TOEFL iBT Iranian Test-Takers' Oral Language Performance: A Comparison between Independent and Integrated Speaking Tasks

Ali A. Ariamanesh
University of Isfahan, Isfahan, Iran
aa.ariamanesh@fgn.ui.ac.ir

Hossein Barati
University of Isfahan, Isfahan, Iran
barati@fgn.ui.ac.ir

Manijeh Youhanaee
University of Isfahan, Isfahan, Iran
youhanaee@fgn.ui.ac.ir

Bio-Profiles

Ali A. Ariamanesh is a PhD holder in applied linguistics. He received his doctorate from University of Isfahan, Iran. As an assistant professor, his research areas include second language assessment and evaluation, teaching language skills, and SLA.

Hossein Barati is associate professor of applied linguistics at University of Isfahan, Iran. He received his PhD from University of Bristol, England. Language testing and assessment, teacher education, and programme evaluation are amongst his areas of research.

Manijeh Youhanaee got her PhD from University of Essex, England. She is associate professor of applied linguistics at University of Isfahan, Iran. She has coauthored "*A Descriptive Dictionary of Theories of Generative Grammar*" and has published a number of articles on the acquisition of different English syntactic properties.

Abstract

The present study investigated the speaking module of TOEFL iBT with an emphasis on the dichotomy of independent and integrated tasks. The potential differences between the two speaking conditions were intended to be explored based on the oral performance elicited from a group of Iranian test takers. To collect the required data, a simulated version of TOEFL iBT characterized with all test rubrics as practiced by ETS was applied. A total number of 224 oral samples were transcribed and then quantified in terms of complexity and fluency by some internet-based programs including *Coh-Metrix 3* and *Syllable Counter*. Moreover, the accuracy of each oral performance was manually computed for comparison purposes. The results revealed that the two speaking contexts were significantly different in terms of linguistic accuracy as well as two complexity measures representing latent semantic analysis and the mean number of modifiers per noun phrase. However, the two speech-eliciting contexts did not trigger any significant difference in the participants' oral language performance in terms of fluency and two other complexity measures, namely lexical diversity and left-embeddedness. The findings may have some implications for all TOEFL iBT stakeholders concerning the speaking subtest.

Keywords: TOEFL speaking, Independent task, Integrated task, Complexity, Accuracy, Fluency

1. Introduction

Second language speaking assessment seems to be a challenging and complicated endeavor when it comes to integrative test tasks, measuring oral production along with other skills such as reading and listening. Fulcher (2003, p. 84) brings up integrative speaking tasks and contends that such assessment activities, unlike the independent ones, require test takers to reflect the prompting information in their oral performance. In this regard, Fulcher suggests although integrative test tasks are more real-world like, more investigation is required to ascertain how precisely L2 speaking is measured against the stimulus skills and memory capacity. By the same token, Brown (2004, p. 140) claims that the isolation of oral language production tasks from the aural comprehension ones is very difficult, and also inauthentic, since the two oral skills are closely inter-related. Nevertheless, Brown remarks such integration may challenge the accuracy and precision of speaking assessment if the test designer fails to clearly specify those influential factors triggered by the aural input. Similarly, Luoma (2004, p. 43) distinguishes between stand-alone (the term she uses for independent) and integrated tests to assess the speaking skill. From her vantage point, the stand-alone test tasks are those which focus only on oral interaction or production. The integrated tasks, however, are those comprising a combination of other skills, say listening and reading, with speaking. Luoma further spotlights test authenticity as the primary motivation behind skills integration. Yet, she asserts that integrated speaking tasks may increase the cognitive load imposed on test takers, which can consequently decrease their quality of oral language performance.

Independent and integrated test tasks are also addressed by Carr (2011), where he defines the latter as the tasks in which test takers need to process additional information in order to complete a task. Moreover, Carr (2011, p. 99) implies that integrated test tasks need to reflect the target language use (TLU) domain for which test takers are being tested. Thus, such tasks must be sampled from the most typical language-mediated activities the test takers are very likely to encounter in the TLU domain. According to Carr, therefore, integrated speaking tasks are more communicative and real-life like in nature. Similarly, Wigglesworth and Frost (2017) contend that recent language tests have moved from independent tasks, addressing individual language skills, toward integrated ones in which skills are combined for authenticity and validity purposes. This tendency to skills integration has necessarily increased the complexity of test tasks, which may put some added burden on L2 test takers (Shohamy et al., 2017, p. 121). Nonetheless, Wigglesworth and Frost claim that integrated tasks are more beneficial to examinees as they are much fairer compared to independent tasks. They believe integrated tasks mitigate biases related to topic familiarity by providing the required input through the prerequisite texts. Despite these advantages, Wigglesworth and Frost (2017) confess the impact of task input as well as the way test takers make use of prompting information are not yet fully known.

