Analysing the oral performance of EFL learners in the testing and the laboratory contexts

Beatriz Santana Perera
Patricia Arnaiz-Castro
Universidad de Las Palmas de Gran Canaria

Received: 2022-05-04 / Accepted: 2023-02-24
DOI: *https://doi.org/10.30827/portalin.vi41.24590*
ISSN paper edition: 1697-7467, ISSN digital edition: 2695-8244

ABSTRACT: The differential effects of planning on the performance of English-as-a-second-language learners in the laboratory and the testing context have been discussed extensively. However, research which compares data from both contexts is scant. The present study aimed at examining and comparing the impact of the testing and laboratory contexts on learners' performance. For this purpose, two groups of Spanish intermediate learners of English as a foreign language (60 in total) were asked to narrate a story based on a sequenced set of pictures under a careful online planning condition, which is the condition that has received the least attention in the testing context. An analysis in terms of complexity, accuracy, lexis, and fluency revealed statistically significance differences just in the fluency parameter, specifically in long pauses. Nevertheless, a clear trend towards the testing context was observed in all dimensions. The research findings will be of particular interest to practitioners trying to design oral tasks that faithfully adhere to the existing requirements for EFL production in both contexts. The paper ends with a discussion of possible reasons for the findings and suggests avenues for further research.
Keywords: careful online planning, testing context, laboratory context, EFL, oral performance

Análisis de la producción oral de aprendices de ILE en contextos de examen y laboratorio

RESUMEN: El impacto de la planificación en el desempeño de los estudiantes de inglés como segunda lengua tanto en el contexto de laboratorio como en el de examen ha sido objeto de numerosas investigaciones. Sin embargo, son escasos los estudios que comparan datos de ambos contextos. El objetivo del presente estudio fue examinar y comparar el impacto del contexto de laboratorio con el de examen en el desempeño de aprendices españoles de inglés como lengua extranjera (ILE). Para ello, se pidió a dos grupos de estudiantes españoles de nivel intermedio de inglés (60 en total) que narraran una historia ilustrada en una secuencia de imágenes en condiciones de planificación en tiempo real, que es la condición que menos atención ha recibido en contexto de exámenes. El análisis de medidas de complejidad, precisión, léxico y fluidez mostró diferencias estadísticamente significativas en un solo parámetro de fluidez; el de las pausas largas. Sin embargo, se observó una clara tendencia de mejora en el contexto de examen en todas las variables analizadas. Estos hallazgos pueden resultar útiles para los profesionales que desean diseñar tareas orales que se ciñan fielmente a los requisitos actuales para la producción de ILE en ambos contextos. El artículo finaliza con una discusión de las posibles razones de los hallazgos y sugiere vías para futuras investigaciones.

Palabras clave: planificación en tiempo real, contexto de examen, contexto de laboratorio, ILE, producción oral

1. Introduction

Research findings in task-based language teaching (TBLT) have consistently proven that task planning has beneficial effects on the second language (L2) learners' oral performance dimensions such as complexity, accuracy, fluency (Bui, Skehan, & Wang, 2018; Ellis, 2005) and lexis (Santana-Perera & Arnaiz-Castro, 2018). Task planning has been operationalized under different conditions, namely pre-task planning, the planning that occurs before the task is performed, i.e. learners are given time to either prepare or rehearse the task before its performance; and online planning, the planning that takes place during the task performance, i.e. learners are instructed to perform the task within a given limited time (pressured online planning), or are provided with unlimited time to perform the task (careful online planning) (Ellis, 2005).

The majority of task-based planning studies have drawn on information processing theories that claim that humans have limited processing capacity (Anderson, 1995; VanPatten, 2002). In L2 learning research, the trade-off hypothesis posits that learners, especially those with a low proficiency level, may have even more difficulties in attending both to meaning (i.e. the content of learners' speech) and form (i.e. the quality of the language they produce) and need to decide how to allocate their attentional resources by prioritizing performance in certain aspects of language at the expense of others (Skehan, 2014; Skehan & Foster, 2007; Yuan & Ellis, 2003). Furthermore, the study of online planning has been framed within Levelt's (1989) speech production model, which is a process consisting of three underlying stages –conceptualization (within which the conceptual content of speech is generated), formulation (when the needed lexicon and syntactic and grammar structures are retrieved to encode the content), and articulation (when the speaker transforms the content into overt speech). Another major incorporated component of speech production is self-monitoring, through which oral output is controlled by constantly self-reviewing both internal and external speech (Levelt, Roelofs & Meyer, 1999).

