RESLA
Revista Española de Lingüística Aplicada

https://matrix.aesla.org.es/RESLA
ISSN 0213-2028 | e-ISSN 2254-6774

Grammatical complexity and lexical diversity in official oral examinations at B2 of the CEFR: the case of FCE and Official Language Schools in Spain

Raúl Azpilicueta-Martínez * & Martin Majercik-Kubjatkova

Public University of Navarra, Pamplona/Iruñea, Spain

Abstract: Official examinations of English for speakers of other languages resort to substantially different tasks in order to assess oral proficiency. However, there is a lack of studies specifically comparing the language features elicited in oral tasks from different official examinations with the same population. This study compares measures of grammatical complexity and lexical diversity in the peer-peer interactions of 10 adult L1-Spanish learners of English performing one task from Cambridge's First Certificate in English (FCE), and a B2 task from the Escuela Oficial de Idiomas (EOI, Spain) speaking tests. The results reveal statistically significant differences both at both grammatical and lexical levels. The ratio of dependent clauses to the total number of clauses was statistically significantly higher in the FCE task. Conversely, the EOI task elicited a significantly more varied lexicon from participants than the FCE task. The study sheds light on the potential of the FCE and EOI tasks to generate comparatively higher levels of grammatical complexity and lexical diversity respectively and highlights the need for further empirical research comparing different oral task samples from official language tests within the same population.

Keywords: *CALF, EFL, FCE, oral proficiency testing, interaction.*

Complejidad gramatical y diversidad léxica en exámenes orales oficiales de nivel B2 del MCER: el caso del FCE y de las Escuelas Oficiales de Idiomas en España

Resumen: Los exámenes oficiales de inglés para hablantes de otras lenguas recurren a tareas sustancialmente diferentes para evaluar la competencia oral. Sin embargo, faltan estudios que comparen específicamente las propiedades lingüísticas elicitadas en tareas orales de distintos exámenes oficiales con la misma población. Este estudio compara medidas de complejidad gramatical y diversidad léxica en las interacciones entre pares de 10 adultos aprendices de inglés con L1 española al realizar una tarea del First Certificate in English (FCE) de Cambridge y una tarea de nivel B2 de las pruebas orales de la Escuela Oficial de Idiomas (EOI, España). Los resultados revelan diferencias estadísticamente significativas tanto en el nivel gramatical como en el léxico. La proporción de oraciones subordinadas respecto del número total de oraciones fue estadísticamente mayor en la tarea del FCE. Por el contrario, la tarea de la EOI elicitó de los participantes un léxico significativamente más variado que la del FCE. El estudio arroja luz sobre el potencial de las tareas del FCE y de la EOI para generar, respectivamente, niveles comparativamente más altos de complejidad gramatical y diversidad léxica, y subraya la necesidad de continuar la investigación empírica que compare distintas tareas orales de exámenes oficiales dentro de la misma población.

Palabras clave: *CALF, inglés como lengua extranjera (EFL/ILE), FCE, evaluación de la competencia oral, interacción.*

How to cite: Azpilicueta-Martínez, R., & Majercik-Kubjatkova, M. (2026). Grammatical complexity and lexical diversity in official oral examinations at B2 of the CEFR: the case of FCE and Official Language Schools in Spain. *Revista Española de Lingüística Aplicada, 39*(1), 212-232. https://doi.org/10.58859/resla.867

*Corresponding author: raul.azpilicueta@unavarra.es

AESLA
Asociación Española de Lingüística Aplicada

1. Introduction

The assessment of oral proficiency has been rightly labelled as "an extremely difficult skill to test, as it is far too complex (a skill) to permit any reliable analysis to be made for the purpose of objective testing" (Heaton, 1988, p. 146). In contrast with the testing of other linguistic skills, the nature itself of the assessment of Foreign / Second Languages (FL/SL) speaking seems to present significant challenges having to do with its reliability, validity, being live and requiring the presence of an examiner, as well as cost and time-efficiency considerations (Foot, 1999).

Perhaps consequently, speaking tests from ESOL (English for Speakers of Other Languages) institutions worldwide resort to different testing formats, or tasks, with which to assess oral proficiency. Interestingly, a growing body of research (e.g., Foster & Tavakoli, 2009; García-Ponce & Tavakoli, 2022; Gilabert, 2007; Ortega, 1999; Skehan, 2009) has been providing increasing empirical evidence of how task variations at different levels, e.g. kind of gap within the task, task goal, linguistic requirements or task complexity, can potentially affect the language output of same-level learners significantly. In other words, different oral tasks can provide different levels of performance in different oral measures. At the same time, different authors have warned about the excessive importance granted to grammatical measures (grammatical accuracy and complexity) in oral test ratings, as mixed results have questioned their validity to consistently differentiate between adjacent proficiency levels (e.g. Iwashita et al., 2008; Seedhouse et al., 2014). This fact has raised awareness about the need to resort to more reliable indicators of oral proficiency, such as vocabulary knowledge (Hsieh & Wang, 2019; Iwashita et al., 2008; Kuiken et al., 2019; Yan et al., 2020) at the same time it begs for further research in the field, since a high number of the existing studies have been based on subjective ratings from test performances, that is, not on empirical data from language proficiency interview transcripts (Iwashita et al., 2008).

At the same time, there is a predominance of monologic task-based studies (e.g., De Jong & Perfetti, 2011; Tavakoli, 2011), and, with a few exceptions (e.g. Foster & Skehan, 1996; García-Ponce & Tavakoli, 2022), research on dialogic task types and oral proficiency in the L2 remains limited. Similarly, studies on the effect of task type on oral performance in an L2 have tended to focus on task types often resorted to by teachers and researchers (García-Ponce & Tavakoli, 2022), and not on oral tasks from officially ESOL exams. This is important because testing formats have proven to exert a direct washback effect on the teaching that takes place prior to their being implemented (Heaton, 1988; Perrone, 2010).

In light of the above, there is a need for specific studies which i) compare the oral features triggered by different official ESOL tests, ii) use dialogic tasks iii) collect data from the same population, and iv) analyse empirical data by means of transcripts rather than more or less subjective ratings. The present exploratory study attempts to shed some light on the above lacunae by comparing grammatical complexity and lexical diversity

measures elicited by two different dialogic oral tasks pertaining to two officially recognised ESOL language testing institutions in Spain with the same population, using computational analysis with transcribed interactions.