The status quo of L2 speaking assessment with independent and integrated tasks implies a number of drawbacks. For instance, TOEFL iBT speaking subtest includes both independent and integrated tasks to assess English speaking for academic purposes. Despite its popularity, few studies (Barkaoui et al., 2013; Biber & Gray, 2013; Kyle et al., 2015; Huang et al., 2018) have meticulously explored the iBT speaking module with an emphasis on the dichotomy of independent and integrated tasks. Therefore, the current study aimed to investigate how the nature of task may invoke changes in the quality of oral language produced by TOEFL iBT Iranian test takers. Essentially, we intended to unveil in what ways the independent speaking task and the integrated ones are differently measuring the underlying speaking construct. A short review of the TOEFL iBT speaking tasks already presented by ETS can be useful to illuminate the current study’s main purpose.

Task 1: An independent question that asks respondents to make a personal choice between two mostly contrasting ideas and then defend their choice by some detailed reasoning.

Task 2: An integrated task with a read-listen-speak sequence on a campus-based topic during which test takers are required to talk about the presented opinions in the listening prompt.

Task 3: An integrated task similar to the previous one in terms of skills-sequence but with an academic topic. This time, respondents need to transfer the gist of the lecturer's points.

Task 4: An integrated task comprising a listen-then-speak sequence on an academic topic. Here, test takers should summarize the lecturer's ideas, usually including two examples.

The present study, therefore, embarked on making a detailed comparison between the total integrated speaking tasks and the independent one with the hope of revealing whether the integrative nature of the former causes any significant differences in the test-takers' oral language performance. Far from the type and magnitude of task input, integrated speaking tasks are claimed to be more beneficial to test takers as such test tasks provide the required content to speak about (Wigglesworth & Frost, 2017). This advantage is claimed to mitigate topic-related biases, which may be experienced while attempting independent speaking tasks. Given this, the test-takers' oral language productions are supposed to show more quality in iBT integrated speaking tasks than the independent one. However, if the integrated tasks are thought to be muddied (Weir, 1990, 1993, 2005), this conjecture may also hold true that the test takers are likely to feel freer when speaking to the independent task rather than the integrated ones. This is because the latter could run the risk of muddied measurement due to skills integration.

2. Background to the Study

Second language test takers may fail to well understand the prerequisite prompting texts, especially the aural ones, based on which they are expected to speak through integrated tasks. This uncertainty could raise the possibility of muddied measurement, which was first argued by Weir (1990). This criticism might be tenable to the degree to which the speaking ability is underestimated due to test-takers' deficiencies in other skills, explicitly reading and listening, while responding to such integrated tasks as presented in TOEFL iBT speaking subtest.

Concerned with the potential drawback, Brown, Iwashita, and McNamara (2005) concluded that the TOEFL integrated speaking tasks, especially the listening-speaking one, may expect some higher mental demand from the test takers. Likewise, Frost, Elder, and Wigglesworth (2011) studied an integrated listening-speaking task to examine how the stimulus content can influence the respondents' quality of performance. Their results lent clear support to the positive correlation between the test-takers' proficiency and how successfully they managed to take advantage of the task input. In other words, Frost et al. (2011) concluded the participants with different language abilities benefitted significantly differently from the task input when responding to an integrated speaking task. In another study, Barkaoui, Brooks, Swain, and Lapkin (2013) studied the TOEFL iBT independent and integrated speaking tasks, aiming to explore the strategies their participants used when coping with the mentioned tasks. Barkaoui et al. reported that their test takers had used more varied strategies when attempting the integrated tasks than in case of the independent ones. Consequently, they concluded as the number of skills involved in speaking test tasks increases, more strategies are required by test takers to create their intended oral output. Resorting to this conclusion, Barkaoui et al. (2013) validated the inclusion of integrated tasks for L2-speaking assessment.

