Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 13, No. 2, October 2023
ISSN 2476-5880

TOEFL iBT Speaking Subtest: The Efficacy of Preparation Time on Test-Takers' Performance

Ali Akbar Ariamanesh[1]*, Hossein Barati[2], Manijeh Youhanaee[3]

Received: 20 August 2022 Accepted: 30 December 2022

Abstract
The present study investigates the efficacy of preparation time in four speaking tasks of TOEFL iBT. As the current pre-task planning time offered by ETS is very short, 15 to 30 seconds, we intended to explore how the test-takers' speaking quality would change if the preparation time was added to the response time, giving the respondents a relatively longer online planning opportunity. To this aim, two groups of TOEFL iBT candidates were studied under pre-task and online planning conditions. Totally, 384 elicited speaking samples were first transcribed and then measured in terms of complexity, accuracy, and fluency (CAF). The results yielded by a series of One-way MANOVA revealed the online planning group significantly outperformed the pre-task planning group in terms of accuracy and fluency across all four speaking tasks. Although with less robustness, the online planners had significantly higher speech complexity represented by lexical diversity and left-embeddedness. The results obtained through this study may challenge the efficacy of the currently provided preparation time in TOEFL iBT speaking subsection.

Keywords: TOEFL speaking; Pre-task planning; Online planning; CAF; Iranian test takers

1. Introduction

An important aspect of second language (L2) oral production concerns the effects of planning both before and while performing a task (Skehan, 2016). The provision of planning time has been reported to raise the levels of complexity, accuracy, and fluency (CAF) in L2 production (Khatib & Farahanynia, 2020; Li et al., 2014; Nunan, 1999; Skehan, 2016; Tavakoli & Skehan, 2005). Planning prior to task achievement can potentially be beneficial to L2 learners by enhancing their attentional capacity, which consequently allows them to attend to the linguistic features of their performance (Elder & Iwashita, 2005). Likewise, L2 speakers may access their declarative knowledge through planning and thereby retrieve the required lexico-grammatical information (Kaplan, 2010). Ellis (2005, 2008, 2009) reports when L2 learners are given pre-task planning time, their discourse tends to be more fluent and complex. He also distinguishes between *pre-task* and *within-task* planning before and while performing task-based production, respectively. Pre-task planning is further subcategorized into *rehearsal* (practicing before the main performance) and *strategic* planning (deciding upon the what and

[1] First affiliation, Email: University of Isfahan, Iran, aa.ariamanesh@fgn.ui.ac.ir
[2] Second affiliation, Email: University of Isfahan, Iran, barati@fgn.ui.ac.ir
[3] Third affiliation, Email: University of Isfahan, Iran, youhanaee@fgn.ui.ac.ir

how of performance). Within-task planning, in turn, can be either *pressured* or *unpressured* based on the given time.

The two major forms of planning are thought to contribute rather differently to the quality of learners' production. Pre-task planning is probably more effective during the conceptualization level of speech production and, therefore, leads to more complex and fluent discourse (Ellis, 2008; Skehan, 2016). Online planning, however, may be more effective when formulating ideas as well as monitoring one's oral production, triggering more accuracy (Ellis, 2005, 2008). The inclusion of planning time, therefore, seems to have the potential to decrease competition or trade-off effects (Kaplan, 2010; Skehan, 2014) among salient aspects of L2 oral production. As mentioned earlier, pre-task planning is more likely to enhance fluency and complexity than accuracy. Implied in this claim is that fluency and complexity are probably in competition with accuracy, while undertaking more complex oral production tasks.

What seems to still remain blurred is the interaction effect between the context of task performance and different planning types. That means various forms of planning can operate differently in classroom and testing conditions (Elder & Iwashita, 2005; O'Grady, 2019). Thus, the prime impetus for the present study was to examine the efficacy of preparation time in the TOEFL iBT speaking tasks. Currently, ETS offers the following time rubric for the speaking tasks in focus. Task 1 allows 15 seconds preparation followed by 45 seconds response time, while tasks 2 and 3 include 30- and 60-seconds preparation and response time, respectively. In task 4, the sequence comprises 20 seconds preparation and 60 seconds speaking time. Explicitly, the existing preparation time-limits in the iBT speaking tasks seem too short to be of much positive effect. They might even be somehow stress-making to the test takers, attending simultaneously to the time countdown and the content of their oral performance.

2. Review of Literature

How planning influences L2 oral performance has been addressed in a number of studies (Elder & Iwashita, 2005; Ellis, 2009; Inoue & Lam, 2021; Khatib & Farahanynia, 2020; Lam, 2019; Li et al., 2014; Mehnert, 1998; Nitta & Nakatsuhara, 2014; O'Grady, 2019; Tavakoli & Skehan, 2005; Wigglesworth, 1997, 2000; Wigglesworth & Elder, 2010; Yuan & Ellis, 2003). Wigglesworth (1997) studied the efficacy of pre-task planning in a testing context and found those participants who benefitted from one-minute planning outperformed those without planning. Similarly, the impact of pre-task planning was studied by Mehnert (1998), where she observed that pre-task planning positively correlated with fluency and lexical density of the participants' oral output. As for accuracy and complexity, the 1- and 10-minute planning were found effective, respectively. She further suggested there may be some competition between accuracy and complexity during the preparation time, which means the given time is used to improve either accuracy or complexity. By the same token, Wigglesworth (2000) explored pre-task planning and task familiarity and observed a more beneficial effect for the latter. She finally inferred that test takers might attend more to the content of their production while planning, which can lead to more fluency at the price of accuracy.