Regarding L2 learners' linguistic performance, pre-task planning has proved to be beneficial notably to complexity, fluency (Tavakoli & Skehan, 2005), and lexis (Santana-Perera & Arnaiz-Castro, 2018). Since careful online planning provides language learners with additional time, they may be more capable of overcoming processing and attentional limitations. The findings have yielded evidence that careful online planning leads to focusing on form, which results in enhanced accuracy (Panahzadeh & Asadi, 2019; Saeedi, 2020) and complexity (Ahmadian, 2012a, 2012b) in oral performance, but they also have proven to be detrimental to lexical achievement (Santana-Perera & Arnaiz-Castro, 2018; Santana-Perera, 2020) and fluency (Ahmadian, 2012a, 2012b; Atai & Nasiri, 2017; Ahmadian, Tavakoli, & Dastjerdi, 2015; Saeedi, 2020; Yuan & Ellis, 2003).

2. Literature Review

2.1. Planning in a testing context

As seen above, a preponderance of studies has provided strong evidence of the benefits of planning on second language speech production. Given that tests should encourage test takers to produce their best possible performance (O'Grady, 2019; O'Sullivan, 2012), it stands to reason that planning should be included in testing processes. Furthermore, if language productions in tests are expected to be representative of the wide variety of real-world language, both planned and unplanned conditions need to be considered (Wigglesworth & Elder, 2010). However, and as asserted also by Li, Chen and Sun (2015), not much research has focused on the setting where planning takes place.

Pre-task planning has been the planning condition researchers in the language testing field have devoted their efforts the most, but in contrast to the results obtained in TBLT studies, the data obtained in language testing are inconsistent. Another difference to be considered between these two areas lies in the parameters traditionally employed to measure performance. As described above, TBLT has used the discourse analytic performance measures of complexity, accuracy, lexis, and fluency, whereas language testing has relied on raters' summative scores. The former, described as "objective, quantitative and verifiable" measures (Housen, Kuiken & Vedder, 2012, p. 2), identifies and records variations more transparently than the latter. Raters' judgments, even those made by trained raters, are inevitably biased by personal impressions (Nitta & Nakatsuhara, 2014).

The first author to examine the effects of planning on the test scores awarded to narrative tasks was (Wigglesworth (1997), who explored the effects of pre-task planning on the oral production of learners taking the access (Australian assessment of communicative English skills) test. The results are different for the two measures used. For discourse level measures, the planning time was beneficial, but for the scores given by raters it was not. The results in the study by (Iwashita, McNamara & Elder., 2001), on the contrary, do not vary from measure to measure. The 193 Asian-born pre-university and university EFL students were required to perform narrative tasks based on picture prompts. Neither the results of the analysis of discourse features nor the scores assigned to candidates' test performances by trained raters were higher for the planned condition. These findings mirror those in Wigglesworth and Elder's (2010) study. The two authors explored the effects of pre-task planning time on the performance of 90 intermediate and advanced level candidates in the International English Language Testing System oral interview. No differences were identified in their performance after the analysis of either the score or the discourse measures. It should be noted, however, that in neither of these studies did the amount of planning time exceed three minutes, whereas, as O'Grady indicates, "the most common amount of planning time in TBLT research is 10 minutes". More recently, O'Grady (2019) looked at the effect different lengths of pre-task planning time had on the performance of 47 Turkish students in their university entrance exam and observed that the extra pre-task planning time conditions (5 minutes and 10 minutes) led to significantly better scores. As in the previous studies mentioned, no differences were detected based on the scale used.

Other studies have examined the effect on test performance of strategic planning using discourse analytic measures exclusively. Tavakoli and Skehan (2005) found in their study with 80 elementary and intermediate adult female learners studying English at an educational association in Tehran, Iran, that the 5-minute planning time given for the narrative task based on picture series led to much better accuracy and better complexity and fluency than the non-planning condition. The success of accuracy over the other two parameters was also observed in the study of similar characteristics undertaken by Li et al. (2015) with 95 intermediate-level ESL Chinese university students. According to the authors, this finding may be justified by the fact that an assessment setting may encourage students to focus on accuracy more than a classroom or laboratory context.