Biber and Gray (2013) studied the TOEFL iBT independent and integrated speaking tasks and found that their participants at all proficiency groups varied their linguistic choices across the two task types, i.e., independent and integrated. They further argued since the independent tasks were more personal, the test takers used more informal style than when they were involved in the integrated tasks, which were normally more informational. Similarly, Kyle, Crossley, and McNamara (2015) set to compare the TOEFL iBT independent and integrated speaking tasks. Their study mainly focused on the lexical and cohesive features of the elicited oral responses. Overall, Kyle et al. (2015) found a partial difference between iBT independent and integrated speaking tasks in terms of some linguistic features observed in the participants' oral productions. In particular, the independent and integrated academic tasks triggered more distinct performances, which can be rooted in the scientific context of the latter. Through a similar exploration to compare independent and integrated speaking tasks, Huang, Hung, and Plakans (2018) studied four samples of TOEFL iBT integrated speaking tasks with a reading-listening-speaking pattern. Besides, four independent speaking tasks as well as some topical knowledge tests were administered. Ultimately, their study confirmed the significant role played by topical knowledge, particularly, in integrated tasks. In other words, those participants with richer topical knowledge on the stimulus texts benefited more from the content provided in the integrated tasks. This finding implies that topical knowledge can affect test-takers' speaking performance in integrated tasks more significantly than in independent counterparts.

Frost, Wigglesworth, and Clothier (2020) examined the effect of proficiency on L2 oral performance in a reading-listening-speaking task from TOEFL iBT. In particular, Frost et al. intended to explore in what ways the ideas presented by the prompting texts were reported by the examinees. By far, the participants across different proficiency groups were observed to have reproduced more 'single ideas' directly extracted from the stimulus texts than either 'summarized' or 'synthesized' information. They also reported those test takers with higher proficiency outperformed the lower ones in summarizing the input information. Besides, Frost et al. (2020) emphasized the fact that their low-proficiency participants experienced more comprehension challenges with the prompting texts. Certainly, such difficulties can prevent lower proficiency test takers from drawing on the concepts and ideas presented via the prerequisite texts. In a recent study, Roever and Ikeda (2021) concentrated on the TOEFL iBT speaking tasks to unveil the extent they may predict test-takers' interactional competence for authentic purposes. They observed a relatively high correlation between performance on the iBT speaking tasks and the interactional activities, consisting of extra-linguistic and strategic skills. Although Roever and Ikeda concluded that TOEFL iBT speaking subsection is generally a valid measure to provide information about the test-takers' interactional competence, they did not clarify how the independent and integrated tasks were different in that relation.

The previous studies on the TOEFL iBT speaking subtest, as reviewed above, have addressed a number of critical issues, including the bifurcation of independent and integrated tasks. Nevertheless, what seems to be lacking in this research area pertains to a paucity of information on the mentioned task types completed by a real group of TOEFL candidates, especially in the Iranian context. In addition, it seems the two speaking conditions have rarely been compared in terms of the complexity, accuracy, and fluency (CAF) triad.

3. Research Questions

The main impetus enforcing the completion of this study was to compare the iBT integrated speaking tasks with the independent one, detecting how each test-taker's speaking quality would be different across the two task types. Compromised with this aim, the following research questions (RQs) were addressed to expose the potential differences.

RQ1: Do EFL test takers perform significantly differently on TOEFL iBT independent and total integrated speaking tasks in terms of complexity?

RQ2: Do EFL test takers perform significantly differently on TOEFL iBT independent and total integrated speaking tasks in terms of accuracy?

RQ3: Do EFL test takers perform significantly differently on TOEFL iBT independent and total integrated speaking tasks in terms of fluency?

4. Methods

4.1. Participants and instrumentation

The required speaking samples were collected from 56 TOEFL iBT Iranian test takers in three official TOEFL centers located in three different cities in Iran. To assure the representativeness of the sample, we randomly selected the participants from among a variety of majors, different university levels, and both genders. The selected participants had been studying for around 6 months on average to get prepared for the TOEFL exam when they were given a simulated TOEFL iBT. The trial test was precisely comparable to the real exam in terms of both content and test-method facets (Bachman, 1990) and was administered to the participants a few days before their scheduled main exam. The deployed data-collection instrument comprised a copy of test inputs from past administrations of the operational TOEFL iBT given by ETS. Moreover, the environment of the computer-mediated trial test was fully identical to the real exam (see Table 1 for the reliability of the simulated test). As to their age range and majors, the recruited participants were between 22 to 35 years old (mean: 27.33), and all were either undergraduate or postgraduate university majoring students from humanities, sciences, medical care, and especially engineering. It should be mentioned that we decided not to control their specific fields of study for the sample to represent the Iranian population of TOEFL candidates.