Yuan and Ellis (2003) embarked on studying the possible differences in L2 oral language production under pre-task and online planning conditions. They reported that pre-task participants showed higher levels of grammatical complexity, lexical diversity, and speech

fluency than those from the online planning group. On the other hand, online planning promoted accuracy more distinctively. They further discussed since the online planners had more time to access their L2 explicit knowledge, they gained more accuracy. Yuan and Ellis (2003) observed trade-off effects between measures of accuracy and fluency, where online planners prioritized accuracy over fluency, and the pre-task planners did the reverse. Focusing on a testing context, Elder and Iwashita (2005) studied L2 speaking in conjunction with pre-task planning. By far, they concluded that pre-task strategic planning had little effect on the test-takers' oral language performance. Explicitly, the three discourse criteria (CAF) were not significantly different across the plus and minus planning conditions. In another study, Tavakoli and Skehan (2005) conducted a factorial study on L2 oral production with planning time, task structure, and proficiency as the independent variables. Based on their results, speech accuracy and complexity were found to have loaded together. Besides, Tavakoli and Skehan (2005) found that fluency increased significantly with pre-task planning. Similarly, positive effects of pre-task planning were reported on both complexity and accuracy, with a more significant impact on the latter. They further raised the testing context as a potential reason why the test takers paid more attention to accuracy.

Ellis (2008, pp. 496-497) summarizes the major findings of six studies on the role of planning on L2 oral performance. Accordingly, pre-task planning was reported to have positively affected fluency and complexity throughout all six studies. The results for accuracy, however, were less consistent as two of the studies did not find pre-task planning to be much effective. On the same premise, Ellis (2009) concluded the comparatively limited pre-task planning time under testing conditions may have restricted its efficacy. At length, he summarized the following points concerning the effects of planning on L2 production.

- Strategic planning has a more positive effect on fluency than either complexity or accuracy. This is primarily associated with access to ready-made linguistic repertoire achieved during pre-task planning.
- When L2 speakers prioritize complexity or accuracy during pre-task strategic planning, they gain higher quality in one at the expense of the other (trade-off effects).
- Online (within-task) planning has the potential to enhance complexity and accuracy, while it has shown little effect on the fluency aspects whatsoever.

Wigglesworth and Elder (2010) concentrated on the effectiveness of planning time in the IELTS speaking module. They concluded that pre-task planning had no significant effect on their test-takers' spoken discourse, both based on ratings and the CAF measures. It was further suggested that within-task planning might reduce the potential benefits of pre-task planning, especially, if test takers have enough time to respond to an oral task. In contrast, Li et al. (2014) reported that the availability of pre-task planning made their participants perform better, both quantitatively and qualitatively. They also found accuracy as the discourse feature that most benefited from pre-task planning. Moreover, Li et al. (2014) observed that fluency and lexical diversity improved with pre-task planning. However, syntactic complexity was not stable across different planning periods. The impact of pre-task planning was also investigated by Nitta and Nakatsuhara (2014) in a 'paired-format speaking assessment'. Contrary to Li et al., pre-task planning was reported by Nitta and Nakatsuhara to have had little beneficial effect

on the L2 speakers' performance in interactional decision-making tasks. Lam (2019), in turn, explored L2 interactional competence under extended- and short-planned conditions. It needs to be noted that Lam experimented with 4-5 hours of preparation time in the extended mode, which cannot normally be indicative of spontaneous L2 performance. By far, Lam (2019) concluded that pre-task planning may be taken by L2 learners for preparing their speaking content, implying it can be more beneficial to fluency.

On the inquiry about how pre-task planning impacts L2 speaking performance, O'Grady (2019) investigated different planning lengths in a testing context. He reported that the less-proficient participants gained more improvement in line with the provision of pre-task planning. Because O'Grady did not find any significant improvement in his participants' scores after increasing the preparation time, he remarked how pre-task planning affects the CAF measures may not similarly influence human raters. Moreover, he proposed the effects of pre-task planning in educational settings should not be expanded to language testing contexts where scores represent test-takers' ability. O'Grady (2019), however, claimed that the availability of preparation time may support performance in complicated speaking tasks like those with unfamiliar topics and obligatory content. In a recent study, Inoue and Lam (2021) explored how extended pre-task planning could influence iBT test-takers' performance in the academic listening-speaking task of the TOEFL speaking subsection. The possible effects of the experimented planning time (90 seconds) were measured via scores, content accuracy, the CAF triad, and the test-takers' self reports. Having compared the candidates' oral performances under the operational preparation time (20 seconds) and extended planning (90 seconds), Inoue and Lam (2021) did not observe any significant difference between the two planning conditions in terms of the assigned scores, content accuracy, and the discourse features (CAF). They finally advocated the existing preparation time in the iBT academic listening-speaking task, claiming that extended pre-task planning may not be much helpful for the TOEFL test takers to enhance their speaking quality.

The reviewed literature reveals the fact that studies on the effects of different planning types, explicitly pre-task and within-task (online) planning, have found varying results. For instance, pre-task planning was reported by some scholars (Li et al., 2014; Mehnert,1998; Tavakoli & Skehan, 2005; Wigglesworth, 1997; Yuan & Ellis, 2003) to produce beneficial effects on L2 oral performance. However, other studies (Elder & Iwashita, 2005; Inoue & Lam, 2021; Nitta & Nakatsuhara, 2014; O'Grady, 2019; Wigglesworth & Elder, 2010) observed little or no significant effect of pre-task planning on learners' oral discourse.