2. Task type and oral performance

Oral proficiency tasks can be categorised depending on different criteria, including performance conditions (e.g., availability of planning time, type of interlocutor), the demands they place on test takers, as well as their specific design features (Hsieh & Wang, 2019). Some of the best-known inventories include Prabhu's (1987), who established three main categories depending on the kind of gap presented, comprising information-gap tasks, reasoning-gap tasks and opinion gaps. Ellis (2009), by contrast, classifies tasks depending on their linguistic requirements, and draws a line between focused tasks, i.e. tasks in which the focus lies on "using some specific linguistic feature" (p. 223) and unfocused tasks, in which there is no explicit focus on a particular language feature. Finally, Pica et al. (1993) classified tasks depending on their goal, and differentiates between divergent or convergent tasks. Divergent tasks do not involve task performers to reach a common agreement, whereas convergent tasks imply the learners reaching a shared consensus.

The relationship between task features and conditions, and measures of oral proficiency has been the focus of a growing body of research since the 1980s (e.g., De Jong & van Ginkel, 1992; Kormos, 2014; McNamara, 1990; Skehan, 1998; Michel et al., 2019). Research on the effect of task variations and oral performance have provided mixed results. On the one hand, there are those studies in which little or no significant differences have been reported. Iwashita et al. (2001), for example, compared different task performance conditions (including familiarity of task information, abstractness of information and nature of the operation required) for the same task type (an individual narrative task) and reported no significant influence in the candidates' task performance. In a a smaller pilot study, Tuzcu & Yalçın (2019) analysed different task conditions (careful online planning and pressured online planning) and also reported no significant differences between groups. Similarly, a study by Brown et al. compared two independent and three integrated TOEFL iBT speaking tasks and reported little differences between tasks (Brown et al., 2005).

On the other hand, there are studies revealing how task variations at different levels, e.g. kind of gap within the task, task goal, linguistic requirements or task complexity can affect the language output of same-level learners significantly (Foster & Skehan, 1996; Gilabert, 2007; Kuiken & Vedder, 2007). García-Ponce et al. (2018) analysed complexity, accuracy and fluency measures during uncontrolled paired interactions in three tasks, namely personal information, narrative and negotiation tasks and highlighted how, concurring with previous research (Skehan, 1998; Tavakoli & Foster, 2011), different tasks may imply different load levels on learners' attention, therefore "causing the learner to make choices on being complex, being accurate and/or being fluent" (p. 84).

A study by Hsieh & Wang (2019) investigated the impact of two task types from the TOEFL Junior Test (a picture narration task and an integrated listen/speak task, both individual, not paired) on measures of grammar, vocabulary, content and fluency, in young learners, and their results revealed an impact on all variables with the exception of fluency. Their findings also suggested that the effect of task type on performance was most marked on lexical measures, and that, regarding grammatical complexity, "test takers tended to mimic the sentence they heard in the input material of the listen/speak task" (p. 44), yielding a comparatively more complex discourse than the picture narration task, since the language in that input material was more complex than the one produced by the young learners in their study.

In view of the mixed findings described, it is clear that more research is needed in order to shed light on the effect that specific task variations might have on the oral performance of same-level learners.

3. Grammatical and lexical measures in oral proficiency testing

One of the most commonly accepted frameworks of linguistic measurements in research today is CALF[1] (former 'CAF'), comprising complexity, accuracy, lexis and fluency (Tavakoli, 2018). Within this framework, scholars have looked into a myriad of aspects with which to analyse and measure oral language performance (e.g., Biber & Gray, 2013; Pallotti, 2009). The connection between CALF and ESOL oral proficiency testing is noticeable by the presence of grammatical, lexical and fluency performance indicators in rating scale repertoires across oral proficiency tests, of which the two first constitute particularly central elements (*Cambridge English: Understanding Results Guide*, 2014). Within these, the grammatical aspect has outweighed other elements (vocabulary, fluency, pronunciation) and has acted as the strongest determinant factor for global scores among raters across proficiency levels in oral testing (Higgs & Clifford, 1982; Iwashita et al., 2008; McNamara, 1990), with exceptions like the English for Academic Purposes (EAP) test (Brown et al., 2005).

In a review of the existing literature, Iwashita et al. (2008) highlighted that, across different proficiency levels, "grammatical accuracy is the principal determining factor for raters assigning a global score" (p. 27), with a more variable specific weight of other measures (e.g. vocabulary, pronunciation, fluency or appropriateness). The same study (Iwashita et al., 2008) analysed the differences between five different individual oral tasks across five proficiency levels at the the TOEFL iBT exams and relied on a range of aspects including grammatical complexity and vocabulary measures. Grammatical complexity, that is, the degree to which production in the TL reflects grammatically complex and advanced structures (Richards, 2015), was operationalised via the *T-unit complexity ratio*, the *Dependent clause ratio* and *Verb phrase complexity*, while meausures for vocabulary included an analysis of tokens (i.e., words) and types (i.e., range). Their results revealed that the grammatical turned out to be a measure which did not always help differentiate between

[1] Although the terms "CALF" has come to replace "CAF" in recent years, the latter term has been maintained when making reference to studies in which "CAF" (i.e. not including lexis) was analysed.

adjacent proficiency levels, that is, the lowest-level students in their study did not always produce the least complex grammar. In fact, results showed no significant differences in increasing complexity across levels for the *T-unit complexity ratio* and the *Dependent clause ratio*, while differences for *Verb phrase complexity* were only significant when related to the number of utterances produced, and yet, their effect size was "marginal" (p. 37). Vocabulary measures, however, revealed a significant linear increase across levels, with medium effect sizes. The study's findings highlight how fluency, pronunciation, and lexical measures like number of tokens might constitute more reliable aspects than grammatical aspects when it comes to characterising distinct levels of proficiency. The authors criticise "an exaggerated emphasis" on the specific weight of the grammatical aspect, also "in the attitudes and behaviour of many learners and teachers", and underline the need to pay more attention to "production features and vocabulary knowledge" (p. 47). Seedhouse et al. (2014) assessed grading criteria in IELTS using CALF measures and did report a linear correspondence between grammatical accuracy (measured as errors per 100 words) and score, whereas higher grammatical complexity proved to be an less reliable oral proficiency indicator, since, in their study, grammatical complexity rates at lower proficiency levels (low C1 of the CEFR) were higher than those at C1-C2 of the CEFR, i.e., there was not a "clear linear progression" throughout the bands (p. 22). However, lexical measures in the study included number of words and these revealed a direct linear correspondence with proficiency levels (p. 22), consistent with previous findings (Brown, 2006a). Yan et al. (2018) analyzed the oral production of speakers performing an individual oral task from the APTIS computer-based test, and reported how both lexical and grammatical complexity did increase with the levels associated with the CEFR. The scope of their study allowed them to establish a mapping to the CEFR, in which more proficient speakers produce more subordination and lexical variety. This result was confirmed by Kuiken et al. (2019). A linear correspondence between measures of grammatical (including accuracy and complexity) and lexical measures and oral proficiency was also reported in Hsieh and Wang (2019), in a study comparing two task types from the TOEFL Junior test with young learners. In conclusion, whilst lexical measures seem to associate more closely with variations across oral tasks and proficiency levels, grammatical measures have provided mixed results, what highlights the need for more research in the field.

