Chabahar Maritime University

Iranian Journal of English for Academic Purposes ISSN: 2476-3187
IJEAP, (2017) Vol. 6 Issue. 2 (Previously Published under the Title: Maritime English Journal)

Input-induced Variation in EFL Learners' Oral Production in Terms of Complexity, Accuracy, and Fluency

[1] Mehdi Karami
[2] Manoochehr Jafarigohar*

IJEAP- 1806-1232

[3] Zia Tajeddin
[4] Afsar Rouhi

Abstract

Researchers have extensively studied phenomena that affect a second language learner's oral production while there is scant evidence about input-related factors. Accordingly, the present study sought to investigate how variation in oral production is caused by the input they receive from different course materials. To this end, the study included a micro-evaluation study of three course materials and a quasi-experimental pretest-posttest design with three groups of participants (N = 72) instructed with three different course materials (*New Headway*, *American English File, and Top Notch & Summit*). Speech samples elicited through picture-cued oral narrative tasks at three data collection times were quantitatively assessed for complexity, accuracy, and fluency (CAF). A one-way Multivariate Analysis of Variance (MANOVA) was used to compare the means of CAF scores. With respect to pretest-posttest differences on measures like the average number of subordinate clauses per AS-unit for grammatical complexity, "D" index for lexical complexity, percentage of error-free clauses for accuracy, and number of dysfluencies for fluency, the results indicated that different course materials have insignificant effects on the variation in grammatical complexity but a clear effect on the variability in lexical complexity, accuracy, and fluency. Micro-evaluation of the three course materials revealed that this variability might well be attributed to the characteristics of the speaking tasks in the textbooks. One important implication is that EFL materials developers can provide the learners with the experience of speaking tasks with particular features if they want to promote gains in a special dimension of oral performance (e.g., fluency).

Keywords: Input, Course Materials, Oral Production, Complexity, Accuracy, Fluency

1. Introduction

Oral production ability has a special status in second language (L2) pedagogy. In communicative language teaching, L2 learners are encouraged to engage in copious amounts of spoken language in the classroom (Hughes, 2011). Variability in oral production is a complex phenomenon which can be attributed to multiple sources both from the outside world and from inside a self-organizing dynamic system (Verspoor, Lowie, & van Dijk, 2008). Among these sources, input provided by different learning environments has a direct impact on how the foreign language is used (Skehan & Foster, 1997). All language-learning theories, especially connectionist and emergentist theories, affirm that learning is first and foremost driven by input (de Bot, Lowie, & Verspoor, 2005), and that lack of exposure to input leads to the decline of connections of bits of information in mind and accordingly to a weakening of the network as a whole (Verspoor, Lowie, & de Bot, 2008). Input is very important in causing variability

[1] PhD student of TEFL, karami_m58@yahoo.com; Department of Teaching English as a Foreign Language (TEFL), Payame Noor University (PNU), Tehran, Iran

[2] Associate professor, jafari@pnu.ac.ir; Department of Teaching English as a Foreign Language (TEFL), Payame Noor University (PNU), Tehran, Iran

[3] Professor, zia_tajeddin@yahoo.com; Department of Foreign Languages, Allameh Tabataba'i University, Tehran, Iran

[4] Associate professor, afsar.rouhi@pnu.ac.ir; Department of Teaching English as a Foreign Language (TEFL), Payame Noor University (PNU), Tehran, Iran

in language performance because it contains the linguistic data needed for a learner's L2 system to reorganize itself constantly in order to find equilibrium (Barcroft & Wong, 2013). With regard to the significance of input, it can be predictable that classroom input—mainly course materials—might play a key role in speaking performance among language learners because some of these textbooks can even contribute to the failure of many L2 learners (Tomlinson, 2008), and over time this will cause variation in speaking among them since up to 90 percent of classroom time is mediated by textbooks (Thornbury, 2014). Despite this, it is curious that classroom-based input rarely, if ever, features in discussions of factors that influence L2 oral production (Verspoor et al., 2008). Ellis (2009) reviewed the studies on L2 learners' oral production, and input factors were not among the variables investigated. Thornbury (2014) has already made the point that "there have been relatively few studies into the impact of coursebooks on learning" (p. 101).

Another problem is that even in the most comprehensive models of L2 speaking assessment (Fulcher, 2003), course materials have never been directly considered as a variable in the performance of test takers. Abilities of the test taker, task conditions, goals, topics, the assessment criteria, and the raters have all been known to have a role in any inferences that are made about the test taker, while a fairer model of speaking assessment should also involve documentation of systematic effects of classroom input on the learners' production of language. Hence, exploring input material factors may provide a source of valuable understanding about variation in EFL learners' oral production, and the present study aimed at investigating how this variation is caused by the input they receive from different sources.

2. Literature Review

2.1. The Construct of Speaking

Oral production/speaking has for a long time retained a very important status in the language classrooms, but its measurement has also proved to be a major difficulty because there have not been any constructs serving as a basis for assessing specific features of speaking performance (Ellis, 2003). Nevertheless, an investigation of the assessment scales of language proficiency (ACTFL, 2012, Council of Europe, 2001; IELTS, 2007) and the frameworks for speaking assessment (Fulcher, 2003) reveals that lexical and grammatical complexity, accuracy, and fluency (CAF) are the most relevant linguistic components of the construct of oral production that have become widely available as dependent variables in L2 research since the 1990s (Housen & Kuiken, 2009). Skehan (1998) argued that CAF are effective indexes for measuring performance on a particular speaking task. They have special empirical and theoretical status. Factor analyses have identified CAF as distinct areas of L2 oral performance (Norris & Ortega, 2009). "Linguistic complexity and accuracy are placed within the range of linguistic competences […], while fluency is regarded as an index of strategic competence" (Czwenar, 2014, p. 82).