3. Research Question

Congruent with the aim to investigate the efficacy of the current preparation time in the TOEFL iBT speaking tasks, the following research question was formulated. It needs to be clarified that the CAF features were analyzed simultaneously in each task to expose how they would change with respect to one another.

> *RQ*: Does pre-task vs. online planning in TOEFL iBT speaking tasks make any significant difference to the complexity, accuracy, and fluency of the test-takers' performance?

4. Method

4.1.Participants and Setting

Altogether, 96 TOEFL iBT Iranian candidates participated in this study to provide the required data. Among them, 56 test takers (28 females & 28 males) belonged to the first group called pre-task planning, and the remaining 40 candidates (14 females & 26 males) formed the online planning group. All participants were native speakers of Persian, who had been studying for their main TOEFL exam for several months (mean: 6.66 months) and ranged from 21 to 36 years old (mean: 27.33 years). As for their majors and educational levels, the randomly-selected participants were from humanities, medical fields, and mainly sciences and engineering majors, holding both undergraduate and graduate university degrees. Based on the scores given by ETS (following their operational TOEFL test), the mean speaking score of the participants turned out to be around 24 (24.10 for the pre-task & 23.97 for the online planners), with a score range of 20-28. Of course, to reach the comparability of the two groups in terms of oral proficiency, we had to discard 14 participants (with extreme scores) from the 110 candidates who were tested at the outset of the study.

4.2.Instrumentation

Aiming to compare the test-takers' oral language performance under pre-task and online planning conditions, the following two instruments were applied to elicit the intended oral data.

I) A trial version of the TOEFL iBT exam with all rubrics currently applied by ETS in the operational test. With regard to the trial test's content, all test inputs were extracted from the previous administrations of the test, accessible in the official guide books issued by the ETS organization. In fact, the trial test had specifically been written by a TOEFL preparation institute in Iran to simulate the main test's conditions for the prospective iBT test takers. The internal consistency of this instrument turned preferably high (Cronbach's alpha: 0.80) based on the pre-task planners' performance.

II) A specifically written software resembling the TOEFL iBT speaking subtest except for the time rubric. For each speaking task, the preparation time was added to the response time in the sense that the online planners had only a combined response time starting right from the end of each task's input, without any separate preparation time. The online-planning participants were supposed to start speaking as soon as the response time began. Thus, they had 60 seconds in task 1, 90 seconds in tasks 2 and 3, and 80 seconds in task 4 as the speaking time. Meanwhile, between every two successive tasks, there was a fifteen-second transition time for the respondents to rest. Cronbach's alpha for this instrument was obtained as 0.79.

It should be clarified that note-taking was allowed throughout the four tasks in both instruments. Furthermore, by designing the two conditions of speaking offered by the two instruments, the focus became to investigate the inclusion versus exclusion of a separate preparation time when the total time allocated was held constant across the two situations.

4.3.Procedures

The two data-collection instruments were conducted to the participants some days (10 days on average) before their scheduled operational TOEFL exam. It needs to be clarified that

the first instrument was given to only the pre-task planners who took the simulated test at three different TOEFL preparation centers located in three cities of Iran. In fact, the test was to estimate their skills at the end of their preparatory program prior to their target test. Once the first round of data collection was finished, the online planning software, comprising TOEFL iBT four speaking tasks, was given to the online planning participants who, similarly, came from different cities in Iran. The members of both groups were requested to share the speaking scores they would receive from ETS in the real exam. These scores were considered to ensure the comparability of the two groups regarding their speaking skills.

4.4.Measurements

All 384 transcribed responses were quantified in terms of complexity, accuracy, and fluency (CAF). More specifically, there were four measures representing complexity, two measures accounting for accuracy, and one measure standing for fluency. To analyze complexity, the online software Coh-Metrix 3 (Graesser et al., 2004; McNamara & Graesser, 2012) was deployed. From the numerous variables that Coh-Metrix 3 computes, four measures were selected to represent complexity (Ariamanesh et al., 2020, 2022).

Latent Semantic Analysis (*LSA*): It indicates semantic overlap between sentences within a text. Among different LSA measures, we selected LSASSp, representing the mean overlap among all sentences in a paragraph (McNamara et al., 2014).

Measure of Textual Lexical Diversity (*MTLD*): It refers to the diversity of unique content and function words within a text in proportion to the total number of words in that text (McNamara et al., 2014). MTLD, therefore, is not dependent on text length.

Syntactic Complexity as Left Embeddedness (*SYNLE*): It portrays the mean number of words before the main verb. McNamara et al. (2014) contend that the complexity of a text rises when the mean number of words before the main verbs increases.

Syntactic Complexity of Noun Phrases (*SYNNP*): It shows the mean number of modifiers per noun phrases. SYNNP centers on the idea that a positive correlation exists between the length of NPs in a text and its complexity (McNamara et al., 2014).

