Iranian Journal of Applied Linguistics (*IJAL*)

Vol. 23, No. 2, September 2020, 33- 62

TOEFL iBT Integrated Speaking Tasks: A Comparison of Test-Takers' Performance in Terms of Complexity, Accuracy, and Fluency

Ali A. Ariamanesh, Hossein Barati*, Manijeh Youhanaee

Faculty of Foreign Languages, University of Isfahan, Isfahan, Iran

Abstract

This study compares three integrated tasks of the TOEFL iBT speaking subtest in terms of complexity, accuracy, and fluency. To this end, a group of TOEFL iBT Iranian candidates took a simulated TOEFL iBT some days prior to their real exam. The collected oral responses were first transcribed and then quantified using software such as 'Syllable Counter' and 'Coh-Metrix3' for fluency and complexity, respectively. For accuracy, however, the responses were tallied manually. The results revealed the responses to the three speaking tasks were significantly different in terms of fluency. The difference in the accuracy index also turned significant, though the pairwise comparisons showed some inconsistencies. As for the selected complexity measures, lexical diversity, the mean number of modifiers per NP, and latent semantic analysis all showed significant differences between tasks 2 and 3 on the one hand and task 4 on the other. Left-embeddedness, however, revealed no significant difference among the three tasks. The results may support the influential role of prompting texts in such integrated speaking tasks.

Keywords: TOEFL integrated speaking, Complexity, Accuracy, Fluency, Iranian test-takers

Article Information:

Received: 23 May 2020 **Revised:** 11 July 2020 **Accepted:** 28 August 2020

Corresponding author: Faculty of Foreign Languages, University of Isfahan, Azadi Square, Isfahan, Iran, PC: 817467344. **Email:** barati@fgn.ui.ac.ir

1. Introduction

One topic of interest in research on speaking assessment is whether speaking ability should be measured independently or in conjunction with other skills, namely reading and listening. This raised a dichotomy of independent speaking tasks, which focus only on oral language production, and integrated ones that are supposed to enrich the oral language performance by preceding reading and listening tasks (Wigglesworth & Frost, 2017). The TOEFL iBT integrated speaking tasks, which are supposed to reflect the academic uses of English, have similarly been devised with the same assumption. Such integrated tasks are essentially expected to serve authenticity and validity purposes (Brooks & Swain, 2014; Brown & Abeywickrama, 2018; Carr, 2011; Farnsworth, 2013; Ockey et al., 2014). The canon of such test tasks, therefore, subsumes skills-integration aimed at enhancing the generalizability of the assessment outcomes to the target language use (TLU) domain (Bachman & Palmer, 2010; Carr, 2011).

During the evolution of integrated techniques, for example, Weir (1990) raised authenticity and the use of context as two features of integrated test tasks, which are claimed to replicate reality and extended contextualization, respectively. Contrary to these expectations, he believed integrated test tasks probably approach validity at the expense of reliability. Douglas (1997), in turn, argued both logically and practically, it is impossible to test speaking ability independently. Such a claim can also be inferred from Brown and Abeywickrama (2018, p.156). Yet, some scholars, as Brown et al. (2001), claimed the integrated form of speaking assessment increases the cognitive load on examinees. As a result, the added burden may lead to a less satisfying speaking performance (Kormos et al., 2020).

It seems TOEFL iBT speaking subtest is intended to cater for the predictability of its results by resorting to skills-integration (Luoma, 2004; Shohamy et al., 2017; Taylor, 2011). However, as the degree of integration in test tasks increases, it turns more complex to accurately estimate test-takers'

abilities, specifically in the case of oral language performance (Brown, 2004; Fulcher, 2003). Following this potential drawback, the current study was designed to explore how the oral performance by TOEFL iBT test-takers is affected by stimulus texts in the integrated speaking tasks.

1.1. TOEFL iBT Integrated Speaking Tasks

The integrated speaking tasks investigated in this study comprise the following characteristics (adapted from ets.org/toefl).

> *Task Two*: It comprises a read-listen-speak sequence on a campus-related issue. This task needs respondents to talk about the speaker's opinions in the listening part, which are based on the reading prompt.
>
> *Task Three*: This task is similar to the previous one for including a read-listen-speak pattern. Of course, task 3 revolves around an academic/scientific topic. The respondents should convey the gist of the lecturer's comments on the reading prompt.
>
> *Task Four*: As an integrated activity, task 4 takes a listen-speak form, where respondents are required to summarize the lecturer's main points. Similar to the previous task, this speaking attempt is on an academic/scientific topic.

Tasks 2 and 3 are comparable based on the sequence of skills they entail, while they are different on the grounds of the central topic. Based on the latter case, tasks 3 and 4 are more similar because they both go around some scientific topics. Of course, the quality of the oral language produced by the respondents might vary in line with how well they are able to decode the input texts in the first place. Accordingly, the tasks in focus may cause varying levels of complexity, accuracy, and fluency in one's speech due to their capabilities in reading, and especially, listening skills. Given this, the present study was intended to explore how the integrated nature of these tasks influences the speaking performance in testing conditions.