It is worth mentioning that the vast majority of the studies hitherto mentioned have been carried out with monologic tasks frequently used by teachers and researchers, since, with a few exceptions (e.g. Foster & Skehan, 1996; García-Ponce & Tavakoli, 2022), research with interactive task types from oral proficiency tests remains particularly scant. This is important, since tests have gradually incorporated dialogic or group interaction as regular tasks in their repertoire (Galaczi & Taylor, 2018), especially in the upper half of the Common European Framework of Reference, or CEFR (Azpilicueta-Martínez, 2017). To complicate matters further, researchers have highlighted the need for more research on oral proficiency testing based on empirical data, and not on more or less subjective ratings (Iwashita et al., 2008). Finally, there is a conspicuous lack of studies specifically comparing language features generated by different ESOL official tests with the same population.

In light of the body of research previously examined, the following research questions will be addressed:

1. What are the similarities and differences in the grammatical complexity of adult EFL learners while performing the FCE and the EOI interactional tasks?
2. What are the similarities and differences in the lexical diversity of adult EFL learners while performing the FCE and the EOI interactional tasks?

4. Method

4.1 Participants

A total of 10 adults participated in the study, six of whom were female and four male. Participants were enrolled in the EOI[2] B2 blended learning group, in which they attended a 150-minute session once a week. Due to the Covid-19 pandemic, lessons were entirely online, and had a focus on the oral production skill. Access to the course was granted by having an official CEFR B1 certificate, or by passing an entry-level examination for the B2 carried out. All participants but one whose L1 was Portuguese, shared Spanish as their L1. The age of the participants ranged between 29 and 45 (mean age: 37). They had a similar linguistic background, that is, they had stopped learning EFL after their compulsory education, and had decided to take on the blended course in order to update their command of English.

4.2 The tasks

The tasks in the study were part of two public official samples, namely Cambridge's FCE, and EOI, available on their respective official websites. Due to the specific [it does not look good to talk about limited scope, as it may be interpreted as having little value] scope of the study, only one task from each of these tests was selected. The tasks chosen were the peer-peer interaction parts given the positive findings of this layout reported in recent research, including the eliciting of a) a wider range of features of interaction (Azpilicueta-Martínez, 2017; Brooks, 2009; Ducasse, 2008; Ducasse & Brown, 2009, 2011; Galaczi, 2004, 2014) and b) less "institutional" or formulaic talk than monologic tasks (Van Lier, 1989), in addition to the information provided in section 2.1.

The FCE task selected (see Appendix 1) corresponds to Part 3, the "two-way collaborative task, which involves the two test-takers in a two-way discussion" (Galaczi, 2008, p. 93). This text-prompted task has a duration of four minutes including the presentation of the task. The participants' production is expected to last approximately three minutes. The prompt consists of a small concept map with a central question (ideas to attract more tourists to the town) and seed ideas for the speakers to discuss and agree on. The FCE task was managed following following Cambridge's script, which sets the time limit, the way each task has to be presented, as well as the tester's intervention depending on the participants' production, ensuring equal treatment for all examinees.

[2] The Escuela Oficial de Idiomas is one of the most important ESOL institutions in Spain, including 449 local offices across the country: https://t.ly/NCXls

The EOI subtask (see Appendix 2), referred to as "co-production"[3], is also a text-prompted activity which lasts between four and five minutes, including the presentation. The prompt is a card which includes the premise (travelling abroad with an NGO, in this case) plus four bullet points related to be topic. The premise and three of the bullet points are identical for both participants, yet one of the bullet points is different. The similarities and differences between both tasks, including the classifications described in the literature review, can be noted in Table 1.

Table 1. Differences between the FCE and the Spanish EOI tasks.

Commonalities		
Element	**FCE / EOI**	
Kind of gap	Opinion/reasoning gap	
Task goal	Convergent	
Monologic/Dialogic	Dialogic	
Tester interference	Minimal	
Differences		
Element FCE		**EOI**
Expected production time	3 minutes	4-5 minutes
Linguistic requirement	Semi-focused: make suggestions (using second conditional, included in prompt)	Unfocused
Number of sections	Two • Discussion on why the ideas presented might lead to a specific outcome • Jointly choose one idea out of five	One • Discussion on the possibility of performing a specified action
Preparation time	Discussion: 15 seconds Common choice: one minute	Discussion: one minute
Visibility of written instructions for participants	No	Yes, in same paragraph as central issue
Text layout	Concept map	Paragraph + Bullet points
Point of discussion worded as	Question (*why would these ideas attract...?)*	Description (*a friend you both have in common...)* + Statement (*have a conversation about the possibility...)*
Length of written prompt	Total number of words: 26	Total number of words: 72 (participant A), 74 (participant B)
	Point of discussion words: 10	Point of discussion words: 47
	Related points: 2-5 words each	Related points: 2-9 words each
Number of specific points to cover	5 (all closed)	4 (3 closed; one open)
Assignment of specific points to each participant	No	Yes
Assignment of roles	No	Yes Participant A starts conversation Participant B closes conversation

[3] Appendix: Guía de Examen de Pruebas de Certificación de las Escuelas Oficiales de Idiomas. Nivel intermedio B2.

4.3 Procedure

Participants undertook both tasks consecutively in a single session, as two online oral activities in their English language classes. The FCE task was undertaken first, and the EOI task was carried out immediately afterwards.