To measure accuracy, the transcripts were scrutinized for possible erroneous forms, explicitly those of grammar, lexicon, and discourse (Ellis, 2008). Concerning pronunciation, the deviant forms in the participants' responses were specified during the transcribing phase. In this category, the errors of word pronunciation and stress position were included, not those related to intonation and the other prosodic features. As for grammar, the deviations pertinent to articles, structures, inflections, prepositions, etc., were spotlighted. Concerning vocabulary, the detected errors included the prompting keywords misunderstood by the test takers (e.g., '*distinction*' for '*extinction*'), basically incorrect forms (e.g., *effectly*), and nonsense words (e.g., something partially pronounced similar to '*malachite*' yet not nearly close to it). The misused cohesive ties (e.g., '*although* & *but*' used concomitantly) were also specified. Ultimately, the following technique was applied to measure the accuracy of each response (adapted from Ellis & Barkhuizen, 2005).

*Accuracy = 100 – [(number of errors of all types / number of all words) * 100]*

Regarding the fluency measure, the ratio of uttered syllables to the number of seconds was computed (Ellis, 2008). Basically, two sets of criteria are used to estimate fluency: 1) The temporal facets, including speech rate, and 2) The repair phenomena (Ellis, 2009; Jong, 2018; Yan et al., 2020). Focusing on the latter approach, we discarded the words or phrases that had successively been repeated by the respondents from the transcripts, but the revisions and self-corrections were kept. In order to count the number of produced syllables, each transcript was pasted into an online tool called SYLLABLE COUNTER (syllablecounter.org), and then, this formula was conducted to quantify the fluency of each speaking sample (Ellis, 2008).

*Fluency = (Total number of syllables / Total number of seconds) * 60*

The three integrated speaking tasks (2, 3, & 4) were further analyzed for content accuracy (Frost et al., 2021) because the respondents were required to reflect the central concepts conveyed by the stimulus texts. Accordingly, we considered such criteria as the number of prompting key points transferred and how well the main ideas from the task input were summarized or paraphrased (Frost et al., 2011). Based on what TOEFL iBT respondents are usually asked to produce throughout the integrated speaking tasks, the stimulus texts were divided into two lines of ideas in each task. In order to ensure how the main ideas are summarized, a relatively large number of suggested responses to the integrated speaking tasks from official TOEFL iBT guide books were meticulously studied. Aiming to quantify the content accuracy of the integrated oral responses, therefore, we reviewed the transcripts and computed a percent score for each response according to the number as well as accuracy and transparency of the transferred key ideas. It should be mentioned that all procedures of quantifying the content accuracy were undertaken by two of the researchers in the present study, with inter-rater reliability (Manzano, 2022) of around 0.91 for both groups.

4.5.Analysis

The numerical data was organized in SPSS (26) and a series of One-way MANOVA (Bachman, 2004; Pallant, 2020; Tabachnick & Fidell, 2013) was conducted to compare the two speaking conditions in terms of the CAF measures. This analytical route was selected since there was one between-groups independent factor, planning type with two levels (pre-task & online), and three dependent variables (CAF), altogether with seven subcategories. Statistically speaking, MANOVA is recommended to be used because it does not run the risk of an 'inflated type I error' (Pallant, 2016, p. 151). In the meantime, conducting MANOVA was justified in this study since the dependent variables were conceptually relevant as the trade-off effects were already proved to exist among the mentioned discourse features (Ellis, 2009; Kaplan, 2010; Skehan, 2014; Yuan & Ellis, 2003, among others).

5. Results

5.1.Speaking Task One

A One-way MANOVA was conducted to compare the pre-task and online planning conditions in terms of CAF in the first iBT speaking task. Table (1) depicts the related descriptive statistics, where the online planning group showed higher mean scores than the pre-task planning group in five measures. Further, planning type turned out to be significant with a large effect size, F (6, 89) = 3.73, p = .002, Wilks' Lambda = .79, partial eta squared = .20.

 ISSN 2476-5880

Table 1
Descriptive Statistics of the CAF Measures in Speaking Task 1

CAF Measures	Planning Type	Mean	Std. Deviation	N
Fluency	Pre-task Planning	160.2321	33.58625	56
	Online Planning	180.8098	35.82576	40
Accuracy	Pre-task Planning	92.0409	4.46055	56
	Online Planning	93.8887	3.00668	40
Lexical Diversity	Pre-task Planning	55.0570	12.85743	56
	Online Planning	63.2195	14.73665	40
Left Embeddedness	Pre-task Planning	4.0593	2.02043	56
	Online Planning	4.9443	1.73672	40
Modifiers per NP	Pre-task Planning	.6680	.19479	56
	Online Planning	.7048	.20871	40
LSA	Pre-task Planning	.2493	.12000	56
	Online Planning	.2402	.09502	40

Table (2) reveals that the two groups of participants had significantly different oral performances in terms of fluency, accuracy, lexical diversity, and left-embeddedness. Regarding all these four measures, the online planning group had significantly higher mean scores than the pre-task planning group when responding to the TOEFL iBT independent speaking task 1.

Table 2
Tests of Between-Subjects Effects in Speaking Task 1

Source	Dependent Variable	df	Mean Square	F	Sig.	Partial Eta Squared
Planning Type	Fluency	1	9880.218	8.285	.005	.081
	Accuracy	1	79.673	5.176	.025	.052
	Lexical Diversity	1	1554.630	8.321	.005	.081
	Left Embeddedness	1	18.274	5.020	.027	.051
	Modifiers per NP	1	.031	.781	.379	.008
	LSA	1	.002	.157	.693	.002

 ISSN 2476-5880

5.2.Speaking Task Two

For iBT integrated speaking tasks (2, 3, & 4), content accuracy was added to the six dependent variables when comparing the two planning types. The descriptive statistics of the CAF subcategories obtained from One-way MANOVA are presented in Table (3). Also, planning type was found significant in the second iBT speaking task, having a large effect size, F (7, 88) = 6.72, p = .000, Wilks' Lambda = .65, partial eta squared = .34.