2. Literature Review

Since its first administration in 2005, the speaking subtest of TOEFL iBT has been an exciting forum for research in language testing (Brooks & Swain, 2014; Brown & Ducasse, 2019; Crossley & Kim, 2019; Cumming et al., 2005; Frost et al., 2011, 2019, & 2020; Huang et al., 2016 & 2018; Kyle et al., 2015; Lee, 2005). In one of the earliest studies in this field, Cumming et al. (2005) investigated the integrated tasks of the TOEFL test and found that the majority of respondents with lower proficiency had trouble understanding the stimulus texts to produce their ideal discourse. This situation certainly poses problems to test-takers if they fail to properly get the ideas from the preceding texts. Aimed to explore this challenge, Lee (2005) carried out a study on the TOEFL speaking prototype tasks. Lee concluded that when there are two distinct aspects of language serving as stimulus, say listening and reading, to trigger a third construct (speaking), the reliability of the total score representing oral language production might be called into question. In other words, any deficiency in each of the stimulus skills can interfere with the subsequent speaking output. Likewise, Frost et al. (2011) explored an integrated listening-speaking task and found a direct relationship between test-takers' speaking proficiency and their success in carrying over the stimulus key ideas to the following oral performance.

Brooks and Swain (2014) endeavored to examine the validity argument of TOEFL iBT speaking tasks. They intended to find out to what extent scores on the TOEFL speaking tasks reflect students' real academic oral language productions. Having compared a group of participants' oral productions during TOEFL iBT, in-class, and out-of-class settings, they found that the participants were most grammatically complex as well as most inaccurate throughout the speaking tasks of TOEFL. Brooks and Swain, however, did not discuss the possible effects of the integrated speaking tasks on the respondents' performance. Kyle et al. (2015) studied the TOEFL iBT speaking module with an emphasis on how the tasks might elicit different oral productions in terms of lexical and cohesive features. Regarding the integrated speaking tasks, they observed both similarities and differences among them based on various factors.

Similarly, Huang et al. (2016) investigated three integrated speaking tasks with a reading-listening-speaking sequence sampled from TOEFL iBT materials. They focused on participants' topical knowledge, among other features, and observed some fluctuating effects for the specificity of topic. Huang et al. (2016) concluded that the integrated speaking tasks may not monolithically decrease the influence of topical knowledge on test-takers' performance. The recent finding may challenge test fairness as integrated speaking tasks could be in favor of some test-takers.

To explore the role of topical knowledge in speaking assessment, Huang et al. (2018) exploited four integrated reading-listening-speaking tasks from TOEFL iBT materials. Not surprisingly, their study lent support to the significant role played by topical knowledge in the sense that those test-takers with more topical knowledge in relation to the prompting texts benefitted more from the content provided by the task input. Needless to say, this finding raises doubts over the supposed impartiality in language assessment (Bachman & Palmer, 2010). In a similar study, an iBT reading-listening-speaking task was explored by Frost et al. (2019) to disclose the relationships among stimulus content, task demands, and the oral discourse produced by test-takers. For this purpose, Frost et al. (2019) scrutinized the oral language produced by a group of TOEFL iBT test-takers across three proficiency levels. Their findings showed the high proficient participants reproduced more accurate discourses in terms of the ideas covered in the source texts.

In another attempt, Crossley and Kim (2019) set out to investigate how text integration can affect oral language performance. Specifically, relational, propositional, and syntactic features of the source text were addressed to determine in what ways they may affect the following speaking performance. To this purpose, a listen-then-speak task derived from TOEFL iBT was adopted to elicit the participants' oral language productions. The study concluded that the linguistic elements of the source text in general, and its frequency of lexical-propositional elements in specific, were strong predictors of the subsequent oral output. In other words, Crossley and Kim (2019) observed a significant effect by the source text's keywords on the quality of their test-takers' speaking

performance. Besides, a more distinctive role was reported for the form and content of the stimulus text than for the test-takers' individual features, including their working memory capacity. Brown and Ducasse (2019), in turn, found the academic integrated speaking tasks of TOEFL iBT to be relatively valid indicators of oral activities practiced in real academic settings. Of course, Brown and Ducasse did not provide clear information as to how their test-takers' speaking skills varied across the three integrated speaking tasks. In a more recent study, Frost et al. (2020) concentrated on the potential relations between test-takers' comprehension, strategy-use, and their oral language performance in a TOEFL iBT reading-listening-speaking task. Based on their results, Frost et al. (2020) reported on a distinguishing role of proficiency, where more proficient participants were observed to be more successful in summarizing and reproducing the prompting ideas in their speech.

In sum, the body of previous work on TOEFL iBT integrated speaking tasks has addressed the effects of several factors associated with such testing formats. These factors include test-takers' proficiency and topical knowledge, validity argument of the speaking tasks, and the roles played by various aspects of the stimulus texts such as their lexical and textual content. However, what seems to be lacking in this research area is the simultaneous investigation of the three integrated speaking tasks on a real group of prospective TOEFL test-takers. In other words, previous studies have mostly taken one of the integrative patterns (esp., read-listen-speak) and under conditions dissimilar to operational TOEFL iBT. Another impetus that encouraged carrying out the present investigation was the scarcity, if any, of such studies in the Iranian context.

2.1. Measurement of Oral Language Performance

Second language research on speaking measurement reveals that some discourse features such as complexity, accuracy, and fluency (CAF) have frequently been used to quantify L2ers' oral productions (Elder & Iwashita, 2005; Ellis, 2009; Lambert & Kormos, 2014; Leaper & Brawn, 2018; Li et al., 2014; Mehnert, 1998; Nitta & Nakatsuhara, 2014; Tavakoli & Skehan, 2005;

Wigglesworth & Elder, 2010; Yan et al., 2020; Yuan & Ellis, 2003). Although with some variations, majority of the investigations in this area have resorted to a number of common approaches to access fine-grained measures representing respondents' speaking samples. The following is a summary of the commonly-used methods to compute the CAF measures.