Students were randomly paired for each of the tasks. In both cases the instructions for the tasks were explained to them, as well as the time allocated for a) preparation and b) the task itself. Participants were given the chance to ask questions prior to performing the task in order to ensure they fully understood their roles and what was expected of them. In the case of EOI, each dyad was free to choose which participant would start (Candidate 'A') the interaction and who ('Candidate 'B') would finish it.

The FCE and EOI tests comprise different guidelines as for the examiner's participation or support to candidates in the exams, which were carefully followed by the researcher.

4.4 Data coding and analysis

Data collection took place between 6th and 8th April 2020 in the participants' regular online classes. The instrument used for video and audio recording was the *Blackboard Collaborate* software, which was integrated within the EOI's online *Moodle* platform. In order to avoid the much-criticised use of subjective ratings, the present study resorted to computational analysis of written transcripts.

Grammatical complexity was examined using the *Web-based L2 Syntactical Complexity Analyzer,* a tool previously used in written-performance analysis and more recently incorporated in oral performance-based studies (e.g. Hu, 2021). We examined the most useful measures of complexity according to Wolfe-Quintero et al. (1998), as followed by relevant CAF-based studies:

a. Number of clauses per T-unit (or *T-unit complexity ratio*), in which clause is defined as "a production unit containing either a subject and a finite verb or a subject and a finite or non-finite verb form" (Mylläri, 2020, following Lu, 2011, p. 44; Wolfe-Quintero et al., 1998, p. 70). A T-unit refers to "one main clause with all subordinate clauses attached to it" (Hunt, 1965, p. 20).

b. Number of verb phrases per T-unit (or *Verb-phrase ratio*), in which the verb phrase is made of a main verb alone, or a main verb plus any modal and/or auxiliary verbs.

c. Ratio of dependent clauses to the total number of clauses (or *Dependent clause ratio*). Dependent clauses refer to groups of words that contain a subject and verb but do not express a complete thought, as in "I did not do the homework *because I felt asleep"* (dependent clause).

Following Iwashita et al. (2008), lexical diversity was analysed using the *Web Vocabprofile* online tool. This software provides different measurements from the BNC/COCA 29 word family lists, of which 25 are based on frequency and range data, and US lists based on Mark Davies' (Brigham Young University) 450-million-word Corpus of Contemporary American English (2012) (Cobb, n.d.; Davies, 2010; Nation, 2018). Frequency ratios are calculated based on the number of words in each frequency list. We examined the proportions of low and high frequency vocabulary displayed by participants, as well as the diversity in the word families present in each of the different frequency levels.

In order to avoid a misrepresentation of low-frequency words caused by participants using the words provided in the prompts, all the lexical items (including singular or plural derivatives) in the participants' prompts which did not belong to the family list of most frequently used words in the corpora above were eliminated from the data. The lexical items eliminated were the following:

FCE: attract; cameras; nightclub; providing, tourists.

EOI: accompanying; benefits; collaboration; common; enriching; professional.

All interactions were transcribed by one of the researchers into .txt format and subsequently checked by the other researcher in the study. Inter-transcriber reliability reached 99.86%. The statistical analysis was conducted using SPSS 24. Due to the small sample sizes, the non-parametric Wilcoxon signed-rank test was used.

5 Results

This section will begin by presenting the results of the three measures analysed in the first research question, that is, the grammatical complexity present in the FCE and EOI tasks, and will be followed by an analysis of the results regarding lexical diversity, our second research question.

Mean scores obtained in relation to the grammatical complexity revealed higher rates for the EOI task in two of the measures, namely the *T-unit complexity ratio* (C/T) and the *Verb-phrase ratio* (VP/T), although differences were non-significant (z = -0.663, p = .508 for C/T; z = -1.070, p = .285 for VP/T). However, the *Dependent clause ratio* (DC/C) in the FCE task did show a significantly higher rate than the EOI task (z = -2.803, p = .005). This is illustrated in Table 2.

Table 2. Mean values and statistical difference between FCE and EOI regarding grammatical complexity.

	FCE	EOI	*Z* value	*p*-value
C/T	22.55	24.18	-0.663	.508
VP/T	29.89	34.13	-1.070	.285
DC/C	0.47	0.46	-2.803	.005

Note. * = statistical significance (p < .05).

The following examples (1, 2), might help illustrate the significantly higher *Dependent clause ratio* (DC/C) rates in the FCE task by providing the two first turns by one of the participants in both tasks, in what appeared to be a common pattern among participants:

(1) FCE

1. Participant 1: *Erm… Yes, I agree because at the end of the day tourists have to sleep somewhere, so, if there are not enough hotels, you need to have an amount of holiday flats, so I think it is a good idea, and also I think in order to promote an economy it is a good idea to have more shops, tourists erm… can go there to spend their money and also some shops like fashion shops would be a kind of places, interesting places to visit like for example in Madrid.*
2. Participant 2: *Erm… I agree, I agree with you because as I told you before, erm… to have erm… people to come to your city to visit it you need places where these people can sleep, so without hotels, erm… holiday flats you cannot receive people. So, for me it is the most important thing.*

(2) EOI

1. Participant 1: *Okay I am thinking to go with Michael. Michael is our colleague that works for an NGO organisation and I am thinking to go with him to Nicaragua in order to help to the most poor people around the country, helping to build schools.*
2. Participant 2: *Yes, I am thinking going with him because, erm… I like a lot to help other people. Mainly people who have no money to survive and they need out help.*

Note how the sentences in participant 2's turns in the FCE task might include a higher *Dependent clause ratio* due to their being noticeably longer than the ones in EOI, a fact which might have been triggered by the nature of the task itself, that is, giving reasons for or against particular items.

Our second research question attempted to compare the lexical diversity in both tasks, presented in Table 3. Since the EOI task implied a longer duration on the participants' production, the statistical analysis was made by comparing the percentage of each of the frequency levels (ranging from K-1, most frequently used, to K-5, less frequently used) in relation to the total number of tokens in each task.