Table 3
Descriptive Statistics of the CAF Measures in Speaking Task 2

CAF Measures	Planning Type	Mean	Std. Deviation	N
Fluency	Pre-task Planning	164.0714	34.34946	56
	Online Planning	177.4490	27.29005	40
Accuracy	Pre-task Planning	89.9786	4.60958	56
	Online Planning	92.1853	3.08342	40
Lexical Diversity	Pre-task Planning	53.4132	14.28318	56
	Online Planning	61.1415	15.52186	40
Left Embeddedness	Pre-task Planning	4.3202	2.35365	56
	Online Planning	5.0590	2.87585	40
Modifiers per NP	Pre-task Planning	.8002	.19449	56
	Online Planning	.7897	.12831	40
LSA	Pre-task Planning	.1829	.07967	56
	Online Planning	.1968	.07959	40
Content Accuracy	Pre-task Planning	59.4643	21.54654	56
	Online Planning	81.8750	12.38731	40

More specifically, the two planning groups were significantly distinguishable in terms of fluency, accuracy, lexical diversity, and content accuracy. Concerning all of these significant differences, the online planning participants outperformed the pre-task planners. For the remaining three measures, i.e., left-embeddedness, modifiers per NP, and LSA, the differences between the two groups were not significantly different (Table 4).

Table 4
Tests of Between-Subjects Effects in Speaking Task 2

Source	Dependent Variable	df	Mean Square	F	Sig.	Partial Eta Squared
Planning Type	Fluency	1	4175.720	4.178	.044	.043
	Accuracy	1	113.620	6.938	.010	.069
	Lexical Diversity	1	1393.616	6.354	.013	.063
	Left Embeddedness	1	12.737	1.909	.170	.020
	Modifiers per NP	1	.003	.088	.768	.001
	LSA	1	.005	.710	.402	.007
	Content Accuracy	1	11718.936	34.950	.000	.271

5.3.Speaking Task Three

Similar to the analytical route conducted to the previous two tasks, a One-way MANOVA was run to compare the oral performances under the two planning conditions in TOEFL iBT speaking task 3. The descriptive statistics of the dependent variables for the two groups are summarized in Table (5). As for the independent factor, multivariate tests of pre-task versus online planning revealed that planning type had a significant effect in task 3, with a very large effect size, F (7, 88) = 22.11, p = .000, Wilks' Lambda = .36, partial eta squared = .63. Furthermore, the tests of between-subjects effects disclosed the online planners significantly outperformed the pre-task planners in terms of fluency, accuracy, left-embeddedness, and content accuracy. Regarding lexical diversity, however, the pre-task planning group showed a significantly higher mean value (Table 6).

Table 5
Descriptive Statistics of the CAF Measures in Speaking Task 3

CAF Measures	Planning Type	Mean	Std. Deviation	N
Fluency	Pre-task Planning	145.7143	34.52520	56
	Online Planning	180.7493	22.94444	40
Accuracy	Pre-task Planning	87.3525	4.38884	56
	Online Planning	92.4740	3.17792	40
Lexical Diversity	Pre-task Planning	59.6205	20.17523	56
	Online Planning	50.0957	13.95527	40
Left Embeddedness	Pre-task Planning	3.8889	2.22009	56
	Online Planning	5.0103	1.93828	40

Modifiers per NP	Pre-task Planning	.7455	.15505	56
	Online Planning	.7220	.16998	40
LSA	Pre-task Planning	.2027	.08648	56
	Online Planning	.2162	.05843	40
Content Accuracy	Pre-task Planning	49.1071	20.82628	56
	Online Planning	86.6250	7.87625	40

Table 6
Tests of Between-Subjects Effects in Speaking Task 3

Source	Dependent Variable	df	Mean Square	F	Sig.	Partial Eta Squared
Planning Type	Fluency	1	28640.470	31.272	.000	.250
	Accuracy	1	612.028	39.587	.000	.296
	Lexical Diversity	1	2116.836	6.637	.012	.066
	Left Embeddedness	1	29.338	6.604	.012	.066
	Modifiers per NP	1	.013	.496	.483	.005
	LSA	1	.004	.742	.391	.008
	Content Accuracy	1	32843.757	117.501	.000	.556

5.4.Speaking Task Four

The oral responses to the fourth iBT speaking task were analyzed, conducting a One-way MANOVA, where planning type as the between-groups factor showed a significant effect with a very large effect size, F (7, 88) = 14.99, p = .000, Wilks' Lambda = .45, partial eta squared = .54. Examining the mean scores computed for the dependent variables (CAF measures) across the planning conditions (Table 7), and the significant points of difference (Table 8), we found that the online planning participants significantly outperformed the pre-task planners in terms of fluency, accuracy, lexical diversity, and content accuracy. Concerning the LSA measure, conversely, the pre-task planning group showed a significantly higher mean score.

In summary, throughout all four iBT speaking tasks, the type of planning (pre-task vs. online) was found to have had a significant impact on the participants' speaking performance, which was measured by a group of CAF variables. With respect to the detected significant differences, almost all the time, the online group presented a higher speaking quality. Actually, in terms of only two measures, i.e., lexical diversity in task 3 and LSA in task 4, the pre-task group showed a better oral performance.