- *Syntactic Complexity*: Measured as the ratio of the number of clauses to the total number of Communication (C-) Units (Elder & Iwashita, 2005; Ellis, 2009; Lambert & Kormos, 2014). Likewise, syntactic complexity has been operationalized as the number of clauses per Analysis of Speech (AS-) Unit (Leaper & Brawn, 2018; Li et al., 2014; Nitta & Nakatsuhara, 2014; Tavakoli & Skehan, 2005).
- *Lexical Complexity*: Mostly computed in terms of lexical-diversity indices such as measures of textual-lexical diversity (MTLD) computed by software like Coh-Metrix (Li et al., 2014; Nitta & Nakatsuhara, 2014; Yan et al., 2020).
- *Lexico-Grammatical Accuracy*: Calculated by counting the number of errors of different types per 100 (Mehnert, 1998; Nitta & Nakatsuhara, 2014) and also by the percentage of error-free clauses per all produced clauses (Leaper & Brawn, 2018; Wigglesworth & Elder, 2010; Yuan & Ellis, 2003).
- *Fluency*: Mainly measured in terms of speech rate as the number of articulated syllables per minute (Ellis, 2009; Lambert & Kormos, 2014; Leaper & Brawn, 2018; Li et al., 2014; Tavakoli & Skehan, 2005). Speech fluency has similarly been quantified based on the mean duration of pauses (Tavakoli & Skehan, 2005; Wigglesworth & Elder, 2010) as well as such repair phenomena as repetitions and revisions (Elder & Iwashita, 2005; Ellis, 2009; Lambert & Kormos, 2014; Leaper & Brawn, 2018; Li et al., 2014).

The main motivation to carry out the current study was to explore how TOEFL iBT integrated speaking tasks examine one's speaking ability. Particularly, we aimed to investigate the degree to which the test-takers' oral language

production could be subject to variations rooted in the prompting texts. To shed light on the mentioned curiosity, the following questions were addressed in the present study.

1. Do EFL test-takers complete TOEFL iBT integrated speaking tasks significantly differently in terms of complexity?
2. Do EFL test-takers complete TOEFL iBT integrated speaking tasks significantly differently in terms of accuracy?
3. Do EFL test-takers complete TOEFL iBT integrated speaking tasks significantly differently in terms of fluency?

3. Method

3.1. Participants and Data Collection

To collect the required data, a simulated TOEFL iBT sampled from the past administrations of the actual test was run at three official TOEFL centers in Iran, where around 80 iBT candidates participated. The trial test was exactly a copy of the real exam, both for the content and test rubrics. In fact, it was a part of the prospective iBT test-takers' preparatory program administered nearly ten days before their scheduled main exam. The oral language samples elicited from the participants were first carefully transcribed and, then, some were discarded due to an incomplete response to one of the tasks. Finally, 56 test-takers (28 female & 28 male) remained as the participants of the study.

According to Test and Score Data Summary (2020) published by ETS, the total mean score of all TOEFL test-takers throughout the preceding year was 83. However, the mean of the graduate-level students (similar to our participants) during the same year was 86-87. Therefore, we selected the participants who had received overall scores between 80 and 95 in order to make a representative sample. Moreover, all of the selected participants were Persian native speakers at the graduate level (both MA & PhD) from different university majors, mostly engineering and sciences, in Iran. Although the study constituted a within-group design, the scores assigned to the participants' speaking by the

institute that administered the trial TOEFL were compared with their scores given by ETS following their main exam. This comparison was aimed at assuring the comparability of the participants' performance in the trial and real TOEFL exam. Fortunately, we found that the two mean scores were not significantly different (both around 24 out of 30). Furthermore, the reliability of the oral responses elicited by the simulated TOEFL turned out to be preferably high (Cronbach's Alpha: 0.808).

3.2. Data Quantification

All oral responses (224 samples) delivered to the three integrated speaking tasks were meticulously transcribed, during which the mispronunciations and stress mispositions were detected. In the next stage, the transcripts were measured in terms of the quality criteria including complexity, accuracy, and fluency (Ellis & Barkhuizen, 2005; Ellis, 2008; Skehan, 2009). For complexity, we used the online version of Coh-Metrix 3 (Graesser et al., 2004; McNamara & Graesser, 2012). The online tool provides a table consisting of 106 measures, which represent various features of the inserted text. Following the guidelines provided in *Automated Evaluation of Text and Discourse with Coh-Metrix* (McNamara et al., 2014), four complexity measures were selected.

Latent Semantic Analysis (LSA): It is a co-reference measure that provides semantic overlap between sentences within a paragraph or paragraphs within a longer text. Because each transcript consisted of only one paragraph, we selected LSASSp that represents the mean overlap among all sentences within a paragraph.

Measure of Textual Lexical Diversity (MTLD): It measures the diversity of unique words (both content & function) occurring in a text in comparison to the total number of words in that text (McNamara et al., 2014). MTLD was selected since, firstly, it takes into account all words, and secondly, it is not dependent on text length.

Syntactic complexity as the mean number of words before main verb (SYNLE): It refers to 'left embeddedness' and is believed to measure complexity because when the mean number of words before the main verb increases, the complexity of the text increases too (McNamara et al., 2014).

Syntactic complexity as the mean number of modifiers per noun phrase (SYNNP): Like the previous measure, SYNNP is also expected to signify complexity because the higher the density of NPs, the higher the level of complexity (McNamara et al., 2014).

As for the next quality criterion, i.e., accuracy, each transcript was reviewed for any traces of ill-formedness, including grammatical, lexical, discourse-based, and pronunciation-related deviant forms. It was mentioned earlier that the pronunciation errors were specified during the transcribing phase. For grammatical errors, different problems related to articles, word order, tenses, pluralization, etc., were considered. Likewise, lexical errors of different types such as the stimulus texts' keywords misunderstood by the respondents (e.g., '*finalogical*' instead of '*phonological*') and basically ill-formed words (e.g., *renewated*) were detected. Additionally, those errors related to inappropriate cohesive ties (e.g., '*however*' used for denoting some result) were tallied. Finally, having been inspired by Ellis and Barkhuizen (2005), we used the following formula to compute the linguistic accuracy.