In general terms both tasks were consistent in displaying an inversely proportional gradient between the frequency level and the number of tokens shown by participants. In other words, the percentage rates decreased as complexity of frequency levels increased, with the exception of the higher number of K-5 than K-4 tokens in the FCE task. The participants did not resort to any K-5 tokens in the EOI task. In descriptive terms, both tasks displayed a clear predominance of K-1 frequency level tokens, even more so in the case of the EOI, where a significant difference over the FCE task was observed (z = -1.988,

p = .047), which might relate to the EOI's overall higher number of tokens. The other statistically significant difference took place at the K-3 level (z = -1.960, p = .050), again with higher values for the EOI task. The rest of scores showed higher rates for the FCE task at the K-2 frequency level, and higher rates for the EOI task in the K-4 frequency level.[4]

Table 3. Values and statistical difference between FCE and EOI regarding lexical diversity.

	FCE	**EOI**		
Total number of tokens (mean)	121	154.4		
Frequency level (% over total)	FCE	EOI	*Z* value	*p*-value
K-1	88.74	92.28	-1.988	.047
K-2	3.68	3.05	-1.274	.203
K-3	0.33	1.08	-1.960	.050
K-4	0.1	0.18	-0.535	.593
K-5	0.3	0	-1.604	.109

Note. * = statistical significance (p < .05).

However, the analysis of frequency levels alone might fail to provide a complete picture of the lexical diversity displayed by participants in each task. If we analyse the number of different word families within each frequency level provided by the software, we are given a very different reading. With the exception of the K-5 frequency level, the EOI task yielded a substantially higher number of word families used, as illustrated in Figure 1.

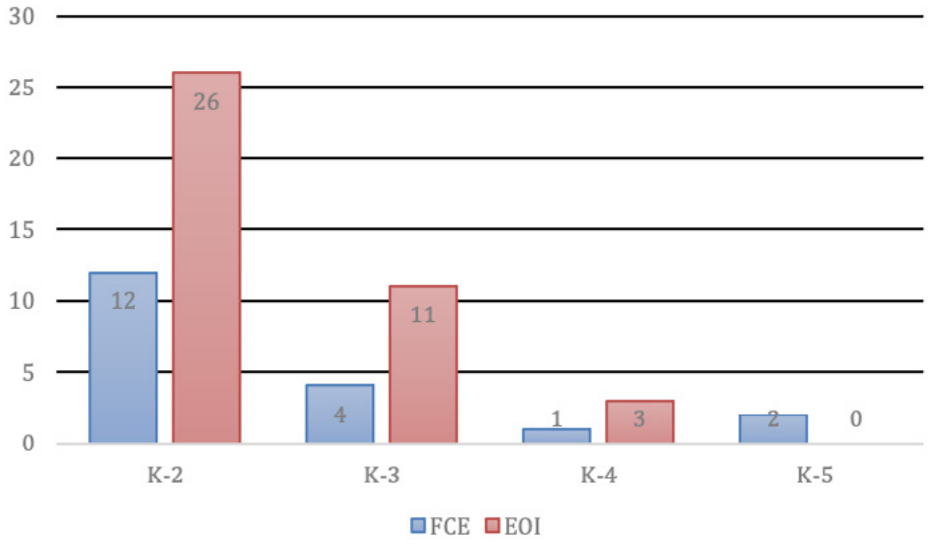


Figure 1. FCE and EOI tasks. Diversity among word families within each frequency level.

[4] Proper nouns do not fit any frequency level in the *Web Vocabprofile* online tool. That is the reason why the addition of all percentages in each task does not add up to 100%.

With the aim of providing a clear reading of the graphic, the K-1 frequency level was not included, although the number of families at that level were also higher for the EOI task (154 word families for the FCE and 195 for the EOI task). This is further illustrated in Table 4, which attempts to provide a more qualitative approach by providing the exact word families and the number of tokens present at level K-2 in each task, with the EOI task eliciting more than twice as many word families as the FCE task. It is worth noting that this phenomenon was not at the expense of EOI data yielding fewer tokens from each word family, since the mean average of instances of the same token was strikingly similar, that is, 3.83 for the FCE task and 3.70 for the EOI task. However, this fact might have been mediated by the EOI being longer, that is, eliciting a larger amount of production from participants.

Table 4. FCE and EOI tasks. K-2 word families and number of tokens (shown in square brackets).

FCE	EOI
BNC-COCA-K2k Families: 12. 1. develop_[1] 2. earn_[1] 3. economy_[1] 4. entertain_[2] 5. example_[4] 6. fashion_[1] 7. hotel_[7] 8. mix_[2] 9. opinion_[6] 10. option_[7] 11. receive_[1] 12. tour_[13]	BNC-COCA-K2k Families: 26 1. active_[6] 2. advantage_[2] 3. advice_[2] 4. attend_[2] 5. boss_[4] 6. career_[2] 7. convince_[2] 8. encourage_[2] 9. example_[2] 10. fantastic_[2] 11. flight_[2] 12. future_[2] 13. increase_[4] 14. opinion_[2] 15. opportunity_[6] 16. option_[2] 17. organize_[12] 18. prefer_[2] 19. profession_[6] 20. project_[16] 21. receive_[4] 22. risk_[4] 23. salary_[2] 24. social_[2] 25. suffer_[2] 26. survive_[2]

The researchers were granted access to the results of the FCE and EOI B2 oral tests from the 2019 examination call at the Centro Superior de Idiomas (CSI) at a Public University in Spain and at one national Official Language School (EOI) in Spain, respectively.

Regarding the oral part of the test, the 2019 FCE results at the CSI displayed a 90,58% pass marks (154 out of 170 administered tests). The 2019 EOI B2 results at the EOI centre revealed an 80,1% pass marks (318 out of 397 administered tests), that is, there were over 10% more candidates getting a 'pass mark' on the oral section of the FCE test. However, it is worth reminding that the tasks in the present study only comprised the interactional subsection from each of the oral section in each test.

6 Discussion and conclusion

The present study compared a number of measures of grammatical complexity and lexical diversity in the oral production of learners performing two oral interaction tasks from two official language tests at level B2 of the CEFR: Cambridge's FCE and EOI in Spain.

Results have revealed significant differences in some of the measures from both grammatical complexity and lexical diversity. Regarding grammatical complexity, the FCE task specifically yielded a significantly higher rate in the *Dependent clause ratio* (DC/C) than the EOI task. Conversely, the EOI task included significantly higher vocabulary rates than the FCE task at the K1 and K3 frequency levels. This was further ratified by substantial differences in the word families within those levels.