To our knowledge, the only study that has focused on online planning in a testing context is the one conducted by Panahzadeh and Asadi (2019). The authors measured the performance of 14 intermediate EFL female students from a private language institute in Tehran, Iran, in the classroom and the testing context. The results obtained from the analysis of learners' production based on given topics showed that in the two contexts under study, unpressured online planning led to significantly higher scores in grammatical range and accuracy than pressured online planning. On the other hand, and although statistical analyses were not conducted to compare the results of both contexts, the authors detected that in the testing context, learners used simpler lexical items than in the classroom context.

In light of the gap in the literature on online planning in testing contexts, the current study focused on the performance of two groups of university EFL learners in a careful online planning task while being assessed. The data obtained from the analysis were compared with the data obtained from the performance of a similar task in a laboratory context.

Furthermore, this study is the result of the authors' concern on two issues. On the one hand, we have evidence in the literature that oral exams that will affect end-of-course grades may lead to an increase in learners' anxiety levels, affecting their attention span and therefore making an impact on the quality of their output (Arnaiz-Castro & Pérez-Luzardo, 2014; Hewitt & Stephenson, 2012; Horwitz, 2010, 2017; O'Grady, 2019; Pérez Castillejo, 2019; Salehi & Marefat, 2014).

On the other hand, we have the laboratory setting, which is the setting very often used to measure oral performance, and where learners are often reassured that their performance will only be used for research purposes and will not count towards their final course grade (see, for example, (Saeedi, 2020). Therefore, and since as researchers and teachers we tend to draw upon both settings, we found it worthwhile to explore the differences in the performance of learners in both settings to find to which degree each setting exerted a relevant influence on their performance.

3. Method

3.1. Research questions

Based on the results of the research previously discussed, and given that the present study aimed at comparing the performance of participants who planned and performed the oral task in a testing setting with the performance of participants who planned and performed

the oral task in a laboratory in terms of complexity, accuracy, lexis, and fluency (CALF), the following question was formulated:

- Are there significant differences in terms of CALF between the oral task performance of Spanish FL students under the careful online planning condition in the testing setting and the oral task performance of Spanish FL students under the careful online planning condition in the laboratory setting?

3.2. Participants and settings

The participants were 60 Spanish intermediate English as a Foreign Language (EFL) learners (39 females and 21 males). They were all enrolled in the Faculty of Educational Sciences at the University of Las Palmas de Gran Canaria. They were either in their first or in their third academic year of the undergraduate degree in Teacher Training or Social Education. The participants ranged in age from 18 to 23 years, and the mean length of exposure to the English language in a classroom setting was 11.5 years. They were all in a foreign language context with few opportunities to practice the English language outside the classroom. Thirty of the participants in this study were randomly assigned as Group A, who planned and performed the task in a testing setting, and the other thirty were assigned as Group B, who planned and performed the same task in a laboratory setting.

Taking into consideration some authors' conclusions that advanced learners may not benefit from planning in terms of complexity (Kawauchi, 2005), fluency (Yuan & Ellis, 2003), or lexical richness (Nielson, 2014), we decided to select intermediate proficiency learners. For the selection of participants, 120 students were required to take the official PET Cambridge test (level B1 according to the Common European Framework of Reference for Languages (CEFR))[1]. Each participant's EFL class level and their English learning history were also considered for the selection. To ensure that the participants constituted a fairly homogeneous group in terms of their English proficiency, only the participants who achieved a score higher than 8 (on a scale of 0–10) were selected for this study. None of the participants had performed the type of oral narrative task in this study before, nor had they ever planned an oral task under careful online condition.

The recordings were made over a period of one month, whenever the participants were available. They all signed written informed consent forms. Participants were audio recorded with the Audacity (R): Free Audio Editor and Recorder (Version 2.0.6; Audacity Team, 2014) and their performances saved as MP3 files.