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

To measure the accuracy of each transcript, therefore, the ratio of all errors to all produced words was calculated in the unit of 100 to offset the possible effect of text length. Then, the amount of 'inaccuracy' (what the above square brackets yield) was subtracted from 100.

The three integrated speaking tasks were further measured in terms of 'content accuracy' (Frost et al., 2020) since test-takers have to transfer the main ideas presented by the prompting texts. These tasks usually involve a similar question that instructs the respondents to express the two lines of ideas,

examples, ways, etc., set forth by the aural prompt. Therefore, the collected responses were double reviewed (inter-rater reliability: 0.91) to quantify their degree of content accuracy. To this end, some criteria such as the number of key ideas transferred from the stimulus texts and the robustness of the oral summary or reproduction made by the respondents in each task were taken into account (Frost et al., 2011).

Finally, to measure fluency, the ratio of produced syllables per minute (Ellis, 2008) was calculated for each response. It should be clarified that there are basically two lines of analysis to measure speech fluency: a) Based on temporal aspects similar to what we applied and b) Based on the repair phenomena (Ellis, 2009; Yan et al., 2020). In the case of the latter, the words/phrases that had been successively repeated by a test-taker were deleted from the corresponding transcript, but the reformulations and revisions were sustained. In the next step, a free online tool known as SYLLABLE COUNTER (available at 'syllablecounter.org') was exploited in order to access the number of syllables produced by each participant when responding to each task. The following simple formula was then conducted to compute the fluency magnitude of each response.

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

3.3. Data Analysis

The quantified data was inserted into IBM SPSS Statistics (26), and One-Way Repeated-Measures ANOVA (Bachman, 2004; Pallant, 2020) was conducted seven times, each for one of the CAF subcategories. This route of analysis was chosen since the present study was aimed at comparing the complexity, accuracy, and fluency of the oral language produced by a group of TOEFL iBT test-takers who all attempted the three integrated speaking tasks.

4. Results

4.1. Analysis of the Three Speaking Tasks across the Complexity Measures

As an aspect of complexity, the measure of textual lexical diversity or MTLD was taken as a baseline to compare the speaking tasks. Table 1 summarizes the mean and standard deviation values computed for the three integrated tasks in terms of lexical diversity. As can be seen in Table 1, the highest mean score (59.62) was obtained for speaking task 3.

Table 1. *Descriptive Statistics of the Three Tasks for Lexical Diversity*

Speaking Tasks	Mean	Std. Deviation	N
Task2_Lexical Diversity	53.4132	14.28318	56
Task3_Lexical Diversity	59.6205	20.17523	56
Task4_Lexical Diversity	47.0859	12.10784	56

For the lexical diversity measure, there was found a significant difference among the three tasks, Wilks' Lambda = 0.71, F (2, 54) = 10.96, p = 0.000, partial eta squared = 0.28. Also, the pairwise comparisons revealed that while tasks 4 and 2, as well as 4 and 3, were significantly different in terms of lexical diversity, tasks 3 and 2 did not show any significant difference in that relation.

Table 2. *Pairwise Comparisons among the Tasks in Terms of Lexical Diversity*

(I) Lexical_Diversity	(J) Lexical_Diversity	Mean Dif. (I-J)	Std. Error	Sig. [b]	95% Con. Interval Lower	Upper
2	3	-6.207	2.951	.120	-13.495	1.080
	4	6.327*	2.266	.022	.732	11.923
3	2	6.207	2.951	.120	-1.080	13.495
	4	12.535*	2.775	.000	5.683	19.386
4	2	-6.327*	2.266	.022	-11.923	-.732
	3	-12.535*	2.775	.000	-19.386	-5.683

The three integrated speaking tasks were further compared based on the mean number of words before the main verb or left embeddedness (SYNLE). Table 3 summarizes the descriptive data for this complexity measure.

Table 3. *Descriptive Statistics of the Three Tasks for Left-Embeddedness*

Speaking Tasks	Mean	Std. Deviation	N
Task2_Left Embeddedness	4.3202	2.35365	56
Task3_Left Embeddedness	3.8889	2.22009	56
Task4_Left Embeddedness	4.8230	2.01905	56

Multivariate Tests showed there was no significant difference in terms of left-embeddedness among the three tasks, Wilks' Lambda = 0.92, F (2, 54) = 2.22, p = 0.118. In the next phase, the three tasks were compared regarding the mean number of modifiers per noun phrase (SYNNP) as a measure of syntactic complexity. The descriptive information in Table 4 shows the participants' oral performance in terms of SYNNP.

Table 4. *Descriptive Statistics of the Three Tasks for Modifiers per NP*

Speaking Tasks	Mean	Std. Deviation	N
Task2_Modifiers per NP	.8002	.19449	56
Task3_Modifiers per NP	.7455	.15505	56
Task4_Modifiers per NP	.6371	.18557	56

As Table 4 displays, the three speaking tasks were found to be significantly different from each other based on the mean number of modifiers per NP. In other words, SYNNP turned out to be significant, Wilks' Lambda = 0.65, F (2, 54) = 14.14, p = 0.000, partial eta squared = 0.34.

Based on the observed pairwise comparisons (Table 5), the differences between tasks 4 and 2 and also 4 and 3 were significant in terms of the SYNNP measure.