These results allow us to draw two main conclusions. Firstly, although more research is needed, it could be hypothesised that differences in the nature of the tasks might significantly affect aspects of the test-takers' performance regarding oral grammatical complexity, such as the *Dependent clause ratio* (DC/C). Previous research has questioned the validity of grammatical complexity to distinguish between adjacent proficiency levels (e.g., Iwashita et al., 2008). In other words, the same speakers might be producing more dissimilar grammar complexity across two different tasks than speakers from different proficiency levels performing tasks of a similar nature. This finding concurs with the notion that different tasks may imply different load levels on participants' attention, thus pushing them "to make choices on being complex, being accurate and/or being fluent" (García-Ponce et al., 2018: 84). This is not to say that the FCE task necessarily generates a more grammatically complex production among participants than the EOI task overall, since the EOI task provided higher values for the other two measures under analysis (i.e., the *T-unit complexity ratio* (C/T) and the *Verb-phrase ratio* (VP/T) than the FCE task, although such differences were not significant.

Secondly, the results regarding lexical diversity reveal that the EOI task yielded a significantly higher rate than the FCE task at two frequency levels. Results are validated by the fact that both tasks were consistent in their provision of a decreasing gradient of tokens across frequency levels. This finding supports previous research associating variations in language diversity rates to changes in tasks or proficiency levels (e.g., Hsieh & Wang, 2019; Iwashita et al., 2008; Yan et al., 2018).

The findings in the present study are expected to raise awareness among testing institutions about the potential of different task formats to elicit significantly different language features. Similarly, these results provide evidence of the positive potential washback effect of specific official exam formats for specific pedagogical purposes (e.g, the EOI task as an interesting format as a vocabulary activator). We believe that a better understanding of the ways in which different tasks from official exams promote specific language features would constitute an important pedagogical tool which would go beyond mere exam preparation. Conversely, this study also casts doubt on the validity of the grammatical complexity aspect as a reliable indicator of language proficiency, since, at least one of its measures seems to hinge to a great degree on the nature of the task at hand.

The study ultimately strives for more empirical research comparing language samples, rating scales and scoring from interactive oral tasks in official language tests. It also advocates for the use of instruments commonly used in research, such as the *Web-based L2 Syntactical Complexity Analyzer* or the *Web Vocabprofile* online tools as valuable means for ESOL institutions when it comes to standardising and analysing language samples with their examining staff.

Limitations

An obvious methodological issue concerning the current study lies in its limited number of language samples (20), so its findings should be taken with caution, since further studies with larger pools of participants are required in order for these findings to be validated. Another limitation is related to order effect, since the fact that the tasks were undertaken in a particular order (FCE first, EOI second) might have had an impact on performance, although it is important to emphasize that that the choice of that particular order was due to the small pool of participants. In other words, the inherent limitation of sample size would have not eliminated the order effect had the tasks been administered in different orders (e.g., two dyads doing FCE first, three dyads doing EOI first).

Similarly, topical knowledge is known to affect performance (Bachman & Palmer, 1996), and the fact that each task revolved around a particular topic, namely a "Holiday resort" (FCE) and "collaboration with an NGO" (EOI) might have also impinged on the lexical choices made by participants. More research with similar task formats but different topics would be needed in order to confirm or refute the impact of the task factor in the tendencies reported in the present study.

Acknowledgements

We would like to thank Laura Escribano Asín and Alazne Ciarra Tejada, as well as the B2 students that participated in this study, and Centro Superior de Idiomas at Universidad Pública de Navarra. We would also like to thank the reviewers for their time and expertise, which have significantly enriched the quality of our work.

CRediT Author contribution / *Contribución de los autores*

Conceptualization / ***Conceptualización:*** Azpilicueta-Martínez and Majercik.
Formal Analysis / ***Análisis formal:*** Azpilicueta-Martínez and Majercik.
Methodology / ***Metodología:*** Azpilicueta-Martínez.
Writing / ***Redacción:*** First draft: Majercik; last draft: Azpilicueta-Martínez.
Research dataset / ***Datos de investigación:*** Majercik.

Funding, data availability, and copyright / *Financiación, disponibilidad de datos y derechos de autoría*

Funding / ***Financiación:*** No funding / *No se ha recibido financiación.*
Conflict of interest / ***Conflicto de intereses:*** The authors have no conflicts of interest to declare. Both co-authors have seen and agree with the contents of the manuscript and there is no economic interest to report. We hereby certify that the submission is original work and is not under review at any other publication / *Los autores declaran no tener ningún conflicto de intereses. Ambos coautores han revisado el contenido del manuscrito y están de acuerdo con él, y no hay ningún interés económico que declarar. Por la presente certificamos que el trabajo presentado es original y no se encuentra en proceso de revisión en ninguna otra publicación.*
Data availability statement/ ***Declaración de disponibilidad de datos:*** The two tasks (FCE and EOI) used to collect data in the study are open access and available for the public at: (FCE): https://www.cambridgeenglish.org/Images/CER_6168_V1_APR19_Cambridge_English_First_Handbook_WEB_v3.pdf (Accesed July 2022). (EOI): https://www.educacion.navarra.es/web/dpto/idiomas-plurilinguismo/escuelas-oficiales?inheritRedirect=true (Accesed July 2022) / *Las dos tareas (FCE y EOI) utilizadas para recopilar datos en el estudio son de acceso libre y están a disposición del público en: (FCE): https://www.cambridgeenglish.org/Images/CER_6168_V1_APR19_Cambridge_English_First_Handbook_WEB_v3.pdf (consultado en julio de 2022). (EOI): https://www.educacion.navarra.es/web/dpto/idiomas-plurilinguismo/escuelas-oficiales?inheritRedirect=true (Consultado en julio de 2022).*

Editorial history / *Fechas del proceso editorial*

Received / ***Recibido:*** 14/07/2022
Accepted / ***Aceptado:*** 06/05/2024
Published / ***Publicado:*** 01/04/2026

References

Azpilicueta-Martínez, R. (2017). *Negotiation for meaning and assessment of oral proficiency through paired interactive tasks: Evidence from EFL children and adults at beginner levels of competence* [Doctoral dissertation, Universidad Pública de Navarra]. Academica-e.unavarra.es.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests (Vol. 1)*. Oxford University Press.