Complexity is characterized as "the ability to use a wide and varied range of sophisticated structures and vocabulary in the L2" (Housen, Kuiken, & Vedder, 2012, p. 1). Accuracy may best be defined as the ability to produce "error-free" (Ellis, 2003, p. 42) and "target-like" (Yuan & Ellis, 2003, p. 2) language. Fluency is the production of language in real time with "native-like rapidity" (Housen et al., 2012, p. 2) and without "undue pausing or hesitation" (Ellis & Barkhuizen, 2005, p. 139) or "reformulation" (Ellis, 2003, p. 342).

2.2. Studies on CAF of Oral Production

The relevant factors that affect the development of CAF in L2 oral production are diverse in nature (Housen et al., 2012), but this study sought to build upon the body of research on the learning context to explore the variable effects of course materials, as a critical input element in the EFL learning environment, on L2 oral production.

Segalowitz and Freed (2004) investigated whether at home (AH) and studying abroad (SA) learning contexts differentially supported gains in oral performance. The AH group included 18 English speaking students studying Spanish at the University of Colorado in the US. The SA group consisted

of 22 students from the United States commencing a semester at the Universidad de Alicante in Spain. With respect to pretest-posttest differences on oral fluency measures, the results indicated greater gains for the SA students.

One research hypothesis formulated in the study by Tavakoli and Foster (2008) was that similar patterns of results associated with lexical diversity, accuracy, and fluency of oral production would be obtained for learners based within the target language environment (London) and those based outside it (Tehran). The learners in both contexts were in typically communicative classrooms in which they had plenty of speaking and listening activities. They were asked to retell stories from picture prompts. The comparison of the data from the London and Tehran cohorts showed that the learning environment had little discernible effect on accuracy or fluency but a clear effect on lexical diversity.

In the study by Pérez-Vidal and Juan-Garau (2011), it was hypothesized that because each of the formal instruction (FI) at home and SA contexts had differential patterns of input exposure, their effects on the participants' oral development would also be different. The study included 20 full-time Catalan/Spanish EFL undergraduates who spent a compulsory three-month sojourn in an English-speaking university. A role-play task was used to tap into oral performance. In order to contrast gains obtained during FI with gains obtained abroad, a pre-test/post-test design was applied for each of the two contexts. The findings revealed a significant improvement as a result of SA in the participants' fluency of oral performance. Additionally, accuracy improved significantly. In contrast, grammatical complexity only showed a tendency toward significant improvement in oral performance.

Mora and Valls-Ferrer (2012) also carried out a study comparing the effects of FI at home and SA learning periods on the speaking skills of advanced-level Catalan-Spanish learners of English. The 3-month SA term was a requirement, and most of them (88%) spent it in the United Kingdom and Ireland. The oral data analyzed were obtained from a guided interview conducted in pairs administered at three different times over a 2-year period. Gains obtained after two terms of FI at home were then compared with those obtained after the SA term. The results provided evidence of robust gains in oral fluency, moderate improvement in accuracy, and no gains in complexity occurring during an SA period; and lack of such gains during FI.

Although these studies reveal that exposure to input in natural language learning conditions (SA) can be more effective than FI settings (AH) in improving oral production, especially for fluency and accuracy, many other questions remain regarding what roles various kinds of instructions, input, and contexts can play in this regard. For example, there is still a need for comparing gains in oral performance obtained from different course materials at different FI settings at home (language schools in EFL contexts). To the best of the researchers' knowledge, this area has not been investigated so far.

Recognizing FI or the classroom as the main source of language input for EFL learners, this study was another step in a continuous line of inquiry aiming at investigating classroom input factors that have hardly ever been examined before for their effects on variation in L2 speaking performance. The significance of this study is paramount as it may contribute to the modification of assessment models of L2 oral performance and address a number of issues related to designing speaking tasks in instructional materials. The main research purpose was therefore to examine whether variability in EFL learners' CAF of oral production can be caused by different sources of input. The following research questions were addressed to determine where exactly this variability occurs:

1. To what extent is the variation observed in the grammatical complexity of EFL learners' oral production induced by different course materials?
2. To what extent is the variation observed in the lexical complexity of EFL learners' oral production induced by different course materials?
3 To what extent is the variation observed in the accuracy of EFL learners' oral production induced by different course materials?
4. To what extent is the variation observed in the fluency of EFL learners' oral production induced by different course materials?

3. Methodology

The present study included a micro-evaluation of three course materials and a quasi-experimental pretest-posttest design consisting of one categorical independent variable—classroom input—with three different course materials and three continuous dependent variables of oral production, i.e., participants' CAF scores on oral narrative tasks administered before and after the intervention period.

3.1. Participants

For a MANOVA design with a 95% confidence level, 5% confidence interval, and a large effect size, the sample size was calculated by G*Power to be 72; therefore, three groups of Persian-speaking advanced learners of English (37 female and 35 male) instructed via three different course programs participated in this study (Table 1). The average length of English language learning experience was four years (*SD* = 2.2). Their average age was 17.5 years (*SD* = 3.51), and the age range was 16-30 years; no participant older than 30 was included to avoid confounding effects from their biological age (for aging effects, see Birdsong, 2014) and to ensure proper continuity in the length of exposure.

Table 1. Description of Course Programs and Participants

Course program	Year of publication	Levels	Participants' Level	*Male* (*n*)	*Female* (*n*)
1. *New Headway*	2015 (4th edition)	A1-C1	C1	12	12
2. *American English File*	2014 (2nd edition)	A1-C1	C1	11	13
3. *Top Notch & Summit*	2012 (2nd edition)	A1-C1	C1	12	12

3.2. Instrumentation

Picture-cued oral narrative tasks were used to elicit speech samples at three data collection times. In an oral narrative task, the speaker has a cartoon-strip story that can be visually depicted in 6-8 pictures, and the primary aim of the test taker is to produce a coherent story (Fulcher, 2003).