Conversely, tasks 2 and 3 were not significantly different from one another in this regard.

Table 5. *Pairwise Comparisons among the Tasks in Terms of Modifiers per NP*

(I) Modifiers_NP	(J) Modifiers_NP	Mean Dif. (I-J)	Std. Error	Sig.[b]	95% Con. Interval Lower	Upper
2	3	.055	.034	.340	-.029	.138
	4	.163*	.032	.000	.084	.242
3	2	-.055	.034	.340	-.138	.029
	4	.108*	.030	.002	.035	.182
4	2	-.163*	.032	.000	-.242	-.084
	3	-.108*	.030	.002	-.182	-.035

The next stage in the analysis of the three speaking tasks in terms of complexity pertained to the LSA measure. The three mean values and the corresponding standard deviations are depicted in Table 6 below.

Table 6. *Descriptive Statistics of the Three Tasks in Terms of LSA*

Speaking Tasks	Mean	Std. Deviation	N
Task2_Latent Semantic Analysis	.1829	.07967	56
Task3_Latent Semantic Analysis	.2027	.08648	56
Task4_Latent Semantic Analysis	.2477	.09929	56

As a subcategory of complexity, LSA was found to show a significant difference in distinguishing the three tasks, Wilks' Lambda = 0.79, F (2, 54) = 6.98, p = 0.002, partial eta squared = 0.20.

Once again, the pairwise comparisons in terms of SLA (Table 7) between tasks 4 and 3 as well as 4 and 2 indicated significant differences, whereas the difference between tasks 2 and 3 did not reach the significance level.

Table 7. *Pairwise Comparisons among the Three Tasks in Terms of LSA*

(I) Latent_Semantic	(J) Latent_Semantic	Mean Dif. (I-J)	Std. Error	Sig. [b]	95% Con. Interval	
					Lower	Upper
2	3	-.020	.017	.773	-.063	.023
	4	-.065*	.018	.002	-.109	-.020
3	2	.020	.017	.773	-.023	.063
	4	-.045*	.016	.022	-.085	-.005
4	2	.065*	.018	.002	.020	.109
	3	.045*	.016	.022	.005	.085

4.2. Analysis of the Three Speaking Tasks across the Accuracy Measures

Two dimensions of accuracy, form- and content-based, were addressed to compare the produced responses to the three integrated speaking tasks. Table 8 presents the descriptive statistics of the analyzed data in terms of linguistic accuracy.

Table 8. *Descriptive Statistics of the Three Tasks for Accuracy*

Speaking Tasks	Mean	Std. Deviation	N
Task_2_Accuracy	89.9786	4.60958	56
Task_3_Accuracy	87.3525	4.38884	56
Task_4_Accuracy	88.5129	4.92451	56

The observed results showed that form accuracy was significant in distinguishing the test-takers' oral language performance, Wilks' Lambda = 0.72, F (2, 54) = 10.19, p = 0.000, partial eta squared = 0.27. Also, the pairwise comparisons in terms of accuracy (Table 9) revealed only tasks 2 and 3 were significantly different.

Table 9. *Pairwise Comparisons among the Tasks in Terms of Accuracy*

(I) Accuracy	(J) Accuracy	Mean Dif. (I-J)	Std. Error	Sig.[b]	95% Con. Interval Lower	Upper
2	3	2.626*	.578	.000	1.199	4.053
	4	1.466	.620	.065	-.065	2.997
3	2	-2.626*	.578	.000	-4.053	-1.199
	4	-1.160	.626	.207	-2.706	.385
4	2	-1.466	.620	.065	-2.997	.065
	3	1.160	.626	.207	-.385	2.706

The oral responses to the three speaking tasks were further compared on the basis of content accuracy. Similar to what was obtained for the formed-based accuracy, the highest mean of content accuracy belonged to speaking task 2. The relevant descriptive statistics are displayed in Table 10.

Table 10. *Descriptive Statistics of the Three Tasks for Content Accuracy*

Speaking Tasks	Mean	Std. Deviation	N
Task_2_Content Accuracy	59.4643	21.54654	56
Task_3_Content Accuracy	49.1071	20.82628	56
Task_4_Content Accuracy	57.2321	17.86107	56

The oral responses to the three tasks were found significantly different in terms of content accuracy, Wilks' Lambda = 0.79, F (2, 54) = 6.97, p = 0.002, partial eta squared = 0.20. The pairwise comparisons (Table 11) indicated the differences between tasks 3 and 4 as well as 3 and 2 were significant. In contrast, responses to tasks 2 and 4 were not significantly different in terms of content accuracy.

Table 11. *Pairwise Comparisons among the Tasks in Terms of Content Accuracy*

(I) Content_Accuracy	(J) Content_Accuracy	Mean Dif. (I-J)	Std. Error	Sig.[b]	95% Con. Interval	
					Lower	Upper
2	3	10.357*	2.985	.003	2.986	17.728
	4	2.232	2.477	1.000	-3.884	8.349
3	2	-10.357*	2.985	.003	-17.728	-2.986
	4	-8.125*	2.486	.006	-14.263	-1.987
4	2	-2.232	2.477	1.000	-8.349	3.884
	3	8.125*	2.486	.006	1.987	14.263

4.3. Analysis of the Three Speaking Tasks Based on Fluency

Measured as the number of produced syllables per minute, the fluency of the participants' oral performance when speaking to the three integrated tasks was compared. The descriptive statistics in terms of fluency are presented in Table 12 below.