As the trial test was the final step of their preparation process for the main exam, the participants were not informed in advance that their responses to the speaking part would be used for research purposes. Once it was administered, however, they were all informed that their oral productions were to be used for research purposes on the condition of anonymity if they agreed. At the outset of the study, the participants comprised up to more than 80 prospective TOEFL iBT candidates who took the simulated test and received scores mostly between 80 to 95 in that test. Regarding the process of scoring, it should be clarified that the responses to the reading and listening items were objectively machine-scored by the software. However, the speaking and writing responses were double-rated by two experienced local scorers who had been teaching and testing TOEFL iBT for several years. Furthermore, the inter-rater reliability of the scores assigned to the productive responses was very high (0.93). Finally, 28 female and 28 male participants were selected from the original group, while the remaining candidates were excluded due to some extreme scores or incomplete performance in one or more of the speaking tasks. Additionally, the mean of their overall scores in the trial test was close to the mean score announced by ETS throughout 2019 for graduate-level test takers (mean: 86-87, Educational Testing Service, 2020). Thus, the participants of the present study had scores around the TOEFL iBT mean score, which elucidates their global proficiency in English.

4.2. Data quantification procedures

A total of 224 speaking samples produced by 56 participants were meticulously transcribed with all false starts, repetitions, revisions, etc. Afterward, each oral sample was once more listened to and the mispronunciations including wrong stress positions were specified. Also, during the second phase, the mere repetitive words or phrases which were successively produced by a respondent were discarded. Still, the reformulations and self-corrections were kept in each transcript. Having completed the transcribing stage, we deployed the following approaches to quantify each sample in terms of the quality criteria known as CAF as the baseline to compare the tasks under exploration.

To measure complexity, the online webtool of Coh-Metrix 3 (Graesser et al., 2004; McNamara & Graesser, 2012) was exploited. Among the various measures the program computes, we applied four of them (Ariamanesh et al., 2020) for the intended purposes. These measures are explained by McNamara, Graesser, McCarthy, and Cai (2014) as follows.

Latent Semantic Analysis (*LSA*): It provides semantic overlap between sentences within a paragraph or a longer text. Coh-Metrix 3 delivers eight measures related to LSA. As each transcript consisted of only one paragraph, we selected the LSASSp measure that represents the mean overlap among all sentences within a paragraph.

Measure of Textual Lexical Diversity (*MTLD*): This measure indicates the diversity of unique words (content & function) to the total number of words in a text. There are four lexical diversity measures provided by the program, but MTLD was selected since it is not dependent on the length of text.

Syntactic Complexity: The mean number of words before the main verb, or Left-Embeddedness (*SYNLE*): McNamara et al. (2014) believe the complexity of a text positively correlates with the mean number of words before main verbs. Therefore, SYNLE can be considered as an indication of syntactic complexity.

Syntactic Complexity: The mean number of modifiers per noun phrase (*SYNNP*): Similar to the previous measure, SYNNP can show complexity as there must be a direct relationship between the mean length of noun phrases within a text and its complexity. In other words, the higher the density of NPs, the higher the complexity.

As for the accuracy index, the transcripts were analyzed for any erroneous forms, including the grammatical, lexical, discourse, and pronunciation deviations. Regarding the last category, we enlisted only the mispronunciations at the syllable and word level. Among the frequently detected errors were those related to word order, tenses, articles, plural forms, nonsense words, misunderstood words from the stimulus texts, and inappropriate discourse markers and cohesive ties. Following this stage, the ratio of detected errors to the whole words in each transcript (Ellis & Barkhuizen, 2005) was calculated to measure the degree of 'inaccuracy' in each sample. Hence, we applied the following formula to compute each transcript's accuracy.

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

To measure speech fluency, the number of produced syllables in the unit of minute was calculated for each speaking sample (Ellis, 2008). Basically, there are two approaches to measuring speech fluency: i) Based on temporal aspects similar to what was used in this study, and ii) Based on the repair phenomena (De Jong, 2018; Ellis, 2009; Yan et al., 2020). As to the latter approach, the words or phrases immediately repeated by the respondents were deleted from the transcripts, yet the reformulations that were attempted to revise their productions were sustained. Once the transcripts were refined, a free online tool available at syllablecounter.org was applied to calculate the number of syllables produced by each participant in each speaking task. The following formula was then conducted to compute the fluency level of each oral production (Ellis, 2009).

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

4.3. Analysis

To address the RQs formulated in the present causal-comparative investigation (Ary et al., 2019), the quantified data was inserted into IBM SPSS Statistics (26) and a series of Paired-Samples T-Test was conducted to compare the participants' oral responses to the independent speaking task with those to the three integrated ones. In this direction, the mean of each evaluative measure representing the CAF criteria across the three integrated speaking tasks was computed, and then it was compared with the corresponding measure standing for the independent task.