It is impossible to measure the effect of different EFL course materials on the learners' oral production when other variables such as initial proficiency of the learners and amount of exposure to language are tightly associated to learners' oral performance (de Bot et al., 2005; Munoz, 2014). Therefore, upon entering the current study, participants took Test of Spoken English or TSE (Papajohn, 2005) to ensure that they are at the same level of proficiency. Language Contact Profile (Freed, Dewey, Segalowitz, & Halter, 2004) or LCP, a socio-educational background questionnaire, was also used to record the participants' linguistic backgrounds and amount of contact/interaction in English with native and non-native speakers.

In many EFL programs, these are mainly the textbooks that determine the kind of methods being used (Richards, 2014; Thornbury, 2009). However, some professional teachers go beyond the textbook methodology, and this will affect the oral performance of language learners. Therefore, it was decided that a the Teachers' Sense of Efficacy Survey (Lee, 2009) was the most feasible and appropriate instrument to have the three teachers in the current study filled to ensure they possess the same sets of qualifications related to English teaching confidence, personal teaching confidence, attitudes toward English, attitudes toward the current English education policy and practices, and English language proficiency. The extent of richness of input provided in each classroom and the degree of textbook adaptation by teachers in groups could also affect the participants' oral production ability. Thus, semi-structured interviews were conducted taking a variety of forms of teacher adaptation of the EFL courses into account—modifying content, adding or deleting content, reorganizing content, addressing omissions, modifying tasks, and extending tasks (Richards, 2001).

3.3. Data Collection Procedure

3.3.1. Course materials

The selection of course programs met a set of preconditions: (a) course materials had to be among the most widely used materials in the EFL context; (b) they had to spread across A1-C1 in CEFR framework; and (c) course materials with different learning objectives had to be selected.

New Headway (Soars, Soars, Hancock, & Williamson, 2015), *American English File* (Latham-Koenig & Oxenden, 2014), and *Top Notch & Summit* series (Saslow & Ascher, 2012) are commonly utilized EFL teaching series worldwide for young adults and adults (Table 1). To ascertain if these textbooks adopted different orientations to teaching oral production, a micro-evaluation (see Ellis, 1998) of their speaking tasks had to be undertaken using two frameworks for analysis of task characteristics. According to the framework proposed by Skehan (2001), speaking tasks vary in the extent of their contribution to complexity, accuracy, or fluency as to whether they require information that is familiar to L2 learners or not. Some tasks are dialogic compared with others where extended turns are required. The timeline for the information underlying some tasks is clearly identifiable. In some tasks, a simple decision has to be made, while in other tasks, the case a learner argues during a task has to predict other possible outcomes. Some tasks require participants simply to reproduce the information. Others require some degree of on-line computation.

Table 2. Summary of The Effects of Task Characteristics on Complexity, Accuracy, and Fluency (Skehan, 2001)

Task characteristic	Accuracy	Complexity	Fluency
Familiarity of information	No effect	No effect	Slightly greater
Dialogic vs. monologic tasks	Greater	Slightly greater	Lower
Degree of structure	No effect	No effect	Greater
Complexity of outcome	No effect	Greater	No effect
Transformations	No effect	Planned condition generates greater complexity	No effect

Ellis's (2003) framework of task features was also reviewed to address the contextual factors as well. Sometimes the input to the task takes the form of a picture which must then be communicated verbally to the hearer (contextual support). A second input factor concerns the number of features that need to be manipulated by the speakers. The individual learner's familiarity with a particular topic will also affect oral production. Shared tasks typically involve decision-making and thus require argumentation, whereas split-information tasks result in description. Production is also influenced by whether learners are asked to carry out a single or a dual task demand. Open tasks are those where the participants know there is no pre-determined solution. Closed tasks are those that require students to reach a single solution. The inherent structure of the outcome refers to whether the product the task elicits exists in some kind of pre-structured form or not. Discourse mode is the specific rhetorical structure of an oral task that reflects its overall communicative function.

Table 3. Task Design Features Affecting Learner Production (Ellis 2003)

Design Variable	Fluency	Accuracy	Complexity
A. Input variables			
1. Contextual support	Tasks with contextual support	Tasks with no contextual support	Tasks with no contextual support
2. Number of elements	Tasks with few elements		Tasks with many elements
3. Topic	Tasks that generate conflict, tasks that are familiar		
B. Task conditions			
1. Shared vs. split information			Shared information tasks
2. Tasks demands	Tasks that pose a single demand		Tasks that pose a single demand
C. Task outcomes			
1. Closed vs. open tasks	Closed tasks	Open tasks	Open tasks with divergent goals
2. Inherent structure of the outcome	A clear inherent structure	A clear inherent structure together with opportunity for planning	
3. Discourse mode			Narrative task > descriptive task > Argument discussion Narrative > argument

Skehan's (2001) and Ellis's (2003) frameworks supported some generalizations that were applied to while coding 409 speaking tasks from *New Headway*, 532 from *American English File*, and 658 from *Top Notch & Summit*. Speaking activities were coded and categorized into task features that affect complexity, accuracy, or fluency. A sample of 10% of the total corpus, randomly selected, was subjected to coding validation by an experienced SLA researcher. The inter-coder agreement, calculated as percentage of identical coding, proved to be greater than 88%.