3.3. Task and task conditions

A tight narrative structured task based on picture compositions (sets of coherently structured picture prompts) was used for both groups to collect data. This type of task was chosen as previous studies have shown that clear and tightly structured tasks (as in pic-

[1] The official Cambridge test consists of four reading exercises, four writing exercises, four listening exercises, and two speaking exercises. It took two hours and twenty minutes for the participants to complete the four parts to the test.

ture-based narrative tasks) have the potential to enhance the output of learners in terms of CALF and therefore be conducive to the development of a second language (Ahmadian et al, 2015; Tavakoli & Foster, 2008; Wang & Skehan, 2014). The picture set was taken from the English learning text book by Soars & Soars (1993) and was the one used for the careful online planning task by Santana-Perera (2020) and Santana-Perera and Arnaiz-Castro (2018).

The task required participants to narrate orally the story in one unique session. To that end, participants received undetailed guided instructions. They were required to carry out the task after seeing the pictures for 50 seconds but were given unlimited time to formulate and monitor their speech plans as they performed the task. The task story had a set of 11 picture prompts and was about a man who had a serious height complex and visited a therapist to get advice.

3.4. Design and Procedure

The participants were first divided into two groups. One group of 30 students constituted Group A and another group of 30 students conformed Group B. The participants in Group A were told that the study was being carried out for foreign language research purposes and that their performance would not have any repercussion on their test scores, whereas the participants in Group B were aware that the task was the speaking part of their final test. The information provided to all the participants was related to the task (i.e., the set of pictures). Also, they were given a written introductory sentence for the picture set, not only for the participants to use as an icebreaker but also to encourage the use of the given verb tense. The task instructions, given in Spanish, were:

- You will be given a set of picture prompts which tell a story. Please watch the set of pictures for 50 seconds and narrate the story orally immediately after. Imagine that you are telling the story to someone who has not seen it and is very eager to know all the details, so be as detailed as possible in narrating the story. You may take as much time as you need to complete the task. Therefore, if you notice that you have made a mistake, either grammatically or syntactically, you may repair your errors. Likewise, if you remember any word or expression that you should have used but you did not use, you may go back and reformulate your narration.

3.5. Measures

Measures of complexity, accuracy, and fluency have been used extensively as dependent variables to evaluate speech performance (e.g. Foster & Skehan, 1996; Ahmadian et al., 2015; Kawauchi, 2005; Ahmadian & Tavakoli, 2011; Nielson, 2014; Ortega, 1999; Tavakoli & Foster, 2008; Tavakoli & Skehan, 2005; Yuan & Ellis, 2003). In this study, we considered a fourth measure, namely lexis, as this represents a form of complexity that needs to be assessed as an independent dimension (Bui et al., 2018).Taking into consideration this need and following Santana-Perera and Arnaiz-Castro (2018), the measures indicated below were selected and used to assess the CALF dimensions.

3.5.1. Dependent variables: CALF measures

- **Complexity**
  Syntactical complexity: amount of subordination: the ratio of clauses to AS-units in the participants' oral production (see Foster et al., 2000 for a rationale behind choosing AS-units). Incomplete sentences were excluded.
- **Accuracy**
  Percentage of error-free clauses: errors relating to prepositions, pronouns, word order, comparative adjectives, subject omission, and lexical choice were considered. Lexical errors were counted when the words were inappropriate or did not exist in English. Errors were counted only once even though they were repeated throughout the narration.
  Percentage of correct verb forms: tense, aspect, modality, and subject-verb agreement were considered. The use of historical present was not considered an error in this study. In both cases, errors were excluded where the learners succeeded in self-repair.
- **Lexis**
  Lexical sophistication: defined as the appropriate use of low- frequency vocabulary items (Malvern, Richards, Chipere & Durán, 2004). The measure of lexis was operationalized using the lexical computational tool VocabProfile. The total number of words produced by each participant was inserted in VocabProfile, and the number of less frequent words was obtained from considering the second 1000-word list and the subsequent ones.
- **Fluency**
  Breakdown fluency: medium pauses (between 3 and 4 seconds) and long pauses (more than 4 seconds).
  Repair fluency: number of self-repairs, number of reformulations, number of hesitations, number of repetitions, and number of false starts.