5. Results

Prior to conducting the main analyses, the quantified data collected by the simulated TOEFL iBT was checked for internal consistency. As Table 1 portrays, the reliability level was preferably high (Pallant, 2016).

Table 1: *Reliability Statistics of Four TOEFL iBT Speaking Tasks across Six CAF Measures*

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	Number of Items
0.808	0.697	24

The independent and three integrated speaking tasks were then compared based on six CAF measures, including complexity with four measures, accuracy with only one holistic measure, and fluency as speech rate. Having computed the mean of each CAF measure for the total integrated tasks, we conducted Paired-Samples T-Test (Pallant, 2016) three times to expose the possible differences between the two speaking conditions. The following subparts present the tabulated results obtained from SPSS outputs.

5.1. Comparison of the two speaking conditions: Complexity

The independent and integrated speaking tasks were compared in terms of four complexity measures including lexical diversity, left-embeddedness, mean number of modifiers per NP, and LSA. The descriptive statistics of the complexity measures for the two speaking situations are summarized in Table 2. Paired differences between the two speaking contexts revealed the mean number of modifiers per NP and LSA were significantly distinguishing the independent task from the total integrated ones. For the former measure, the integrated tasks had a higher mean score, whereas for the latter, the independent task showed a higher mean value. The pointed significant differences can be seen in Table 3.

Table 2: *Descriptive Statistics of Independent vs. Integrated Tasks in Terms of Complexity*

Task Type		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Task1_MTLD	**55.0570**	56	12.85743	1.71815
	Integrated_MTLD	**53.3732**	56	10.88626	1.45474
Pair 2	Task1_Left-Embeddedness	**4.0593**	56	2.02043	.26999
	Integrated_Left Embeddedness	**4.3440**	56	1.26033	.16842
Pair 3	Task1_Modifiers per NP	**.6680**	56	.19479	.02603
	Integrated_Modifiers per NP	**.7276**	56	.11398	.01523
Pair 4	Task1_LSA	**.2493**	56	.12000	.01604
	Integrated_LSA	**.2111**	56	.04868	.00651

Table 3: *Paired Differences bet. Independent and Integrated Tasks in Terms of Complexity*

Task Type	Mean	Std. Deviation	Std. Error Mean	t	df	Sig. (2-tailed)
Task1_MTLD Integrated_MTLD	1.68375	12.95274	1.73088	.973	55	**.335**
Task1_LE* Integrated_LE	-.28476	2.25652	.30154	-.944	55	**.349**
Task1_Modifiers_NP Integrated_Modifiers_NP	-.05958	.18003	.02406	-2.477	55	**.016**
Task1_LSA Integrated_LSA	.03821	.12180	.01628	2.348	55	**.023**

*Left Embeddedness

Because the IBM SPSS software does not calculate 'effect size' statistics at the significant points for t-tests, we used the following formula to compute eta squared values (Pallant, 2016).

Eta squared = $t^2 / t^2 + (N - 1)$

Using this formula, eta squared for both of the significant differences shown in Table 3 was obtained as 0.09, which shows a moderate effect size for the significant differences between the two task types in terms of modifiers per NP and LSA. The comparisons between the two speaking conditions on the basis of the complexity measures are summarized as follows.

A paired-samples t-test was conducted to compare TOEFL iBT independent and integrated speaking tasks in terms of complexity. Among the four measures representing complexity as mentioned above, two of them turned out to be significantly different. Correspondingly, the oral responses to total integrated tasks were significantly more complex in terms of the mean number of modifiers per NP (*M* = 0.72*, SD* = 0.11) than those to the independent task (*M* = 0.66*, SD* = 0.19), *t* (55) = -2.47, sig. (2-tailed) = 0.016, with a moderate effect size, eta squared statistic = 0.09. Based on the LSA measure, the independent speaking task indicated a significantly higher mean value (*M* = 0.24*, SD* = 0.12) than the total integrated tasks (*M* = 0.21*, SD* = 0.04), *t* (55) = 2.34, sig. (2-tailed) = 0.023, with a moderate effect size, eta squared statistic = 0.09.

5.2. Comparison of the two speaking conditions: Accuracy

The oral responses to the independent and integrated speaking tasks were further compared in terms of form-based accuracy. It seems essential to clarify that due to the nature of integrated speaking tasks, expecting respondents to reflect the content presented by the stimulus texts, ETS raters consider content relevance besides linguistic accuracy when rating iBT speaking tasks 2 to 4. On the contrary, the responses to the independent task are not required to reflect any given prompting input. Owing to this difference, we considered only the linguistic accuracy of the elicited productions. For this purpose, the two task types were compared by conducting a Paired-Samples T-Test, and the following descriptive statistics were obtained (Table 4).