3.3.2. Intervention period

Before the intervention period, the possible effects of a few learner variables on speaking performance needed to be examined to ensure that upon entering the present study the three groups were equivalent in terms of their oral proficiency and the amount of exposure to and practice of English inside and outside of the classroom. Two experienced raters judged the speaking performance of the participants in TSE using "Rating Scale for the TOEFL Test of Spoken English" (ETS, 2001, p. 29). The results of one-way ANOVA revealed that the actual difference in mean scores between the three groups was quite small, $F(2, 69) = 2.09$, $p = .13$., e*ta squared* = .05. Post-hoc comparisons using the Tukey HSD test also indicated that the mean score for *New Headway* group ($M = 42.92$, $SD = 7.21$) did not differ significantly from either *American English File* ($M = 38.33$, $SD = 9.16$) or *Top Notch & Summit* ($M = 40.210$, $SD = 6.83$). The participants in all groups also filled in the LCP. The median scores for all sub-scales related to speaking and language use were calculated and entered into ANOVA. The actual difference in mean scores was quite small for all sub-scales except for "hours/day speaking in English", $F(2, 69) = 4.36$, $p = .01$, which was moderate (*eta squared* = .06). Moreover, post-hoc comparisons using the Tukey HSD test indicated that *American English File* did not differ significantly from either *New Headway* or *Top Notch & Summit* in any of the sub-scales.

Furthermore, some teacher factors needed to be taken care of. The three teachers in the present study filled in the self-efficacy questionnaire. Comparison was drawn between the raw scores,

and it revealed no significant differences across them in their teaching efficacy (Table 4). Semi-structured interviews were also conducted with the three participant teachers to document what adaptations they had made to the official course of their institute to enrich it. All interviews, conducted in English, were audio-recorded and transcribed. The results of the interview analysis revealed that the three teachers had similar approaches to adapting teaching materials. All indicated that there were only rare cases in which there was a need to change some parts of the textbook to suit the learners' age and cultural background. Nobody mentioned any deleting. All admitted that they did not feel additional practice tasks needed to be added but most felt a need to enrich the syllabus only in terms of vocabulary practice. These results enabled the researchers to study the impact of textbook input on speaking after controlling for the impact of teachers.

Table 4. Important Findings of Teacher Efficacy Questionnaire

Sub-scales	*New Headway teacher*	*English File teacher*	*Top Notch & Summit teacher*
[a] English Teaching Confidence	6.5	7	7.5
[b] Personal Teaching Confidence	5	4.5	5
[b] Attitudes toward English	4	5.5	5
[b] Attitudes toward the current English education policy and practices	4	5	4
[b] Current Level of Speaking	5.5	5.5	5
Sex	Male	Male	Male
Age	20s	30s	30s
Teaching experience (yrs)	8.5	8	11
Highest degree	MA in TEFL	MA in TEFL	MA in TEFL
Attending in-service training programs	Yes	Yes	Yes
English use in a period of English class	90-99%	90-99%	90-99%

[a] max. score = 9; [b] max. score = 6

During the pretest phase of the study, participants in each group met the first author individually and were required to study a picture-cued oral narrative task. One minute was devoted to this to let them have enough time to gather their thoughts about how they would narrate it within 4 minutes. The audio-recorded speech data were transcribed in CHAT format using the CLAN (Computerized Language Analysis) software of the CHILDES program (MacWhinney, 2000). Preparation of the transcripts for coding and analysis began by segmenting each text into AS-units. An AS-unit is "a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either" (Foster, Tonkyn, & Wigglesworth, 2000, p. 365).

After that the oral narrative monologs were graded with an analytic rubric that consisted of grammatical and lexical complexity, accuracy, and fluency as valid measures of quantifying speaking performance. The mean number of subordinate clauses per AS-unit was a measure of grammatical complexity to be tapped into. Lexical complexity was measured by means of the "D" index that has been integrated within CLAN and is computable through the VocD program. D ranges between 10 and 100 and a higher value indicates a more diverse text (McKee, Malvern, & Richards, 2000). Accuracy was measured as the percentage of error-free clauses, which is defined as "The number of error-free clauses divided by the total number of independent clauses, sub-clausal units and subordinate clauses multiplied by 100" (Ellis & Barkhuizen, 2005, p. 150). Fluency was measured by means of calculating the number of dysfluencies. The total number of functionless repetitions/reformulations was divided by the total amount of time expressed in seconds and multiplied by 60 to calculate the number of dysfluencies (Ellis & Barkhuizen, 2005). It should be borne in mind that with fluency a reduction in values clearly represents an improvement. A sample of 10% of the total transcripts (i.e. 74/740 AS-

units), randomly selected, was subjected to coding validation by an experienced second language acquisition (SLA) researcher. The inter-coder agreement, calculated as percentage of identical scoring, proved to be greater than 95% for each measure.

A period of approximately 3 months intervened between the initial oral narrative task and the posttest tasks. During the intervention period, participants, who had been assigned to one of the three course groups, received 48 hours of classroom instruction offered by three different course materials. After the intervention period participants were told that they would do two more oral narrative tasks which was similar to the pretest task in both difficulty level and story content.

3.4. Data Analysis

Analysis of the data in the micro-evaluation study included counting the number and calculating the percentage of speaking tasks with complexity, accuracy, or fluency features in the three course materials. Furthermore, given the nature of the research questions, as well as the 3-by-3 research design, performing a one-way between groups MANOVA compared the overall effect of three course materials on average performance of L2 learners in oral narrative tasks and for the dependent variables–CAF– before and after the intervention period. The alpha for achieving statistical significance was set at .05.

4. Results

4.1. Micro-Evaluation of the Speaking Tasks in the Course Materials

The four research questions of the study could only be addressed if the researchers ascertained that the textbooks used in the three EFL programs differed in terms of orientations to teaching oral production. Thus, before anything else, the results of the micro-evaluation of all the speaking tasks in *New Headway, American English File,* and *Top Notch & Summit* are presented here. Table 5 displays the mean percentage of task types in each course that have been classified as contributing to complexity. The difference in the emphasis on complexity of oral production was not that much different.