3.6. Data Analysis

The audio recordings were transcribed, coded, segmented, and scored by both researchers (first separately and then, a week later, together to ensure that the coding, segmentation, and scoring were conducted reliably) to collect data for the statistical analysis in terms of the above-defined four oral production measures of CALF. The scores were entered into SPSS version 25.0 and checked in terms of normality of distribution via skewness and kurtosis indices. The alpha for achieving statistical significance was set at .05. Unpaired t-tests were then run.

4. Results

The results of each of the four language dimensions analysed will be reported separately for better clarity.

4.1. Complexity

Syntactical subordination was assessed to measure the complexity of the language used by the participants in their oral productions. Table1shows that the participants who planned the task in the testing setting (i.e., Group A) performed better than the participants that planned the task in the laboratory setting (i.e., Group B), although the difference between the two groups did not reach statistical significance (p=.240).

Table 1. *Statistics for Complexity under the Two Planning Settings*

Variable	Planning Setting	N	Mean	SD	Min.	Max.	Sig.
Syntactical complexity	Testing (Group A)	30	156.911	63.107	33.33	237.50	.240
	Laboratory (Group B)	30	114.475	49.171	43.48	190.00	

4.2. Accuracy

Two variables were used to measure accuracy: percentage of error-free clauses and percentage of correct verb forms. As shown in Table 2, again, Group A obtained slightly better results in terms of error-free clauses with no statistically significant differences (p=.943). Likewise, a greater percentage of correct verb forms were observed in the oral productions of the participants who performed in the testing setting, although their scores did not yield statistically significant differences either (p=.078).

Table 2. *Statistics for Accuracy under the Two Planning Settings*

Variable	Planning Setting	N	Mean	SD	Min.	Max.	Sig.
Error-free clauses	Testing (Group A)	30	79.232	12.780	35.29	97.30	.943
	Laboratory (Group B)	30	78.145	11.055	55.88	94.87	
Correct verb forms	Testing (Group A)	30	78.892	9.994	59.09	96.88	.078
	Laboratory (Group B)	30	72.137	17.071	34.33	87.14	

4.3. Lexis

Lexical sophistication was the variable used to measure lexis. The results exhibited in Table 3 show that the participants who planned and performed the task in the testing setting produced a more lexically sophisticated output. Nonetheless, the difference regarding the performance of participants in Group B was not significant (p=.760).

Table 3. *Statistics for Lexis under the Two Planning Settings*

Variable	Planning Setting	N	Mean	SD	Min.	Max.	Sig.
Sophistication	Testing (Group A)	30	3.622	1.227	0.97	5.97	.760
	Laboratory (Group B)	30	2.665	1.071	1.03	4.39	

4.4. Fluency

Breakdown fluency and repair fluency were measured separately. Breakdown fluency variables were medium and long pauses, and repair fluency variables were hesitations, repetitions, reformulations, false starts, and self-repairs. As can be observed in Table 4, the participants in the laboratory setting scored a greater number of medium pauses(p=.983) than the participants in the testing setting, although the difference was not statistically significant. Likewise, the number of long pauses was higher in the laboratory setting, but the difference this time was statistically significant (p=.009).

As for repair fluency, Group A proved to produce a greater number of disfluencies than Group B in all the variables, except in false starts, although once again, no statistical differences were found for any of the repair fluency measures in this study (see Table 5).

Table 4. *Statistics for Breakdown Fluency under the Two Planning Settings*

Variable	Planning Setting	N	Mean	SD	Min.	Max.	Sig.
Medium pauses	Testing (Group A)	30	0.604	0.678	0.00	2.40	.983
	Laboratory (Group B)	30	1.140	1.010	0.00	4.14	
Long pauses	Testing (Group A)	30	0.159	0.348	0.00	1.06	.009
	Laboratory (Group B)	30	0.613	0.749	0.00	2.46	