 ISSN 2476-5880

Table 7
Descriptive Statistics of the CAF Measures in Speaking Task 4

CAF Measures	Planning Type	Mean	Std. Deviation	N
Fluency	Pre-task Planning	155.2857	29.60818	56
	Online Planning	181.0090	26.76106	40
Accuracy	Pre-task Planning	88.5129	4.92451	56
	Online Planning	92.0140	2.55913	40
Lexical Diversity	Pre-task Planning	47.0859	12.10784	56
	Online Planning	54.2453	17.27479	40
Left Embeddedness	Pre-task Planning	4.8230	2.01905	56
	Online Planning	5.5255	2.18063	40
Modifiers per NP	Pre-task Planning	.6371	.18557	56
	Online Planning	.7010	.14232	40
LSA	Pre-task Planning	.2477	.09929	56
	Online Planning	.1720	.06219	40
Content Accuracy	Pre-task Planning	57.2321	17.86107	56
	Online Planning	81.2500	8.67874	40

Table 8
Tests of Between-Subjects Effects in Speaking Task 4

Source	Dependent Variable	df	Mean Square	F	Sig.	Partial Eta Squared
Planning Type	Fluency	1	15439.373	19.060	.000	.169
	Accuracy	1	286.020	16.918	.000	.153
	Lexical Diversity	1	1195.983	5.706	.019	.057
	Left Embeddedness	1	11.514	2.642	.107	.027
	Modifiers per NP	1	.095	3.332	.071	.034
	LSA	1	.134	18.126	.000	.162
	Content Accuracy	1	13460.007	61.769	.000	.397

6. Discussion

As time rubric is an influential test-method facet (Bachman & Palmer, 1996), this study attempted to explore the role of different types of planning time, explicitly pre-task vs. online,

in TOEFL iBT speaking subtest. More clearly, we set to investigate the efficacy of the existing preparation time in the iBT speaking tasks on the test-takers' oral language performance. Given the fact that the preparation time in TOEFL iBT speaking module is very limited, this study aimed to explore possible variations when the preparation time was added to the response time in each task. In line with this impetus, two comparable groups of TOEFL iBT candidates were recruited to respond to the four speaking tasks under pre-task and online planning conditions. Throughout the following subsections, each speaking task is discussed in light of the obtained results under the two planning situations.

6.1.The Independent, Option-Based Speaking Task 1

The results revealed that the speaking quality of the participants significantly increased in terms of fluency, accuracy, and two of the complexity measures, including lexical diversity and left-embeddedness, under the online planning condition. As to the other two complexity measures, modifiers per NP and LSA, the difference between the pre-task and online planners was not significant. Even these two recent measures, which did not show any meaningful difference between the two groups, can be interpreted in favor of online planning since the exclusion of preparation time did not decrease the test-takers' speaking complexity. One possible reason for the better performance of the online planners may relate to the more convenient response time they experienced, through which they were less pressured by time constraints (Ellis, 2009). As the first iBT speaking task is basically opinion-based, it seems a longer online planning time benefits the respondents more than when both the preparation and response time are very short.

6.2.The Integrated Read-Listen-Speak, Campus-Related Task 2

The results obtained for the second speaking task revealed the online planning group was significantly better than the pre-task planning group with regard to fluency, accuracy, and lexical diversity as an aspect of complexity. However, left-embeddedness, modifiers per NP, and LSA, which represent complexity, did not indicate any meaningful differences between the two groups. In fact, the complexity measures of modifiers per NP and LSA were not significantly different in both speaking tasks 1 and 2. An interesting outcome observed in the second task was that the online planners were much more successful in transferring the key ideas from the stimulus texts. This achievement may have been exerted by more response time as well as the rapid transition from receiving the input texts to reproducing them orally. This immediate transition might have reduced their memory demands (Skehan, 2016) when summarizing the prompting ideas.

6.3.The Integrated Read-Listen-Speak, Academic Task 3

The comparisons between the two groups of participants responding to the third speaking task, once more, showed the online planners outperformed the pre-task planners in terms of fluency, linguistic accuracy, as well as content accuracy. These findings exactly replicate those obtained in the previous task. Nonetheless, among the two complexity measures that were found significant in task 3, lexical diversity was higher among pre-task planners, whereas left-embeddedness was higher in favor of the online participants. The measures of

 ISSN 2476-5880

modifiers per NP and LSA were not significantly different between the two groups, which corresponds to the previous task. It seems tenable to claim when pre-task planners focused more on the diversity of their lexical items, their syntactic complexity manifested by left-embeddedness decreased. The case for the online planners appears to have been the reverse, as they chose to make their structures more complex than focus on their lexical diversity. Thus, there might be some trade-off effect (Kaplan, 2010; Skehan, 2014) between lexical diversity and syntactic density.

6.4.The Integrated Listen-Speak, Academic Task 4

The results found in speaking task 4, similarly, disclosed the low efficacy of the currently-offered preparation time in the TOEFL speaking subtest. That is because the online planning respondents experienced a significantly better speech performance in terms of fluency, form and content accuracy, and lexical diversity. The only measure on which the pre-task planners showed a higher mean was the semantic overlap or LSA. The other complexity measures, i.e., left-embeddedness and modifiers per NP, were not significantly different between the two planning groups. An important point concerning lexical diversity and LSA, which were found significant in task 4, is that these two measures are essentially vocabulary-based. Moreover, as the lexical diversity increases, the semantic overlap among the sentences within a text (what LSA measures) may decrease (McNamara et al., 2014). The online planners might have been more successful in reproducing the lexical items offered by the prompts due to the increased online planning opportunity. This gain probably caused less semantic overlap (LSA) in the uttered ideas by the online participants.