Table 5. Comparing Speaking Tasks in the Textbooks for Their Contribution to Complexity

	New Headway		*American English File*		*Top Notch & Summit*	
	n/N	%	*n/N*	%	*n/N*	%
Dialogic tasks	252/409	61.61	323/532	60.90	490/658	74.46
Tasks that need transformations	250/409	61.12	453/532	85.16	599/658	91.07
Tasks with no contextual support	139/409	33.82	182/532	34.24	154/658	23.40
Tasks with many elements	208/409	50.85	265/532	49.81	281/658	42.70
Shared information tasks	99/409	24.20	95/532	17.85	94/658	14.28
Tasks that pose a single demand	215/409	52.43	280/532	52.63	384/658	58.35
Open tasks	118/409	28.85	103/532	19.36	96/658	14.58
Narrative tasks	36/409	8.83	34/532	6.35	20/658	3.03
Mean percentage		40.21		40.78		40.23

Note: *n/N* = proportion of tasks with particular characteristics to total number of tasks in the course; % = percentage of tasks with particular characteristics in the course

Table 6 displays the mean percentage of speaking tasks in each course that have been classified as contributing to accuracy. The difference in the emphasis on accuracy was striking particularly between *New Headway* and *American English File* (*M* = 44.15%, 37.55%). These results show that in comparison to the other textbooks, *New Headway* contained more accuracy-oriented speaking tasks.

Table 6. Comparing Speaking Tasks in the Textbooks for Their Contribution to Accuracy

	New Headway		*American English File*		*Top Notch & Summit*	
	n/N	%	*n/N*	%	*n/N*	%
Dialogic tasks	252/409	61.61	323/532	60.90	490/658	74.46
Tasks with clear inherent structure	214/409	52.32	190/532	35.71	396/658	60.18
Tasks with no contextual support	139/409	33.82	182/532	34.24	154/658	23.40
Open tasks	118/409	28.85	103/532	19.36	96/658	14.58
Mean percentage		44.15		37.55		43.15

Table 7 displays the mean percentage of speaking tasks in each course that have been classified as contributing to fluency. The difference in the emphasis on fluency was also significant. The results of the task analysis revealed that *American English File* and *Top Notch & Summit* were more fluency-oriented (*M* = 59.34%, 71.33%) than *New Headway* with a mean percentage of 56.34% tasks with fluency features.

Table 7. Comparing Speaking Tasks in the Textbooks for Their Contribution to Fluency

	New Headway		*American English File*		*Top Notch & Summit*	
	n/N	%	*n/N*	%	*n/N*	%
Tasks with familiar information	355/409	62.34	463/532	87.03	608/658	92.40
Tasks with clear inherent structure	214/409	52.32	190/532	35.71	396/658	60.18
Tasks with contextual support	208/409	50.92	269/532	50.56	506/658	76.89
Tasks with few elements	200/409	48.89	266/532	50	378/658	57.44
Tasks that pose a single demand	215/409	52.43	280/532	52.63	384/658	58.35
Closed tasks	291/409	71.14	426/532	80.07	559/658	84.95
Mean percentage		56.34		59.33		71.70

4.2. Oral Narrative Tasks

Before proceeding with the MANOVA results of oral narrative tasks, some preliminary assumption tests were conducted. A one-sample Kolmogorov-Smirnov was calculated to assess the normality of distribution of CAF scores. All pretest and posttest scores were greater than .05 and normally distributed. Maximum Mahalanobis value for pretest scores (13.64) and two posttest scores (13.93 and 13.85) was smaller than the critical value (18.47) suggesting multivariate normality in our data file. Generating scatterplot matrices between each pair of the dependent variables and separately for each course group did not show any obvious evidence of non-linearity. The strength of correlations among pretest and posttest scores was checked and most of the coefficients were moderate (.25 to .55). In our data file, Box's M significance value was .032 for pretest scores and .018 for two posttest scores that are larger than .001. The assumption of homogeneity of variance-covariance matrices, therefore, has been met (Pallant, 2007).

No serious violations were noted in the preliminary assumption testing; therefore, the first one-way between-groups MANOVA was performed before the intervention period. In separate

examinations of the results for the dependent variables (Table 8), none of the differences reached statistical significance. Thus, it can be concluded that, before the intervention program, the three groups were almost equivalent in all dimensions of oral production, i.e. CAF.

Table 8. Tests of Between-Subjects Effects on the Separate Dependent Variables for Pretest Scores

Dependent Variable	Index	Type III Sum of Squares	*df*	Mean Square	*F*	*Sig.*	Partial Eta Squared
Grammatical Complexity	mean number of clauses per AS-unit	.12	2	.06	.55	.57	.01
Lexical Complexity	D	597.44	2	298.72	2.51	.08	.06
Accuracy	percentage of error-free clauses	2327.52	2	1163.76	2.42	.09	.06
Fluency	number of dysfluencies	4.77	2	2.38	1.15	.32	.03

After the intervention program, the most important thing the authors needed to know was the exact dimension of oral production that variability occurred more in. Thus, in response to the four research questions, the results of MANOVAs were considered separately for each dependent variable, i.e. grammatical complexity, lexical complexity, accuracy, and fluency.

4.2.1. Research question one

As response to research question one, the difference in grammatical complexity scores of oral narratives between *New Headway*, *American English File*, and *Top Notch & Summit* groups did not reach statistical significance (Table 9). An inspection of the mean scores of two posttests also indicated that the three course groups did not report different levels of grammatical complexity—*New Headway* (*M* = 1.86 and 1.91, *SD* = .32), *American English File* (*M* = 1.89 and 1.94, *SD* = .31), and *Top Notch & Summit* (*M* = 1.84 and 1.89, *SD* = .36). These results together with the results of the micro-evaluation of the speaking tasks, which revealed the three course materials put the same emphasis on complexity of oral production, show that no variation in the grammatical complexity of EFL learners' oral production can be induced by different course materials.