Table 5. *Statistics for Repair Fluency under the Two Planning Settings*

Variable	Planning Setting	N	Mean	SD	Min.	Max.	Sig.
Hesitations	Testing (Group A)	30	15.113	8.850	2.65	40.00	.094
	Laboratory (Group B)	30	8.370	4.564	2.76	17.06	
Repetitions	Testing (Group A)	30	7.189	4.757	0.26	22.22	.547
	Laboratory (Group B)	30	4.866	3.159	0.55	10.51	
Reformulations	Testing (Group A)	30	2.173	1.806	0.00	6.87	.083
	Laboratory (Group B)	30	1.440	1.130	0.00	4.48	
False starts	Testing (Group A)	30	0.185	0.314	0.00	1.20	.568
	Self-repairs	30	0.379	0.373	0.00	1.06	
Self-repairs	Testing (Group A)	30	2.738	1.482	0.49	6.87	.884
	Laboratory (Group B)	30	2.283	1.378	0.00	5.00	

5. Discussion and Conclusions

This study sought to investigate the performance of Spanish EFL learners in an oral task under careful online planning conditions in two different settings, namely, the testing and the laboratory setting. To that end, the complexity, accuracy, lexis, and fluency of their oral narrative speech was measured. The only statistical difference identified in the comparison of learners' performance was found within the fluency variable, specifically in the long pauses. Nevertheless, there was a clear tendency in all the results that deserves to be discussed.

In terms of complexity and accuracy, the testing setting allowed for more syntactic subordination, error-free clauses, and correct verb forms than the laboratory setting. As Pan-

ahzadeh & Asadi (2019) remarked, unlimited time for task performance encourages learners not only to conceptualize their production in a way that other task conditions do not but also to formulate higher quality output, considering they can monitor it while performing. Besides, and along the same lines, Li et al. (2015) pointed out the difference in learners' language behaviour in a testing context from that in a non-testing setting, where fluency and complexity are usually the parameters with higher scores. In this regard, the authors argued that this superiority of accuracy might be of interest for test designers when selecting assessment criteria. In the case of our study, it is fundamental to bear in mind that learners were aware that it would be the last attempt of their utterances that would count towards the final mark, and, consequently, the pressure of their end-of-course grade may have been diminished by the fact that they had several opportunities within one.

With respect to lexis, likewise, the level of lexical sophistication was higher in the testing setting. This result is in contrast with the result in Panahzadeh and Asadi's study (2019) in which learners used simpler lexical items in the testing context than in the classroom context. The authors justify this result by explaining that the learners might not have wanted to take risks. Also, they measured the use not only of a wide vocabulary resource but also of idioms. As for fluency, once again it is the testing context where learners performed better, making fewer medium pauses and significantly fewer long pauses than in the laboratory setting. It is without doubt surprising that it is precisely in the fluency parameter that a statistically significant difference was identified, especially if we take into account the study by Pérez Castillejo (2019) on anxiety and fluency in a final exam context. Pérez Castillejo's findings reveal the strong impact of anxiety on the pause frequency in learners' performance. The reason for the difference in results might be attributed to the fact that in Pérez Castillejo's study, learners had to complete a narrative without planning time. Again, the fact that in our case there was no time pressure might have been crucial. It can be said that the findings of the current study broadly correspond to claims made by Wigglesworth and Elder (2010) that giving test takers opportunities for planning may make them feel more self-confident.

There is little doubt that producing output without any time pressure is a different process from producing output either with time pressure or without strategic planning. If we want tests to recreate as much as possible the cognitive processes that occur when undertaking similar tasks in an academic context (for example, when making an oral presentation) or in real-life situations (Wigglesworth and Elder, 2010), online planning should be incorporated. Furthermore, and due to the critical role of oral language in EFL learning, it is of paramount importance to create conditions that encourage learners to show the best possible version of their output (Swain, 1985) not only in class but also in an exam setting. After all, as stated by (Zhang, 2019, p. 776), "performance is one of the most important outcomes of FL learning".

The need to develop more learner adequate assessment has been underlined by several authors (Gursoy & Arman, 2016; O'Sullivan, 2012; Swain, 1985). Keeping in mind that exam-oriented education systems still abound, and many teachers and students rely mostly on test results when determining the success of the learners (Huang, Chang, Zhi, et al. 2020), the findings of the current study have relevant implications for language teachers and test designers.