Table 12. *Descriptive Statistics of the Three Tasks for Fluency*

Speaking Tasks	Mean	Std. Deviation	N
Task_2_Fluency	164.0714	34.34946	56
Task_3_Fluency	145.7143	34.52520	56
Task_4_Fluency	155.2857	29.60818	56

Multivariate Tests showed a significant difference in terms of the fluency measure across the three tasks, Wilks' Lambda = 0.58, F (2, 54) = 19.11, p = 0.000, partial eta squared = 0.41. Additionally, the pairwise comparisons among the three tasks were all found to be significant.

Table 13. *Pairwise Comparisons among the Tasks in Terms of Fluency*

(I) Fluency	(J) Fluency	Mean Dif. (I-J)	Std. Error	Sig.[b]	95% Con. Interval Lower	Upper
2	3	18.357*	3.001	.000	10.947	25.767
	4	8.786*	3.313	.031	.604	16.967
3	2	-18.357*	3.001	.000	-25.767	-10.947
	4	-9.571*	2.855	.004	-16.622	-2.521
4	2	-8.786*	3.313	.031	-16.967	-.604
	3	9.571*	2.855	.004	2.521	16.622

5. Discussion

5.1. Comparison among TOEFL iBT integrated speaking tasks in terms of complexity

Among the selected complexity indices, two measures are directly related to syntactic complexity (SYNLE & SYNNP), one pertains to lexical diversity (MTLD), and the last one (LSA) falls at the heart of semantic unity. An important point associated with the last two measures is the fact that they can be in negative correlation (McNamara et al., 2014). Given this, it makes sense when the lexical diversity of a text increases, its LSA index is likely to decrease.

The obtained results in the present study revealed that lexical diversity had a significant difference in distinguishing tasks 2 and 3 from task 4. This was the case because task 4 showed the lowest mean score among the three tasks. Despite the fact that task 3 had the highest mean score in that relation, the difference between tasks 2 and 3 was not found to be significant. The result observed for lexical diversity is probably rooted in the amount of provided task input, which is richer in tasks 2 and 3 (preceded by both textual and aural inputs)

than in task 4, being prompted by an aural input only. This claim could validate Cumming et al. (2005) and also Lee (2005), who concluded that any deficiency in each of the prompting skills could interfere with the subsequent speaking performance. More specifically, what the current study concluded about lexical diversity supports Crossley & Kim (2019), who reported a significant effect for the source text's keywords on the quality of test-takers' speaking performance.

An important finding of this study was the fact that no significant difference was found in terms of left-embeddedness among the three speaking tasks. A potential argumentation for this finding might stem from the nature of the three integrated tasks, being mainly different based on the topic and amount of content presented to respondents. Such variations seem to be more lexical and concept-driven than syntactic in nature. Contrary to this claim, the mean number of modifiers per NP, which also accounts for complexity, left a significant difference to differentiate tasks 2 and 3 from task 4. The finding is in full agreement with what was observed about lexical diversity. As argued for lexical diversity, it is possible to relate the lower mean of modifiers per NP in the responses to task 4 to the comparatively lower amount of input provided in this task. This argumentation can be tenable since the textual input in tasks 2 and 3 possibly enriches the respondents with more ideas to enhance their noun phrases with premodification. This conjecture can be substantiated by referring to Crossley and Kim (2019), where they asserted the linguistic elements of the source texts in general, and their frequency of lexical-propositional elements in specific, strongly affect the following oral output. In the meantime, if test-takers fail to get the input ideas from the aural prompt, what may have happened in task 4, their following oral production will certainly decrease in quality (Cumming et al., 2005; Frost et al., 2020; Lee, 2005).

The remaining complexity measure, i.e., LSA, turned out to be significant in distinguishing task 4 from both tasks 2 and 3, which themselves were not significantly different. Concerning LSA, however, task 4 recorded the highest mean value. It was already mentioned at the outset of this part that lexical diversity and LSA can be in a negative correlation. The claimed negative

correlation was fully validated in the case of task 4, showing the lowest lexical diversity as well as the highest latent semantic overlap or LSA.

A concluding result about the four subcategories of complexity is that the responses to speaking task 4 were mainly different from those to tasks 2 and 3. Further, task 4 showed the highest mean score in terms of the left-embeddedness measure. Although this measure did not show any significant difference among the tasks, the fact that the responses to task 4 had the highest mean in terms of left-embeddedness might be originated in the less input provided in this task, which in turn directs test-takers to focus more on their syntactic elaboration. The recent claim seems to be in line with the trade-off effects (Ellis, 2009; Mehnert, 1998; Skehan, 2014; Yuan & Ellis, 2003), which make L2 learners prioritize some aspects of their performance, especially under testing constraints.

5.2. Comparison among TOEFL iBT integrated speaking tasks in terms of accuracy

The results revealed that both form- and content-based accuracy left significant differences among the tasks under investigation. However, the mean difference in terms of form accuracy only between tasks 2 and 3 was significant. The same two tasks were also significantly different in terms of content accuracy. Similarly, the difference between tasks 3 and 4 was significant in terms of the accuracy of content. It needs to be remarked that TOEFL iBT demands test-takers to speak on campus-related issues in task 2, whereas they should speak on scientific/academic topics in task 3. This variation might be the source of the observed difference between the two mentioned tasks based on accuracy. In this direction, Huang et al. (2016) and Huang et al. (2018) confirmed the influential role played by topic to distinguish the quality of the oral language produced in integrated speaking tasks. In the same way, we can conclude that as they are normally more familiar with campus-related topics than those scientific ones, the respondents could pay closer attention to the accuracy of their speech in task

2 than in task 3. In fact, task 2 showed the highest mean among the three tasks when accuracy measures were concerned.