Table 9. Tests of Between-Subjects Effects for Posttest Scores of Grammatical Complexity

Dependent Variable	Index	Test period	Type III Sum of Squares	*df*	Mean Square	*F*	*Sig.*	Partial Eta Squared
Grammatical Complexity	mean number of clauses per AS-unit	Posttest 1	.02	2	.01	.12	.87	.00
		Posttest 2	.02	2	.01	.12	.87	.00

4.2.2. Research question two

Using Bonferroni adjusted alpha levels of .11 and .13 for two posttest scores, the difference in lexical complexity could reach statistical significance between the three course groups $F(2, 69) = 5.60$ and 6.73, $p = .006$ and .002, partial eta squared = .14 and .16 (Table 10). An inspection of the mean scores of two posttests indicated that the *New Headway* group reported higher levels of lexical complexity of oral production—*New Headway* (*M* = 45.12 and 46.64, *SD* = 10.70) compared with *American English File* (*M* = 34.66 and 35.16, *SD* = 13.32) and *Top Notch & Summit* (*M* = 39.26 and

40.26, *SD* = 7.85). Thus, in response to the second research question, it was found that after the intervention program in this study, the variation observed in the lexical complexity of EFL learners' oral production was the result of being instructed with different course materials.

Table 10. Tests of Between-Subjects Effects for Posttest Scores of Lexical Complexity

Dependent Variable	Index	Test period	Type III Sum of Squares	*df*	Mean Square	*F*	*Sig.*	Partial Eta Squared
Lexical Complexity	D	Posttest 1	1318.24	2	659.12	5.60	.00	.14
		Posttest 2	1587.12	2	793.56	6.73	.00	.16

4.2.3. Research question three

Using Bonferroni adjusted alpha levels of .06 and .08 for two posttest scores, difference in accuracy of oral production also reached statistical significance between the groups, $F(2, 69) = 3.56$ and 4.31, p = .03 and .01, partial eta squared = .09 and .11 (Table 11). An inspection of the mean scores of two posttests also indicated that the *New Headway* group reported higher levels of accuracy of oral production—*New Headway* (*M* = 66.63 and 68.63, *SD* = 19.97), *American English File* (*M* = 51.03 *and* 51.28, *SD* = 26.40), and *Top Notch & Summit* (*M* = 53.21 and 54.21, *SD* = 18.58). Thus, in response to the third research question, these results together with the results of the micro-evaluation of the speaking tasks, which revealed *New Headway* contains more accuracy-oriented speaking tasks, show that to a large extent the variation observed in the accuracy of EFL learners' oral production was induced by different course materials.

Table 11. Tests of Between-Subjects Effects for Posttest Scores of Accuracy

Dependent Variable	Index	Test period	Type III Sum of Squares	*df*	Mean Square	*F*	*Sig.*	Partial Eta Squared
Accuracy	percentage of error-free clauses	Posttest 1	3429.27	2	1714.63	3.56	.03	.09
		Posttest 2	4144.29	2	2072.14	4.31	.01	.11

4.2.4. Research question four

Using Bonferroni adjusted alpha levels of .05 and .06 for two posttest scores, difference in fluency of oral production could also reach statistical significance between the groups, $F(2, 69) = 3.23$ and 3.26, $p = .04$ and .04, partial eta squared = .08 and .08 (Table 12). The mean scores of two posttests also indicated that *American English File* and *Top Notch & Summit* groups reported less proportion of dysfluencies (i.e. higher fluency)—*American English File* (*M* = 1.17 and 1.15, *SD* = 1.18), *Top Notch & Summit* (*M* = 1.85 and 1.82, *SD* = 1.49), and *New Headway* (*M* = 2.16 and 2.13, *SD* = 1.4). Thus, in response to research question four, these results along with the results of the micro-evaluation of the speaking tasks, which revealed *American English File* and *Top Notch & Summit* are more fluency-oriented, show that the variation observed in the fluency of EFL learners' oral production was to a large extent induced by different course materials.

Table 12. Tests of Between-Subjects Effects for Posttest Scores of Fluency

Dependent Variable	Index	Test period	Type III Sum of Squares	*df*	Mean Square	*F*	*Sig.*	Partial Eta Squared
Fluency	number of dysfluencies	Posttest 1	12.22	2	6.11	3.23	.04	.08
		Posttest 2	12.15	2	6.07	3.26	.04	.08

5. Discussion

The main objective of the current study was to understand how variation in the learners' CAF of oral production is affected by the input they receive from different course materials. In response to research questions one and two, an attempt was made to probe the effect of speaking tasks and input in different course materials on grammatical and lexical complexity of oral production. No difference was observed between the groups in grammatical complexity. This finding is to be interpreted in light of the nature of tasks used in the course materials to elicit oral production, since the results of the micro-evaluation study revealed that mean percentage of speaking tasks with complexity characteristics was almost equal in the three textbooks and they put the same emphasis on complexity of oral production. However, the findings revealed that the *New Headway* group had a better performance in lexical complexity. As an explanation for this findings, a review of textbook evaluation studies (e.g., Roshan, 2013) showed that the contribution of *New Headway* to lexical complexity of oral production might be due to long reading texts with a great deal of new vocabulary. Although detailed studies that focus on the relationship between input and variation in oral CAF are rare, some of the similar studies (Mora & Valls-Ferrer, 2012; Pérez-Vidal & Juan-Garau, 2011) indicated that grammatical complexity only showed a tendency toward significant improvement after learners were exposed to input in a different L2 context. However, other studies showed that exposing learners to different input in different L2 contexts causes a difference in lexical complexity (Tavakoli & Foster, 2008).