6.5.Comparison with Similar Studies

The findings of the present study seem to be incompatible with those investigations that reported positive effects of pre-task planning on L2 learners' speaking quality (Khatib & Farahanynia, 2020; Mehnert, 1998; Tavakoli & Skehan, 2005; Wigglesworth, 1997). Of course, the pre-task planning time in the spotlighted studies was significantly longer (1 to 10 minutes) than the preparation time given in TOEFL iBT speaking tasks (15 to 30 seconds). Unlike Wigglesworth (2000), who examined 5 minutes pre-task planning, we did not find the trade-off effects between fluency and accuracy throughout the four speaking tasks. Also, our findings contradict those found by Li et al. (2014), claiming that the availability of pre-task planning (up to 5 minutes) made the participants perform more fluently and accurately. As to fluency, once again, our results are not in line with Bui and Huang (2016), who found beneficial effects of pre-task planning (10 minutes). The most judicious deduction as to why the outlined discrepancies occurred may be pertinent to the length of pre-task planning. To make it clearer, the pre-task planning time in TOEFL iBT speaking subtest ranges between 15 and 30 seconds, whereas the mentioned studies provided up to 10 minutes pre-task planning. It would be tenable to claim the contradictions were probably rooted in the much shorter and less effective preparation time (15-30 seconds) given to the pre-task planners in the present study.

Across the three integrated speaking tasks, content accuracy was significantly higher in favor of the online planning condition. This finding is somehow against what O'Grady (2019) reported, where he implied pre-task planning could positively affect speaking tasks with

obligatory content (such as integrated tasks). An appealing argumentation for the higher levels of content accuracy gained by the online planners in the current study can be attributed to the longer response time they were given to transfer the main ideas from the task input.

The outcomes of the present study, however, corroborate Yuan and Ellis (2003), who found that online planning promoted both accuracy and grammatical complexity. One reason for more accurate discourse under online planning condition may pertain to the accessibility of L2 explicit knowledge by test takers when they have more time to reach it (Ellis, 2005). On the inefficacy of pre-task planning in most testing contexts, the present investigation validates the results reported by Elder and Iwashita (2005), Wigglesworth and Elder (2010), and Nitta and Nakatsuhara (2014). These studies reported the inclusion of pre-task planning had little or no positive effect in helping learners promote their L2 oral production. In a similar way, our findings confirm Ellis (2009), contending that pre-task strategic planning has shown less positive effect in testing contexts than in educational ones. The main reason is targeted to the limited pre-task planning time usually offered under testing conditions. This claim was perceivably substantiated by the current study as the existing preparation time in the TOEFL iBT speaking module is very short. The fact that we observed accuracy and, to a lesser degree, complexity increase with expanding the online planning time corresponds to Ellis (2009). Yet, the enhancement of speech fluency with online planning, found in our investigation, does not match his conclusions. Regarding this contrast, it can be justifiable to suggest when the participants were given more time to express themselves, they were freer in formulating their ideas, which increased their rate of delivery.

In sum, integrating the preparation and response time and thereby providing the test takers with a relatively longer online planning opportunity engendered noticeable improvement in their oral performance. As a high-stakes test like TOEFL iBT does not normally allow a preferably long pre-task planning time, and even extending the preparation time to some degree might not assist the test takers (Inoue & Lam, 2021), a less-pressured online planning expereince can be more beneficial and less stress-making to the respondents.

7. Conclusion

This study aimed to inspect the efficacy of the currently-offered preparation time in the TOEFL iBT speaking tasks. Accordingly, two homogeneous groups of TOEFL candidates were compared under pre-task and online planning conditions. By far, we concluded that the online planning situation experimented in this study triggered more speech fluency and accuracy to a great deal, and to a lesser extent, more speech complexity. It should, however, be acknowledged that we examined only one proficiency level, which was around TOEFL iBT's mean score. More comprehensive results could be obtained if different proficiency levels (Kim, 2021) were studied. Another potential limitation of the current investigation pertains to the size of the studied sample (96 participants overall), which may constrain the implications. All in all, although the observed results seem to be against the inclusion of pre-task planning in the TOEFL speaking module, we do not intend to ignore the positive effects of preparation on the test-takers' oral performance. Instead, this study may show the existing preparation time in the speaking subsection is not much helpful. It seems these preparation periods, ranging from 15 to 30 seconds, are too short and hence stress-making to be beneficial for the examinees to

deliver their speaking skills in the high-stakes test. One possible solution to improve the efficacy of the time rubric, as this study revealed, is by expanding the response time, which can lead to a more effective online planning opportunity.

Declaration of Conflicting Interests

There is no conflict of interests related to this investigation.

Funding

This research did not receive any specific grant from funding agencies.