Table 4: *Descriptive Statistics of Independent vs. Integrated Tasks in Terms of Accuracy*

	Task Type	Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Task1_Accuracy	**92.0409**	56	4.46055	.59607
	Integrated_Accuracy	**88.6146**	56	3.83177	.51204

As can be noticed in Table 4, oral responses to the independent speaking task indicated a higher mean score (92.04) than those to the total integrated tasks (88.61). Furthermore, the mean difference between the two speaking conditions, independent vs. integrated, was found to be significant (Table 5).

Table 5: *Paired Differences bet. Independent and Integrated Tasks in Terms of Accuracy*

Task Type	Mean	Std. Deviation	Std. Error Mean	t	df	Sig. (2-tailed)
Task1_Accuracy Integrated_Accuracy	3.42625	3.96216	.52947	6.471	55	**.000**

In terms of linguistic accuracy, the iBT independent speaking task showed a significantly higher mean value (*M* = 92.04*, SD* = 4.46) than the mean score obtained for the total integrated speaking tasks (*M* = 88.61*, SD* =3.83), *t* (55) = 6.47, sig. (2-tailed) = 0.000, with a large effect size, eta squared statistic = 0.43.

5.3. Comparison of the two speaking conditions: Fluency

We defined speech fluency as the number of produced syllables per minute (Ellis, 2008). The speech rate that was computed for each speaking sample was then taken as the main criterion to compare the iBT independent and integrated speaking tasks in terms of fluency. The outputs yielded by running a Paired-Samples T-Test revealed no significant difference between the two speaking conditions, although the independent responses left a slightly higher mean score. Table 6 encompasses the descriptive information related to the fluency of the tasks in focus.

Table 6: *Descriptive Statistics of Independent vs. Integrated Tasks in Terms of Fluency*

	Task Type	Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Task1_Fluency	**160.2321**	56	33.58625	4.48815
	Integrated_Fluency	**155.0238**	56	30.12954	4.02623

The complementary information concerning the fluency measure for the two speech-eliciting conditions is summarized in Table 7. Accordingly, the two speaking task types did not trigger significantly different performances (P > 0.05) in terms of fluency.

Table 7: *Paired Differences bet. Independent and Integrated Tasks in Terms of Fluency*

Task Type	Mean	Std. Deviation	Std. Error Mean	t	df	Sig. (2-tailed)
Task1_Fluency Integrated_Fluency	5.20833	24.83144	3.31824	1.570	55	**.122**

Paired-samples t-test comparison of the two speaking contexts based on speech fluency revealed that the independent speaking task caused a slightly higher mean value (*M* = 160.23*, SD* = 33.58) than the mean score obtained for the total integrated tasks (*M* = 155.02*, SD* = 30.12), but the mean difference was not large enough to make a significant difference, *t* (55) = 1.57, sig. (2-tailed) = 0.122.

6. Discussion

The present study aimed to draw a multi-faceted comparison between the TOEFL iBT independent and integrated speaking tasks. In particular, we intended to see in what ways the two speaking assessment contexts were distinct. To this end, a group of prospective TOEFL iBT Iranian candidates was studied for research purposes. The collected oral responses to the four speaking tasks were then analyzed in terms of the commonly-used discourse criteria known as CAF (Ellis, 2008, 2009). Altogether, there were six quantitative measures representing the CAF criteria that were undertaken to illuminate the comparisons. Throughout the following subparts, the results obtained from the analytical outputs are reviewed in light of the RQs as well as similar studies in the field.

6.1. TOEFL iBT independent vs. integrated speaking tasks in terms of complexity

Among the four complexity measures, the mean number of modifiers per noun phrase and LSA were found to show significant differences between the two speaking contexts. In contrast, the other two complexity measures, i.e., lexical diversity and left-embeddedness, did not indicate any significant difference in distinguishing the independent and integrated speaking tasks. Not surprisingly, the total integrated tasks showed a higher mean for modifiers per NP than the independent task for the same measure. This result could possibly be attributed to the academic context of speaking in two of the integrated tasks (3 & 4). The diversity of prompting concepts and ideas can also be detected in the campus-based speaking task 2, where the test takers are presented with both textual and aural inputs. The pre-modification density of the independent responses is supposed to be lower as the test takers need to conceptualize their original ideas (Levelt, 1989) under the time constraints of testing. This claim validates what Wigglesworth and Frost (2017) proposed in conjunction with the input ideas the test takers can benefit from when being involved in integrated tasks. In a similar vein, the recent finding can be verified in line with Biber and Gray (2013) who found some linguistic differences between the iBT independent and integrated speaking tasks. Yet, a more probable reason for the richer mean of modifiers per NP in the integrated speaking tasks might relate to the fact that the respondents merely carried over the existing noun phrases as presented by the stimulus texts. The recent claim confirms Frost et al. (2020), arguing a similar finding of an iBT integrated speaking task.