Regarding the third research question, the main objective was to investigate the effect of speaking tasks and input with different characteristics in different course materials on EFL learners' accuracy. The findings revealed that the *New Headway* group outperformed the other groups in accuracy as well. A better performance on the part of *New Headway* learners can best be explained by the fact that there is a considerable proportion of speaking tasks in the course with characteristics that encourage L2 learners to produce grammatically and lexically accurate language. Dialogic tasks are associated with greater accuracy, and such effects are due to communication-driven push towards precision, 'creation' of more time to focus on form, as partner is speaking, and recycling of partner's language, both with tendency to re-use correct language and to edit it. Tasks which contain clear inherent structure, especially sequential structure, facilitate task performance by clarifying the macrostructure of the speech event. As a result, the lack of need to engage in large-scale planning frees up attentional resources for on-line planning and higher accuracy. There is also a wealth of research to show that there-and-then tasks (tasks with no contextual support) are associated with greater accuracy because they are more cognitively demanding. In open tasks learners are free to decide on the solution, and this will promote accuracy (for task features, see Ellis, 2003; Skehan, 2001). It should be noted that the findings of this study related to accuracy of oral production are associated with some of the previous studies (Ferrari, 2012; Pérez-Vidal & Juan-Garau, 2011) which similarly showed that exposing learners to different input in different L2 contexts causes a difference in accuracy.

In response to the fourth research question, the effect of speaking tasks and input with different characteristics in different course materials on fluency was investigated. The findings revealed that *American English File* and *Top Notch & Summit* groups did better in fluency. Less proportion of dysfluencies (i.e., higher fluency) during speaking on the part of learners in *American English File* and *Top Notch & Summit* groups can best be explained by the fact that most speaking tasks in these courses have mainly fluency features. Tasks with familiar information will lead to greater fluency, since the easy access to information should make only limited demands on attention, allowing material to be

assembled for speech more easily. The lack of need to engage in large-scale planning in tasks which contain clear inherent structure frees up attentional resources for on-line planning and higher fluency. Here-and-now tasks or tasks with contextual support (a picture, a map, a diagram, etc.) are associated with greater fluency because they are less cognitively demanding. The number of elements and features in a task that need to be manipulated by the speakers will also affect fluency. For example, a story with four females interacting proves more difficult to narrate than a story with only one female and one male character. Tasks that pose a single demand will result in greater fluency. A task that requires learners to describe a route on a map where the route to be taken is marked on the map involves a single task demand and contributes to greater fluency. Closed tasks are those that require learners to reach a single correct solution and are more associated with greater fluency (for task features, see Ellis, 2003; Skehan, 2001). Mora and Valls-Ferrer (2012) and Segalowitz and Freed (2004) are some of the previous studies that similarly found exposing learners to different input in different L2 contexts causes a difference in fluency of oral production.

6. Conclusion and Implications

The overall analysis of the results provides evidence for the impact of course materials on variability in speaking. Some textbooks contribute more to the accuracy of oral production and some others do so for speaking fluency. This varied contribution is due to the fact that there is a considerable proportion of speaking tasks in one textbook which encourage grammatical and lexical accuracy of oral production (e.g., dialogic tasks), while in another textbook there is a wealth of tasks with features that target an improvement in speaking fluency (e.g., here-and-now tasks or tasks with contextual support).

One implication of this study is that EFL materials developers should provide the learners with the experience of speaking tasks with particular features if they want to promote gains in a special dimension of oral performance. Indeed, for them this would mean that ELT materials should provide the learners with dialogic tasks, tasks that need transformations, tasks with no contextual support, tasks with many elements, shared information tasks, tasks that pose a single demand, open tasks with divergent goals, and narrative tasks if the focus is on improving complexity of oral production. These materials should contain dialogic tasks, i.e. tasks with clear inherent structure, tasks with no contextual support, and open tasks if the purpose is accuracy in speaking. Examples of structured tasks are personal information exchange and narrative (Skehan, 2001). Many opinion gap tasks, for example, tasks involving making choices, surveys, debates, ranking activities, and general discussion, are open in nature (Ellis, 2003). ELT textbooks should present the learners with tasks with familiar information, tasks with clear inherent structure, tasks with contextual support, tasks with few elements, tasks that pose a single demand, and closed tasks if the main emphasis is on fluency. Information gap tasks, for example ‘same-or-different’, are typically closed in nature (Ellis, 2003).

The second implication can be for language teaching practitioners who need to be sensitized that a fairer oral communicative test would involve documentation of systematic effects of course material on learners’ variable production of language; otherwise, it can distort the validity of the test scores. For example, while assessing the speaking performance of language learners who have been instructed with a course which gives priority to accurate production of language rather than the ability to convey messages fluently, language educators, in their rating scale, should avoid giving the majority of importance to fluency since it is not fair to assess the learners for what they have not been adequately trained for.

Based on the results of the present study, an intimate understanding of input-related variability in speaking was called for in the realms of research, theory, and practice in SLA. The findings, although significant, have some limitations. The intervention period of this study spanned only two terms (3 months) which was not long enough to discover the actual impact of different task types in different textbooks on variability in oral production. Therefore, a longitudinal study of three different course programs, beginning immediately after beginner learners enter these programs and ending right after they become advanced learners, may possibly provide more complete answers to the complex relationships between task types in the course materials and variability in oral production.

References

ACTFL. (2012). The ACTFL Proficiency Guidelines 2012—Speaking. Retrieved October 10, 2017, from https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012/english/speaking

Barcroft, J., & Wong, W. (2013). Input, input processing and focus on form. In J. R. Herschensohn & M. Young-Scholten (Eds.), *The Cambridge handbook of second language acquisition* (pp. 627-647). Cambridge: Cambridge University Press.