Some limitations of the study, and therefore areas of future research, should be noted. First, the participants in the research were 60 students enrolled in the Faculty of Educational Sciences at the University Las Palmas de Gran Canaria. Future research should be conducted with learners in other faculties and with higher-ability learners as language proficiency has been demonstrated to be a key variable. Another limitation is that data were collected from just one task-type. It would be helpful to conduct a study of similar characteristics with a task that required different cognitive abilities on the part of the candidates, for example, a decision-making task. Finally, although measuring anxiety levels was not within the scope of this study, it is true that this affective factor might have had an impact on learners' attitude. In fact, the literature in the language learning arena presents plenty of evidence that having speaking ability assessed in a final oral exam that can be viewed by learners as a high-stakes test is naturally anxiety-provoking (Hewitt & Stephenson, 2012; Horwitz, 2010, 2017; Pérez Castillejo, 2019; Salehi & Marefat, 2014), and that anxiety may affect their scores. Future researchers could administer an anxiety scale and try to find out, for example, whether learners in the testing setting suffered from facilitating anxiety, which has been shown to improve language performance (Hewitt & Stephenson, 2012) and hence aids acquisition.

6. References

Ahmadian, M. J. (2012a). The effects of guided careful online planning on complexity, accuracy and fluency in intermediate EFL learners' oral production: The case of English articles. *Language Teaching Research*, *16*(1), 129–149.https://doi.org/10.1177/1362168811425433

Ahmadian, M. J. (2012b). The relationship between working memory capacity and L2 oral performance under task-based careful online planning condition. *TESOL Quarterly*, *46*(1), 165–175.

Anderson, J. R. (1995). *Learning and memory: an integrated approach*. John Wiley & Sons.

Arnaiz, P., & Pérez-Luzardo, J. (2014). Anxiety in Spanish EFL university lessons: causes, responsibility attribution and coping. *Studia Anglica Posnaniensia*, *49*(1), 57–76.

Atai, M. R., & Nasiri, M. (2017). An investigation into the effects of joint planning on complexity, accuracy, and fluency across task complexity. *Journal of English Language teaching and Learning*, *20*, 49–74.

Audacity Team (2014). Audacity(R): Free audio editor and recorder (Version 2.0.6) [Computer program]. Retrieved from http://audacity.sourceforge.net/.

Bui, G., Skehan, P., & Wang, Z. (2018). Task condition effects on advanced level foreign language performance. In P.A. Malovrh, A. Benati (Eds.), *The handbook of advanced proficiency in second language acquisition*, (pp.219–237). John Wiley & Sons. https://doi.org/10.1002/9781119261650.ch12.

Council of Europe. Council for cultural co-operation. Education committee. Modern Languages Division (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

Ellis, R. (1987). Interlanguage variability in narrative discourse: style shifting in the use of the past tense. *Studies in Second Language Acquisition*, *9*(1), 1–19.

Ellis, R. (2005). *Planning and task-based performance*. John Benjamins Publishing Company.

Foster, P., & Skehan, P. (1996). The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition*, *18*(3), 299–324.

Gursoy, E., & Arman, T. (2016). Analyzing foreign language test anxiety among high school students in an EFL Context (Note 1). *Journal of Education and Learning*, *5*(4), 190–200.

Hewitt, E., & Stephenson, J. (2012). Foreign language anxiety and oral exam performance: A replication of Phillips's MLJ study. *The Modern Language Journal*, *96*(2), 170–189.

Horwitz, E. K. (2010). Foreign and second language anxiety. *Language Teaching*, *43*(2), 154–167.

Horwitz, E. K. (2017). On the misreading of Horwitz, Horwitz, & Cope (1986) and the need to balance anxiety research and the experiences of anxious language learners. In C. Gknou, M. Daubney & J-M Dewaele (Eds.), *New insights into language anxiety: theory, research and educational implications* (pp. 31–47). Multilingual Matters.

Housen, A., Kuiken, F., & Vedder, I. (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (Vol. 32). John Benjamins Publishing.

Huang, B. H., Chang, Y.-H. S., Zhi, M., & Niu, L. (2020). The effect of input on bilingual adolescents' long-term language outcomes in a foreign language instruction context. *International Journal of Bilingualism*, *24*(1), 8–25.

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.