As for the LSA measure that turned significant, the responses to the independent task showed a significantly higher mean than those to the total integrated tasks. The most probable reason for this result is that the respondents had more repetition of key ideas when speaking from their own minds. This is because LSA concerns semantic overlap between explicit words and words that are implicitly related in meaning (McNamara et al., 2014). Owing to the content provided via task input, the diversity of ideas and concepts may rise in integrated speaking tasks, which leads to a lower LSA level. The discussed difference between the two task types reinforces Huang et al. (2018), concluding that the topical knowledge of test takers enriched by the prompting texts in TOEFL iBT integrated tasks may exert tangible variations in test-takers' oral outputs.

With respect to lexical diversity and left-embeddedness, which did not show any significant difference between the independent and integrated tasks, an unexpected result was obtained. The iBT integrated tasks were thought to cause more lexical diversity as such tasks provide the test takers with textual and aural inputs. Conversely, the lexical diversity of the oral performances was slightly, but not significantly, higher in the independent task than in the total integrated ones. Why the integrated tasks did not give rise to higher lexical variation can be linked to the similar conjecture made by Luoma (2004), Brown et al. (2005), and Shohamy et al. (2017). According to them, such tasks may impose some added cognitive load upon respondents, which may have frustrated our participants when they were trying to speak in response to the iBT integrated tasks.

Finally, left-embeddedness as a syntactic complexity measure, which stands for the mean number of words before main verbs, was not found to significantly distinguish the two speaking situations. In other words, the TOEFL iBT independent and integrated speaking tasks were found to be comparable in terms of this complexity measure. The finding might reflect the fact that, under the pressure of time, the participants attended more to transferring their ideas in a fluent way than making their sentences syntactically dense. The fact may corroborate the underlying tenets of the Trade-off Hypothesis (Skehan, 2014), claiming that different aspects of L2 production (e.g., the CAF triad) may be in competition.

6.2. TOEFL iBT independent vs. integrated speaking tasks in terms of accuracy

To meet the comparability of the two speaking conditions, form-based linguistic accuracy of the oral responses was measured for the planned comparisons. The observed results in terms of accuracy showed a significantly higher mean value for the independent speaking task than the total integrated ones. The finding may lend support to the impression that the integrated speaking tasks entangled the TOEFL test takers in conditions where they produced less accurate forms than in the independent task, where they had more freedom of production. It might be tenable to relate the lower accuracy of the integrated oral performances to the imposition of mainly formal and scientific ideas, which may have been less familiar to the test takers. This argumentation could spotlight muddied measurement (Weir, 1990, 1993, 2005), which implies the fact that any deficiency in the prerequisite texts/skills can lead to lower success in the main skill under assessment.

Another reason for the lower accuracy in the integrated speaking tasks, observed in the current investigation, may have stemmed from the higher modification per NP found in the integrated oral responses. The increase might consequently have decreased the accuracy of the responses to the integrated tasks due to the trade-off effects (Ellis, 2009; Skehan, 2014). In other words, when the test takers tried to enrich their speech complexity by increasing the length of noun phrases through pre-modification, they may have ended up with lower accuracy in their speech. The finding is meant to confirm the mentioned trade-off effects, saying that there exists some sort of competition among different aspects of L2 speech due to test-takers' limited mental capacity in the target language.

6.3. TOEFL iBT independent vs. integrated speaking tasks in terms of fluency

Speech rate standing for fluency was not found to significantly distinguish the independent from integrated speaking productions. Although the speech rate of the independent responses was slightly higher compared to the integrated ones, the two speech-eliciting situations were comparable as long as the fluency index was focused. This comparability, as the results revealed, needs some further hesitation because integrated tasks may be expected to promote fluency. The suggestion centers on the claim that integrated tasks are believed to provide test takers with the required ideas to talk about (Barkaoui et al., 2013; Frost et al., 2020; Wigglesworth & Frost, 2017, among others). In spite of this seeming advantage, the participants in this study showed, though not significantly, a higher speech rate when speaking to the independent task, the context which seems to make L2 speakers more hesitant due to the original conceptualization of ideas not prompted by either textual or aural input. Thus, it might be rational to raise muddied measurement, which was discussed earlier, as one possibility for a slightly lower speech fluency observed in TOEFL iBT integrated speaking tasks. The mentioned hypothesis is consistent with Luoma (2004), Brown et al. (2005), and Shohamy et al. (2017) when they comment on the potential added burden of such integrated (speaking) tasks. This added load may have left the participants in this study unable to take advantage of the ready-made prompting ideas to perform more fluently in the iBT integrated speaking tasks.