References

Ariamanesh, A. A., Barati, H., & Youhanaee, M. (2020). TOEFL iBT integrated speaking tasks: A comparison of test-takers' performance in terms of complexity, accuracy, and fluency. *Iranian Journal of Applied Linguistics (IJAL), 23*(2), 33-62. http://ijal.khu.ac.ir/article-1-3091-en.html

Ariamanesh, A. A., Barati, H., & Youhanaee, M. (2022). TOEFL iBT Iranian test-takers' oral language performance: A comparison between independent and integrated speaking tasks. *International TESOL Journal, 17*(1), 25-43. https://connect.academics.education/index.php/itj/article/view/371

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.

Bui, G., & Huang, Z. (2016). L2 fluency as influenced by content familiarity and planning: Performance, measurement, and pedagogy. *Language Teaching Research, 22*(1), 94-114. https://doi.org/10.1177/1362168816656650

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford University Press.

Elder, C. A., & Iwashita, N. (2005). Planning for test performance: Does it make a difference? In R. Ellis (ed.), *Planning and task performance in a second language* (pp. 219 - 238). John Benjamins Publishing Company.

Ellis, R. (2005). *Planning and task performance in a second language*. John Benjamins Publishing Company.

Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). Oxford University Press.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics, 30*(4), 474-509. https://doi:10.1093/applin/amp042

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford University Press.

Frost, K., Elder, C., & Wigglesworth, G. (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. https://doi.org/10.1177/0265532211424479

Frost, K., Wigglesworth, G., & Clothier, J. (2021). Relationships between comprehension, strategic behaviors and content-related aspects of test performances in integrated speaking tasks. *Language Assessment Quarterly, 18* (2), 133-153. https://doi:10.1080/15434303.2020.1835918

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers, 36*, 193-202. https://doi.org/10.3758/BF03195564

Inoue, C., & Lam, D. M. K. (2021). The effects of extended planning time on candidates' performance, processes, and strategy use in the lecture listening-into-speaking tasks of the TOEFL iBT® Test. *ETS Research Report Series, 2021*(1), 1-32. https://doi.org/10.1002/ets2.12322

Jong, N. H. D. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly, 15*(3), 237-254. https://doi.org/10.1080/15434303.2018.1477780

Kaplan, R. B. (2010). *The Oxford handbook of applied linguistics* (2nd ed.). Oxford University Press.

Khatib, M., & Farahanynia, M. (2020). Planning conditions (strategic planning, task repetition, and joint planning), cognitive task complexity, and task type: Effects on L2 oral performance. *System 93*(1), 1-12. https://doi.org/10.1016/j.system.2020.102297

Kim, P. (2021). Contrasting groups analysis of TOEFL® iBT test cut scores and the common European framework of reference (CEFR) proficiency levels: Kernel density estimation of an English learners' corpus. *International Journal of Language Testing*, *11*(1), 88-102. https://www.ijlt.ir/article_128362.html

Lam, D. M. K. (2019). Interactional competence with and without extended planning time in a group oral assessment. *Language Assessment Quarterly, 16*(1), 1-20. https://doi.org/10.1080/15434303.2019.1602627

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Li, L., Chen, J, & Sun, L. (2014). The effects of different lengths of pre-task planning time on L2 learners' oral test performance. *TESOL Quarterly, 49*(1), 38-66. https://doi.org/10.1002/tesq.159

Manzano, D. L. (2022). Examining the interrater reliability between self- and teacher assessment of students' oral performances. *International Journal of Language Testing*, *12*(2), 128-144. https://doi.org/10.22034/ijlt.2022.157130

McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis*: *Identification, investigation, and resolution* (pp. 188-205). Hershey, PA: IGI Global. https://doi.org/10.4018/978-1-60960-741-8.ch011

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

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. http://www.jstor.org/stable/44486384

Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on paired oral test performance. *Language Testing, 31*(2), 147-175. https://doi.org/10.1177/0265532213514401

Nunan, D. (1999). *Second language teaching and learning*. Heinle & Heinle Publishers.

O'Grady, S. (2019). The impact of pre-task planning on speaking test performance for English-medium university admission. *Language Testing, 36*(4), 505-526. https://doi.org/10.1177/0265532219826604

Pallant, J. (2016). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (6th ed.). Open University Press.

Pallant, J. (2020). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (7th ed.). Routledge.

Shohamy, E., Or, L. G., & May, S. (2017). *Language testing and assessment* (3rd ed.). Springer International Publishing AG.

Skehan, P. (2014). *Processing perspectives on task performance*. John Benjamins Publishing Company.

Skehan, P. (2016). Tasks versus conditions: Two perspectives on task research and their implications for pedagogy. *Annual Review of Applied Linguistics, 36*, 34-49. https://doi.org/10.1017/S0267190515000100

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson Education.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239 - 273). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.11.15tav

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*(1), 85-106. https://doi.org/10.1177/026553229701400105

Wigglesworth, G. (2000). Issues in the development of oral tasks for competency-based assessments of second language performance. In G. Brindley (ed.), *Studies in immigrant English language assessment* (Vol. 1) [Research Series 11]. National Centre for English Language Teaching and Research Macquarie University.

Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly, 7*(1), 1-24. https://doi.org/10.1080/15434300903031779

Yan, X., Kim, H. R., & Kim, J. Y. (2020). Dimensionality of speech fluency: Examining the relationships among complexity, accuracy, and fluency (CAF) features of speaking performances on the Aptis test. *Language Testing, 38*(4), 1-26. https://doi.org/10.1177/0265532220951508

Yuan, F., & Ellis, R. (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. https://doi.org/10.1093/applin/24.1.1