Javad Ahmadian, M., Tavakoli, M., & Vahid Dastjerdi, H. (2015). The combined effects of online planning and task structure on complexity, accuracy and fluency of L2 speech. *Language Learning Journal*, *43*(1), 41–56. https://doi.org/10.1080/09571736.2012.681795

Kawauchi, C. (2005). The effects of strategic planning on the oral narratives of learners with low and high intermediate L2 proficiency. In R. Ellis (Ed.), *Planning and Task Performance in a Second Language* (pp. 143–164). John Benjamins Publishing Company.

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

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75.

Li, L., Chen, J., & Sun, L. (2015). The effects of different lengths of pretask planning time on L2 learners' oral test performance. *TESOL Quarterly*, *49*(1), 38–66.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development. Quantification and assessment. Palgrave Macmillan*.

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition*, *20*(1), 83–108.

Mohammad Javad A., & Tavakoli, M. (2011). The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research*, *15*(1), 35–59. https://doi.org/10.1177/1362168810383329

Newell, A.,& Simon, H. A. (1972). *Human problem solving* (Vol. 104, Issue 9). Prentice-Hall Englewood Cliffs, NJ.

Nielson, K. B. (2014). Can planning time compensate for individual differences in working memory capacity? *Language Teaching Research*, *18*(3), 272–293.

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

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

O'Sullivan, B. (2012). A brief history of language testing. In C. Coombe, P. Davidson, B. O'Sul-

livan & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment*, (pp. 9–19). Cambridge University Press.

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition*, *21*(1), 109–148. https://doi.org/10.1017/S0272263199001047

Panahzadeh, V., & Asadi, B. (2019). On the impacts of pressured vs. unpressured on-line task planning on EFL students' oral production in classroom and testing contexts. *Eurasian Journal of Applied Linguistics*, *5*(3), 341–352.

Pérez Castillejo, S. (2019). The role of foreign language anxiety on L2 utterance fluency during a final exam. *Language Testing*, *36*(3), 327–345.

Saeedi, M. (2020). Task condition and L2 oral performance: investigating the combined effects of online planning and immediacy. *International Journal of Foreign Language Teaching and Research*, *8*(32), 35–48.

Salehi, M., & Marefat, F. (2014). The Effects of Foreign Language Anxiety and Test Anxiety on Foreign Language Test Performance. *Theory & Practice in Language Studies*, *4*(5), 931–940.

Santana-Perera, B., & Arnaiz-Castro, P. (2018). The effects of three planning conditions on the complexity, accuracy, lexis, and fluency in Spanish adult EFL learners' oral production. *The International Journal of Pedagogy and Curriculum*, *25*, 1–18.

Santana-Perera, B. (2020). *El impacto de la planificación de la tarea en la producción oral de aprendices españoles universitarios de inglés como lengua extranjera*. Doctoral dissertation, Universidad de Las Palmas de Gran Canaria.

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

Skehan, P. (2014). Limited attentional capacity, second language performance, and task-based pedagogy. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 211– 26). John Benjamins Publishing Company.

Skehan, P., & Foster, P. (2007). Complexity, accuracy, fluency and lexis in task-based performance: A meta-analysis of the Ealing research. In P. Van Daele, S. Housen, F. Kuiken, M. Pierrard, & I Vedder (Eds.), *Complexity, accuracy, and fluency in second language use, learning, and teaching* (pp. 207–226). University of Brussels Press.

Soars, J., & Soars, L. (1993). *Headway: advanced student's book*. Oxford University Press.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Newbury House.

Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*, *58*(2), 439–473.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–273). John Benjamins Publishing Company.

VanPatten, B. (1990). Attending to form and content in the input: An experiment in consciousness. *Studies in Second Language Acquisition*, *12*(3), 287–301.

VanPatten, B. (2002). Processing instruction: an update. *Language Learning*, *52*(4), 755–803.

Wang, Z., & Skehan, P. (2014). Structure, lexis, and time perspective: Influences on task performance. In P. Skehan (Ed.), Investigating a *processing perspectives on task performance* (pp.155–185). John Benjamins Publishing Company.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, *14*(1), 85–106.

Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly*, *7*(1), 1–24.

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.

Zhang, X. (2019). Foreign language anxiety and foreign language performance: a meta-analysis. *The Modern Language Journal*, *103*(4), 763–781.