7. Conclusions

In this study, we set to compare TOEFL iBT Iranian test-takers' oral language performance in the independent and total integrated speaking tasks. The concluding results the current study managed to reveal are summarized as follows.

The two speaking conditions did not show any significant variations in the participants' speaking quality in terms of lexical diversity and left-embeddedness. Nonetheless, the same speaking contexts were significantly different based on the mean number of modifiers per NP and LSA.

The iBT independent and integrated speaking tasks turned out to be significantly different when linguistic accuracy was focused, with a higher mean value for the former context (independent). The two speech-eliciting situations, however, were not able to distinguish the oral responses in terms of fluency, which was measured as the rate of oral production.

Overall, the TOEFL iBT independent and integrated speaking tasks were comparable based on three CAF measures and separable concerning three other CAF measures. Among those three significant measures, the independent speaking task revealed higher mean values in terms of LSA and accuracy, whereas the total integrated tasks raised the mean number of modifiers per NP. In line with Kyle, Crossley, and McNamara (2015), we can conclude that the two speaking contexts were partially similar and partially different on the basis of various CAF dimensions. The findings, therefore, may have potential implications for an array of stakeholders involved in L2 speaking assessment. Specifically, the implications may inspire both TOEFL iBT trainers and trainees in coping with the speaking tasks more effectively. The discussed findings could also help L2 speech raters (including those affiliated with ETS) when scoring the two speaking task types, especially, the integrated type.

Regarding the limitations of the present study, we can acknowledge the limited sample-size as well as the participants' nationality and L1 background. In addition, it might be argued that the two speaking conditions entail some differences that could impede the comparisons. One discrepancy belongs to the number of tasks in TEOFL iBT speaking subsection, which consists of only one independent but three integrated tasks. The other difference pertains to the response time allowed in the independent task (45 seconds), which is significantly lower than the response time in the integrated tasks (60 seconds in each one). This means TOEFL iBT respondents are more limited both for the number of tasks and the offered response time in the independent task compared to the integrated ones. These variations can be concentrated on for further research on the dichotomy of independent and integrated tasks in the TOEFL iBT speaking module.

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.

Ary, D., Jacobs, L. C., Irvine, C. K. S., & Walker, D. A. (2019). *Introduction to research in education* (10th ed.). Cengage Learning, Inc.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Barkaoui, K., Brooks, L., Swain, M., & Lapkin, S. (2013). Test-takers' strategic behaviors in independent and integrated speaking tasks. *Applied Linguistics, 34*(3), 304-324. https://doi:10.1093/applin/ams046

Biber, D., & Gray, B. (2013). *Discourse characteristics of writing and speaking task types on the TOEFL iBT test: A lexico-grammatical analysis*. TOEFL iBT Research Report 19. Educational Testing Service. https://doi:10.1002/j.2333-8504.2013.tb02311.x

Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test taker performance on English-for-Academic-Purposes speaking tasks (TOEFL Monograph Series MS-29). Princeton, NJ: Educational Testing Service.

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

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

De Jong, N. H. (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

Educational Testing Service. (2020). *TOEFL iBT®: Test and score data summary 2019*. https://www.ets.org/toefl

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: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. (2010). *Practical language testing*. Hodder Education.

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., & 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

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

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

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.

Pallant, J. (2016). *SPSS survival manual* (6th ed.). Open University Press.

Roever, C., & Ikeda, N. (2021). What scores from monologic speaking tests can(not) tell us about interactional competence. *Language Testing, 39*(1), 7-29. https://doi.org/10.1177/02655322211003332

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.

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

Weir, C. J. (1993). *Understanding and developing language tests*. Hemel Hempstead: Prentice Hall.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Wigglesworth, G., & Frost, K. (2017). Task and performance-based assessment, in Shohamy et al. (eds), 2017. Springer International Publishing AG. https://doi.org/10.1007/978-3-319-02261-1_8

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), 485-510. https://doi.org/10.1177/0265532220951508