Biber, D., & Gray, B. (2013). Discourse characteristics of writing and speaking task types on the TOEFL iBT® test: A lexico-grammatical analysis. *ETS Research Report Series*, *2013*(1), 1-128. https://doi.org/10.1002/j.2333-8504.2013.tb02311.x

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing*, *26*(3), 341-366. https://doi.org/10.1177/0265532209104666

Brown, A. (2006a). Candidate discourse in the revised IELTS Speaking Test. *IELTS Research Reports*, *6*, 71-89. IELTS Australia and British Council.

Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks. *ETS Research Report Series*, *2005*(1), 1-157. https://doi.org/10.1002/j.2333-8504.2005.tb01982.x

Byrnes, H. (1987). Proficiency as a framework for research in second language acquisition. *The Modern Language Journal, 71*(1), 44-49. https://doi.org/10.1111/j.1540-4781.1987.tb01054.x

Cambridge English: Understanding results guide. (2014). Cambridge English: Understanding results guide. http://www.gml.cz/prof/zajickova/Cambridge%20exams_information/Understanding%20results%20guide.pdf

Cobb, T. (n.d.). *Web Vocabprofile* [An adaptation of Heatley, Nation, & Coxhead's (2002) Range]. Retrieved January 2021, from http://www.lextutor.ca/vp/

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing, 25*(4), 447–464. https://doi.org/10.1093/llc/fqq018

De Jong, J. H., & Van Ginkel, L. W. (1992). 15 dimensions in oral foreign language proficiency. In L. Verhoeven & J. de Jong (Eds.), *The construct of language proficiency* (pp. 187-207). John Benjamins. https://doi.org/10.1075/z.62.19jon

De Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, *61*, 533-568. https://doi.org/10.1111/j.1467-9922.2010.00620.x

Ducasse, A. M. (2008). *Interaction in paired oral proficiency assessment in Spanish* [Unpublished doctoral dissertation]. School of Languages and Linguistics, Faculty of Arts, The University of Melbourne.

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, *26*(3), 423-443. https://doi.org/10.1177/0265532209104669

Ducasse, A. M., & Brown, A. (2011). The role of interactive communication in IELTS speaking and its relationship to candidates' preparedness for study or training contexts. *IELTS Research Reports*, *12*, 1. https://www.ielts.org/for-researchers/research-reports/volume-12-report-3

Ellis, R. (2009). Task-based language teaching sorting out the misunderstandings. *International Journal of Applied Linguistics*, *19*, 221-246. https://doi.org/10.1111/j.1473-4192.2009.00231.x

Foot, M. C. (1999). Relaxing in pairs. *ELT Journal, 53*(1), 36–41. https://doi.org/10.1093/elt/53.1.36

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, *18*(3), 299-323. https://doi.org/10.1017/S0272263100015047

Foster, P., & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language Learning*, *59*(4), 866-896. https://doi.org/10.1111/j.1467-9922.2009.00528.x

Galaczi, E. D. (2004). *Peer-peer interaction in a paired speaking test: The case of the First Certificate in English* [Unpublished doctoral dissertation]. Columbia University.

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly*, *5*(2), 89-119. https://doi.org/10.1080/15434300801934702

Galaczi, E. D. (2014). *Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? Applied Linguistics*, *35*(5), 553-574. https://doi.org/10.1093/applin/amt017

Galaczi, E. D., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, *15*(3), 219-236. https://doi.org/10.1080/15434303.2018.1453816

García-Ponce, E. E., & Tavakoli, P. (2022). Effects of task type and language proficiency on dialogic performance and task engagement. *System*, *105*, 102734. https://doi.org/10.1016/j.system.2022.102734

García-Ponce, E. E., Mora-Pablo, I., Lengeling, M. M., & Crawford, T. (2018). Task design characteristics and EFL learners' complexity, accuracy and fluency during uncontrolled pair interactions: A naturalistic perspective. *Iranian Journal of Language Teaching Research*, *6*(1), 75-92.

Gilabert, R. (2007). Effects of manipulating task complexity on self-repairs during L2 oral production. *IRAL: International Review of Applied Linguistics in Language Teaching*, *45*, 215-240. https://doi.org/10.1515/iral.2007.010

Heaton, J. B. (1988). *Writing English language tests*. Longman Inc.

Higgs, T. V., & Clifford, R. (1982). The push towards communication. In T.V. Higgs (Ed.), *Curriculum, competence, and the foreign language teacher* (pp. 57-79). National Textbook Company.

Hsieh, C. N., & Wang, Y. (2019). Speaking proficiency of young language students: A discourse-analytic study. *Language Testing*, *36*(1), 27-50. https://doi.org/10.1177/0265532217734240

Hu, X. (2021). Predicting CEFR levels in L2 oral speech, based on lexical and syntactic complexity. *Asia Pacific Journal of Corpus Research, 2*(1), 35–45. https://doi.org/10.22925/apjcr.2021.2.1.35

Hunt, K. W. (1965). *Grammatical structures written in three grade levels (Research Report No. 3). National Council of Teachers of English*. https://files.eric.ed.gov/fulltext/ED113735.pdf

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24-49. https://doi.org/10.1093/applin/amm017

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning*, *51*(3), 401-436. https://doi.org/10.1111/0023-8333.00160

Kormos, J. (2014). *Speech production and second language acquisition.* Routledge. https://doi.org/10.4324/9780203763964

Kuiken, F., & Vedder, I. (2007). Task complexity and measures of linguistic performance in L2 writing. *IRAL: International Review of Applied Linguistics in Language Teaching*, *45*(3), 261–284. https://doi.org/10.1515/iral.2007.012

Kuiken, F., Vedder, I., Housen, A., & De Clercq, B. (2019). Variation in syntactic complexity: Introduction. *International Journal of Applied Linguistics*, *29*(2), 161-170. https://doi.org/10.1111/ijal.12255

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36-62. https://doi.org/10.5054/tq.2011.240859

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing*, *7*(1), 52-76. https://doi.org/10.1177/026553229000700105

Michel, M. C., Révész, A., Shi, D., & Li, Y. (2019). The effects of task demands on linguistic complexity and accuracy across task types and L1/L2 speakers. In Wen & Ahmadian (Eds.), *Researching L2 task performance and pedagogy* (pp. 133-151). John Benjamins. https://doi.org/10.1075/tblt.13.07mic

Mylläri, T. (2020). Words, clauses, sentences, and T-units in learner language: Precise and objective units of measure? *Journal of the European Second Language Association*, *4*(1), 13-23. https://doi.org/10.22599/jesla.63

Nation, I. S. P. (2018, July 5). Information on the BNC/COCA word family lists [Unpublished technical document]. Victoria University of Wellington. https://www.victoria.ac.nz/__data/assets/pdf_file/0004/1689349/Information-on-the-BNC_COCA-word-family-lists-20180705.pdf

Ortega, L. (1999). Task-based language teaching: Sorting out the misunderstandings. *Studies in Second Language Acquisition*, *21*(1), 109-148. https://doi.org/10.1017/S0272263199001047

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590-601. https://doi.org/10.1093/applin/amp045

Perrone, J. M. (2010). *The impact of the First Certificate in English (FCE) examination on the EFL classroom: A washback study* [Unpublished doctoral dissertation]. Columbia University.