Regarding the degree of content accuracy, we found the obtained mean for task 3 was significantly lower than those for tasks 2 and 4. The difference in terms of accuracy between tasks 2 and 3 was already explained to be possibly rooted in the two different speaking contexts (campus-related vs. academic). This reasoning can potentially corroborate Frost et al. (2019), suggesting that the content accuracy of the oral discourse produced in integrated speaking tasks correlates with respondents' success in making sense of the task input. Despite the similarity between tasks 3 and 4 in consisting of academic topics, the respondents showed a significantly lower content accuracy when speaking to task 3 than task 4. This is while the former has a richer task input by presenting the test-takers with both textual and aural texts. The most tenable justification for this seeming paradox may stem from the specificity of the topic presented in task 3 in our simulated test. In fact, the participants had to speak on a topic in biology in task 3, whereas they were expected to speak on a history-related topic in task 4. Consequently, the higher degree of topic specificity of biology may have contributed to lower content accuracy of the oral responses to task 3. This conclusion is certainly in line with similar results reported by Huang et al. (2016), Huang et al. (2018), Frost et al. (2019), and Crossley & Kim (2019) over the effective role of topic and content of the prompting texts in distinguishing the integrated speaking tasks.

5.3. Comparison among TOEFL iBT integrated speaking tasks in terms of fluency

Although speech fluency has several dimensions, including temporal aspects and those related to the repair phenomena (Ellis, 2009; Yan et al., 2020), we focused on the number of syllables produced per minute. It needs to be reminded that the repair phenomena and the related subcategories were also considered

during the transcribing stage. Yet, those aspects related to the number and length of pauses were not considered in this study.

The analyses demonstrated that fluency made a significant difference across the three speaking tasks. The highest mean was observed for the oral responses to task 2, while task 3 was found to have caused the lowest level of fluency. Two points seem to be noteworthy based on the results obtained for the fluency measure. First, the participants may have experienced more convenience when speaking on the campus-driven topic in task 2, entailing more informal as well as commonplace ideas, which in turn, could trigger higher speech fluency. This impression is reinforced by Fulcher and Reiter (2003), reporting that topic familiarity and fluency are positively correlated. Second, the mentioned trade-off effects (Ellis, 2009; Mehnert, 1998; Skehan, 2014; Yuan & Ellis, 2003) may have caused the participants to experience the lowest fluency during task 3. In other words, because task 3 involved a formal and scientific context, and the test-takers were given both textual and aural input, more attention might have been paid by them to transfer the ideas presented by the task input. A proof to support this conjecture is the highest lexical diversity, as a complexity measure, observed for task 3 among the three speaking tasks. Relying on the prompting texts, the respondents are thought to have reached more complexity at the expense of fluency.

As the final remark, task 2 was found to have triggered the highest mean score for both accuracy and fluency, while task 3 showed the lowest mean scores with respect to these measures. These facts could imply speech accuracy and fluency are in agreement. This recent claim, however, is in contrast to Yuan and Ellis (2003), where they reported on some competition between accuracy and fluency.

6. Conclusion

The current study was an attempt to compare the three integrated tasks of the TOEFL iBT speaking module on the basis of the oral language produced by a group of prospective iBT candidates in a simulated TOEFL test. The following is a summary of the main results.

- Regarding complexity, three measures including lexical diversity, the mean number of modifiers per NP, and LSA revealed that tasks 2 and 3 were significantly different from task 4. Across these measures, tasks 2 and 3 did not show any significant difference.
- Left-embeddedness, which falls at the heart of syntactic complexity, did not show any significant difference among the three speaking tasks.
- Speaking tasks 2 and 3 revealed significant differences concerning both form- and content-based accuracy.
- In terms of speech fluency, all pairwise comparisons among the three speaking tasks were found to be significantly different.

The results may lend support to the influential role of prompting texts in TOEFL iBT integrated speaking tasks. The rationale to validate the claim relates to the fact that the two tasks enriched by both textual and aural prompts (tasks 2 & 3) triggered more comparable oral outputs than task 4, which includes an aural stimulus only. In the meantime, the oral responses elicited by tasks 2 and 3 showed different levels of accuracy, which could be attributed to the context of speaking in each (campus-related vs. academic). Given these conclusions, there seems to exist some sort of competition between complexity and accuracy when it comes to integrated speaking tasks in L2 assessment. The concluding results imply some facts as to how skills integration in assessment may contribute to fluctuations in test-takers' speaking performance. The effects could be more or less beneficial to respondents depending on their capabilities to decode the prompting texts. Thus, different stakeholders including test-preparation trainers and trainees, as well as test constructors, could take advantage of the findings obtained on integrated speaking assessment.

With regard to the potential limitations of the present study, it should be acknowledged that we employed a group of TOEFL iBT test-takers who coupled roughly around the TOEFL iBT's mean score. It would be a more inclusive exploration of the tasks in focus if more test-takers at different proficiency levels were studied. Moreover, exploring a variety of topics in such

integrated speaking tasks could yield more precise results as to how the input texts function. Similarly, the observed results in terms of fluency demand further research to meticulously delve into the possible reasons why the iBT test-takers produced oral responses with varying levels of fluency when speaking to the integrated tasks. The amount and topic of the stimulus texts in such integrated speaking tasks can be possible sources of variation in speech fluency.

7. References

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

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.

