. In R. Ellis, ed., *Planning and Task Performance in a Second Language*, 239-273. Amsterdam: John Benjamins.

Tavarkoli, 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. Philadelphia, PA: John Benjamins.

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.

Mehnert, U. 1998. The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition* 20(1), 83-108.

Weigle. S. C. 2002. *Assessing Writing*. Cambridge: Cambridge University Press.

Wendel, J. N. 1997. *Planning and Second Language Narrative Production.* Unpublished doctoral dissertation, Temple University, Japan.

Wigglesworth, G. 1997. An investigation of planning time and proficiency level on oral test discourse. *Language*

Testing 14(1), 85-106.

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.

Van Patten, B. 1990. Attending to form and content in the input: An experiment in consciousness. *Studies in Second Language Acquisition* 12(3), 287-301.

Yuan, F. and R. Ellis. 2003. The effects of pre-task planning and on-line planning on fluency, complexity, and accuracy in L2 monologic oral production. *Applied Linguistics* 24(1), 1-27.

Examples in: English
Applicable Language: English
Applicable Level: Tertiary

Appendix

Writing Rating Scales

Fluency

5 Vocabulary and content words (nouns, verbs, adjectives, and adverbs) are used in various ways, and the content is very consistent and logical, and the length is relatively very long.

4 Vocabulary and content words (nouns, verbs, adjectives, and adverbs) have been tried in various ways, and the content is consistent, logical, and a bit long.

3 Some attempts have been made to use a variety of vocabulary and content words (nouns, verbs, adjectives, adverbs), but most rely on basic vocabulary and content words. The content is consistent and logical, and the length is average.

2 The use of basic vocabulary and content words (nouns, verbs, adjectives, and adverbs) is mostly large, and the content lacks consistency and logic, and is a bit short in length.

1 The use of basic vocabulary and content words (nouns, verbs, adjectives, and adverbs) seems difficult, and is very limited and repetitive. Due to the lack of consistency and logic of the content, it is difficult to understand what is being conveyed, and the length is short.

Accuracy

5 Errors are hardly noticeable.

4 Errors are rare and insignificant to the extent that they are difficult to understand.

3 Manages most common forms, sometimes with errors. There are significant errors that affect your understanding.

2 Linguistic control is limited; and major errors that are difficult to understand frequently occur.

1 Even basic forms of linguistic control are difficult.

Complexity

5 Confidently attempts a variety of verb forms (e.g., passive, auxiliary, tense, and aspect), even if the use is not always correct. Regularly takes risks grammatically to express complex meanings. Occasionally, attempts the use of coordination and subordination to convey ideas that cannot be expressed in a single clause, even if the result is awkward or inaccurate.

4 Attempts a variety of verb forms (e.g., passive, auxiliary, tense, and aspect), although the use is not always correct. Takes a grammatical risk to express a complex meaning. Frequently attempts to use coordination and subordination to convey ideas that cannot be expressed in a single clause, even if the result is awkward or wrong.

3 Mostly relies on simple verb forms, with some attempts to use a variety of forms (e.g., passive, auxiliary, various tenses, and aspect). Partially attempts to use coordination and subordination to convey ideas that cannot be expressed as a single clause.

2 Frequently generates sentence fragments, even when simple sentence structures are required. It is difficult to attempt to express more complex clause relations, and many errors occur when attempted.

1 Mainly creates sentence fragments and simple phrases. Rarely uses grammatical means to convey ideas better.