Pica, T., Kanagy, R., & Falodun, J. (1993). Choosing and using communication tasks for second language research and instruction. In G. Crookes & S.M. Gass (Eds.), *Tasks and second language learning* (pp. 9-34). Multilingual Matters.

Prabhu, N. S. (1987). *Second language pedagogy*. Oxford University Press.

Richards, J. C. (2015). *Key issues in language teaching*. Cambridge University Press. https://doi.org/10.1017/9781009024600

Seedhouse, P., Harris, A., Naeb, R., & Üstünel, E. (2014). The relationship between speaking features and band descriptors: A mixed methods study. *IELTS Research Reports Online Series, 30*. https://eprint.ncl.ac.uk/208149

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press. https://doi.org/10.1177/003368829802900209

Skehan, P. (2009). Modeling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(3), 510-532. https://doi.org/10.1093/applin/amp047

Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers. *ELT Journal*, *65*(1), 71-79. https://doi.org/10.1093/elt/ccq020

Tavakoli, P. (2018). L2 development in an intensive study abroad EAP context. *System*, *72*, 62-74. https://doi.org/10.1016/j.system.2017.10.009

Tavakoli, P., & Foster, P. (2011). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*, *61*, 37-72. https://doi.org/10.1111/j.1467-9922.2011.00642.x

Taylor, L. (2000). Investigating the paired speaking test format. *University of Cambridge ESOL Examinations Research Notes*, *2*, 14-15.

Tuzcu, A., & Yalçın, Ş. (2019). The combined effects of manipulating tasks in two dimensions on L2 speech performance. In 2017 *Second Language Research Forum. Cascadilla Proceedings Project* (pp. 175-184).

Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*, *23*(3), 489-508. https://doi.org/10.2307/3586922

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. University of Hawaii Press.

Yan, X., Kim, H. R., & Kim, J. Y. (2018). *Complexity, accuracy and fluency features of speaking performances on Aptis across different CEFR levels*. https://www.britishcouncil.org/sites/default/files/yan_et_al_b.pdf

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*, *0*(00), 1-26.

Appendix 1

FCE task. Examiner's view

21 Holiday resort

Part 3 4 minutes (5 minutes for groups of three)
Part 4 4 minutes (6 minutes for groups of three)

Part 3

Interlocutor Now, I'd like you to talk about something together for about two minutes. *(3 minutes for groups of three).*

I'd like you to imagine that a town wants more tourists to visit. Here are some ideas they're thinking about and a question for you to discuss. First you have some time to look at the task.

*Place **Part 3** booklet, open at **Task 21**, in front of the candidates. Allow 15 seconds.*

Now, talk to each other about **why these ideas would attract more tourists to the town.**

Candidates *2 minutes (3 minutes for groups of three)* .....

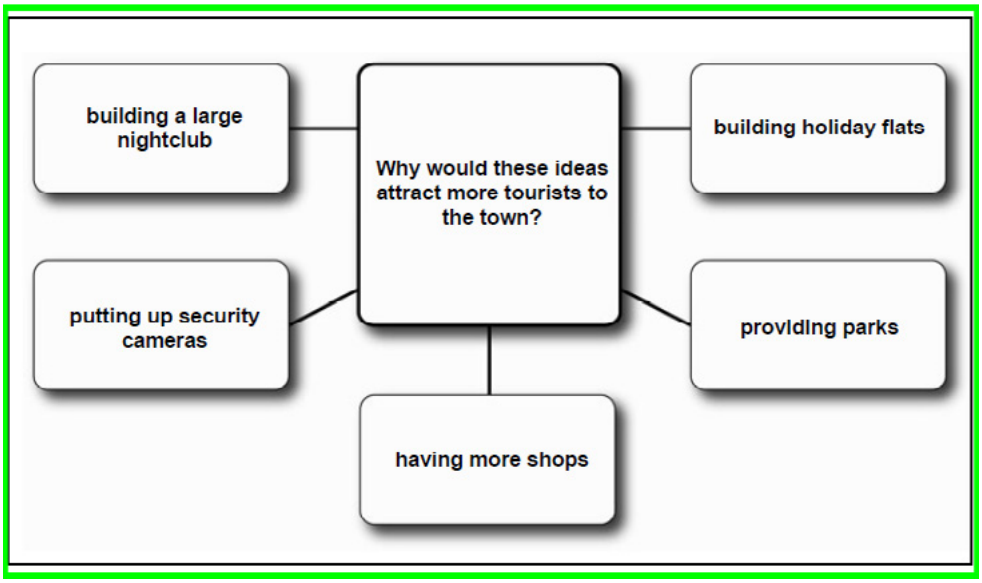
Interlocutor Thank you.

Now you have about a minute to decide **which idea would be best for the town.**

Candidates *1 minute (for pairs and groups of three)* .....

Thank you. (Can I have the booklet, please?) *Retrieve **Part 3** booklet.*

FCE task. Candidates' view



Appendix 2

EOI task. Examiner's and candidates' view

INTERACTION.

(**EXAMINER'S INSTRUCTIONS READ ALOUD**). *A friend, you both have in common, who works for an NGO in a country from the so-called "third world" has asked for your collaboration for 2 months in the Summer. Have a conversation about the possibility of accompanying him/her. Refer to the cues on your card.*

CANDIDATE A.

- An enriching experience.
- Helping others.
- Professional benefits of taking part in such an experience.
- Anything else you want to add?

CANDIDATE B.

- An enriching experience.
- Helping others.
- People in your local area need help.
- Anything else you want to add?