Brooks, L. & Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBT and real-life academic speaking activities. *Language Assessment Quarterly*, *11*, 353-373. https://doi.org/10.1080/15434303.2014.947532

Brown, A., & Ducasse, A. M. (2019) An equal challenge? Comparing TOEFL iBT speaking tasks with academic speaking tasks. *Language Assessment Quarterly,* *16*(2), 253-270. https://doi.org/10.1080/15434303.2019.1628240

Brown, A., McNamara, T., Iwashita, N., & O'Hagan, S. (2001). *Investigating raters' orientations in specific-purpose task-based oral assessment*. TOEFL 2000 Research and Development project report. Educational Testing Service. Unpublished manuscript.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices* (2nd ed.). Pearson Education Inc.

Brown, H. D., & Abeywickrama, P. (2018). *Language assessment: Principles and classroom practices* (3rd ed.). Pearson Education ESL.

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

Crossley, S. A., & Kim, Y. J. (2019). Text integration and speaking proficiency: Linguistic, individual differences, and strategy use considerations. *Language Assessment Quarterly, 16*(2), 217-235. https://doi.org/10.1080/15434303.2019.1628239

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2005). *A teacher-verification study of speaking and writing prototype tasks for a new TOEFL*. TOEFL Monograph Series 26. Educational Testing Service. https://www.ets.org/Media/Research/pdf/RM-04-05.pdf

Douglas, D. (1997). *Testing speaking ability in academic contexts: Theoretical considerations*. TOEFL Monograph Series 8. Princeton, NJ: Educational Testing Service. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.603.3457

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. (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.

Farnsworth, T. (2013). An investigation into the validity of the TOEFL iBT speaking test for international teaching assistant certification. *Language*

Assessment Quarterly, *10*(3), 274-291. https://doi.org/10.1080/15434303. 2013.769548

Frost, K., Clothier, J., Huisman, A., & Wigglesworth, G. (2019). Responding to a TOEFL iBT integrated speaking task: Mapping task demands and test takers' use of stimulus content. *Language Testing*, 1-23. https://doi.org/10.1177/0265532219860750

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. (2020). Relationships between comprehension, strategic behaviors and content-related aspects of test performances in integrated speaking tasks. *Language Assessment Quarterly*. https://doi:10.1080/15434303.2020. 1835918

Fulcher, G. (2003). *Testing second language speaking*. Pearson Education Limited.

Fulcher, G., & Reiter, R. M. (2003). Task difficulty in speaking tests. *Language Testing, 20*(3), 321-344. https://doi.org/10.1191/0265532203lt259oa

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

Huang, H. T. D., Hung, S. T. A., & Hong, H. T. V. (2016). Test-taker characteristics and integrated speaking test performance: A path-analytic study. *Language Assessment Quarterly, 13*(4), 283-301. https://doi.org/10.1080/15434303.2016.1236111

Huang, H. T. D., Hung, S. T. A., & Plakans, L. (2018). Topical knowledge in L2 speaking assessment: Comparing independent and integrated

speaking test tasks. *Language Testing* *35*(1), 27-49. https://doi.org/10.1177/0265532216677106

Kormos, J., Brunfaut, T., & Michel, M. (2020). Motivational factors in computer-administered integrated skills tasks: A study of young learners. *Language Assessment Quarterly, 17*(1), 43-59. https://doi.org/10.1080/15434303. 2019.1664551

Kyle, K., Crossley, S. A., & McNamara, D. S. (2015). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, 1-21. https://doi.org/10.1177/0265532215587391

Lambert, C., & Kormos, J. (2014). Complexity, accuracy, and fluency in task-based L2 research: Toward more developmentally based measures of second language acquisition. *Applied Linguistics, 35*(5), 607-614. https://doi.org/10.1093/ applin/amu047

Leaper, D. A., & Brawn, J. R. (2018). Detecting development of speaking proficiency with a group oral test: A quantitative analysis. *Language Testing, 36*(2), 181-206. https://doi.org/10.1177/0265532218779626

Lee, Y.-W. (2005). *Dependability of scores for a new ESL speaking test: Evaluating prototype tasks*. TOEFL Monograph Series 28. Princeton, NJ: Educational Testing Service. https://www.ets.org/Media/Research/pdf/RM-04-07.pdf

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

Luoma, S. (2004). *Assessing Speaking*. Cambridge University Press.

McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. MaCarthy & 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

Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2014). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing, 32*(1), 39-62. https://doi.org/10.1177/0265532214 538014

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. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510-532. https://doi.org/10.1093/ applin/amp047

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

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

Taylor, L. (2011). *Examining speaking: Research and practice in assessing second language speaking*, Studies in Language Testing (vol. 30). Cambridge University Press.

TOEFL iBT: Test and score data summary 2019. (2020). Educational Testing Service.

Weir, C. J. (1990). *Communicative language testing*. Hemel Hempstead: Prentice Hall.

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

Wigglesworth, G., & Frost, K. (2017). 'Task and performance-based assessment' in: Shohamy, E., Or, L. G., & May, S. (Eds.), *Language testing and assessment* (3rd ed.). Springer International Publishing AG.

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, 00*(0), 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

Notes on Contributors:

Ali A. Ariamanesh is Ph.D. Student in TEFL at University of Isfahan, Iran. He has been studying and doing research at this university since 2017. His research areas cover second language testing and assessment, teaching methodology, SLA, and quantitative research.

Hossein Barati is associate Professor of applied linguistics at University of Isfahan, Iran. He received his Ph.D. degree from University of Bristol, England, in 2005. His research interests include language testing and assessment, teaching and learning language skills, teacher education, and program evaluation.

Manijeh Youhanaee is an associate Professor at University of Isfahan. She has co-authored "*A Descriptive Dictionary of Theories of Generative Grammar*" and has published a number of articles on learning different English syntactic properties by native speakers of Persian. Her areas of interest include acquisition of L2/L3 syntax and issues in teaching and learning English as a foreign language.