Birdsong, D. (2014). Dominance and age in bilingualism. *Applied Linguistics, 35*(4), 374-392. doi:10.1093/applin/amu031

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Czwenar, I. (2014). Analysing spoken language for complexity, accuracy and fluency: Some methodological considerations. In W. Szubko-Sitarek, Ł. Salski, and P. Stalmaszczyk (Eds.), *Language learning, discourse and communication*, (pp. 81-92). Cham: Springer International Publishing.

de Bot, K., Lowie, W., & Verspoor, M. (2005). *Second language acquisition: An advanced resource book*. London: Routledge.

ETS (2001). *TSE and SPEAK score user guide*. Princeton, NJ: Educational Testing Service.

Ellis, R. (1998). The evaluation of communicative tasks. In B. Tomlinson (Ed.), *Materials development in language teaching* (pp. 217-238). Cambridge: Cambridge University Press.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: 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. doi:10.1093/applin/amp042

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

Faul, F. (2014). G*Power (Version 3.1.9.2). Germany: University of Kiel.

Ferrari, S. (2012). A longitudinal study of complexity, accuracy and fluency variation in second language development. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 277-297). Amsterdam: John Benjamins.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354-375. doi:10.1093/applin/21.3.354

Freed, B. F., Dewey, D. P., Segalowitz, N., & Halter, R. (2004). The language contact profile. *Studies in Second Language Acquisition*, *26*(2), 349-356. doi:10.1017/s027226310426209x

Fulcher, G. (2003). *Testing second language speaking*. *Applied linguistics and language study*. London: Longman.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461-473. doi:10.1093/applin/amp048

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency. Complexity, accuracy, and fluency in SLA* (pp. 1-20). Amsterdam: John Benjamins.

Hughes, R. (2011). *Teaching and researching speaking* (2nd ed.). Harlow: Longman.

IELTS. (2007). *IELTS: International English Language Testing System Handbook.*

Latham-Koenig, C., & Oxenden, C. (2014). *American English file* (2nd ed.). New York: Oxford University Press.

Lee, J.-A. (2009). *Teachers' sense of efficacy in teaching English, perceived English language proficiency, and attitudes toward the English language: A case of Korean public elementary school teachers* (Doctoral dissertation). The Ohio State University, Ohio. Retrieved September 10, 2017, from https://etd.ohiolink.edu/

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). New York, NY: Psychology Press.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, *15*(3), 323-338. doi:10.1093/llc/15.3.323

Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly*, *46*(4), 610-641. doi:10.1002/tesq.34

Munoz, C. (2014). Contrasting effects of starting age and input on the oral performance of foreign language learners. *Applied Linguistics*, *35*(4), 463-482. doi:10.1093/applin/amu024

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555-578. doi:10.1093/applin/amp044

Pallant, J. (2007). *SPSS survival manual: A step by step guide to data analysis using SPSS* (3rd ed.). Maidenhead: Open University Press.

Papajohn, D. (2005). *Toward speaking excellence: The MICHIGAN guide to maximizing your performance on the TSE test and other speaking tests* (2nd ed.). Ann Arbor, MI: University of Michigan Press.

Pérez-Vidal, C., & Juan-Garau, M. (2011). The effect of context and input conditions on oral and written development: A Study Abroad perspective. *International Review of Applied Linguistics in Language Teaching*, *49*(2), 157-185. doi:10.1515/iral.2011.008

Richards, J. C. (2001). The role of textbooks in a language program. Retrieved August 10, 2017, from http://www.professorjackrichards.com/articles/

Richards, J. C. (2014). The ELT textbook. In S. Garton and K. Graves (Eds.), *International perspectives on materials in ELT* (pp. 19-36). New York, NY: Palgrave Macmillan.

Roshan, S. (2013). A Critical Comparative Evaluation of English Course Books in EFL Context. *Journal of Studies in Education*, *4*(1), 172-179. https://doi.org/10.5296/jse.v4i1.4990

Saslow, J. M., & Ascher, A. (2012). *Summit 2* (2nd ed.). White Plains, NY: Pearson Education.

Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, *26*(2), 173-199. doi:10.1017/s0272263104262027

Skehan, P. (1998). Task-based instruction. *Annual Review of Applied Linguistics*, *18*, 268-286. doi:10.1017/s0267190500003585

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, and M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching, and testing* (pp. 167-185). London: Longman.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, *1*(3), 185-211. doi:10.1177/136216889700100302

Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*, *58*(2), 479-473. doi:10.1111/j.1467-9922.2008.00446.x

Thornbury, S. (2009). Methods, post-method, and métodos. Retrieved July 10, 2017, from http://www.teachingenglish.org.uk/article/methods-post-method-m%C3%A9todos

Thornbury, S. (2014). English language teaching textbooks: Content, consumption, production. *ELT Journal*, *69*(1), 100-102. doi:10.1093/elt/ccu066

Tomlinson, B. (2008). Language acquisition and language learning materials. In B. Tomlinson (Ed.), *English language learning materials: A critical review* (pp. 4-13). London: Continuum. doi:10.5040/9781474212182.ch-001.

Verspoor, M., Lowie, W., & de Bot, K. (2008). Input and second language development from a dynamic perspective. In T. Piske & M. Young-Scholten (Eds.), *Input matters in SLA* (pp. 62-80). Bristol: Multilingual Matters.

Verspoor, M., Lowie, W., & van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. *The Modern Language Journal*, *92*(2), 214-231. doi:10.1111/j.1540-4781.2008.00715.x

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. doi:10.1093/applin/24.1.1