



Castledown

 OPEN ACCESS

Language Education & Assessment

ISSN 2209-0959

<https://www.castledown.com/journals/lea/>

Language Education & Assessment, 6(1), 1–18 (2023)

<https://doi.org/10.29140/lea.v6n1.1180>

The Use and Impact of Pre-Task Planning Time in the Monologic Task of LanguageCert Speaking Tests



LEDA LAMPROPOULOU

LanguageCert, Greece

leda.lampropoulou@peoplecert.org

Abstract

Extensive oral tasks or monologues of different types (e.g., presentations, storytelling) are often used as second language acquisition tasks in the fields of language learning and language testing. Pre-task planning time is a common provision to test-takers who may use different strategies to prepare their response. High-stakes tests, such as the LanguageCert IESOL suite of tests, include planning time prior to monologic tasks and offer test-takers the opportunity for note-making. While the language assessment literature supports planning time for reasons of face validity and fairness, research studies do not consistently support correlations between planning and performance. The current study examined the differences between the scores of test-takers who used note-making as a strategy and those who did not. The research questions investigated: (i) whether test-takers who make notes during planning time in the monologue task of an L2-English B2 speaking task are awarded higher scores on their spoken performances than test-takers who do not and (ii) test-takers' perceptions of their use of planning time. The findings suggest that making notes did not improve test-takers' performance against any of the rating criteria used in the assessment. It also revealed that most test-takers use their planning time to generate their main propositions.

Keywords: speaking exams, oral tests, planning time, pre-task planning, LanguageCert exams

Copyright: © 2023 Lampropoulou. This is an open access article distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within this paper.

Background and Study Scope

This study has two related focuses. Firstly, it examines whether using note-making during pre-task planning time creates gains for test-takers which are transferable into scoring. Secondly, it aims to provide insights into test-taker perceptions of their use and focus during planning time. In this regard, the background and research literature, and in particular, oral assessment task types and task planning, are first explored; following this, the study is outlined and pedagogical implications are discussed.

Second language (L2) assessments use various task types, frequently in combination, to operationalise the construct of speaking to measure, as accurately as possible, test-takers' L2 spoken ability. Brown (2004) placed these on a continuum of a five-level taxonomy. According to that, the simplest tasks in terms of cognition and ease of completion involve mere repetition of oral input and are classified as imitative. In between task types range from intensive and responsive to interactive, with the length of the expected response, turns, and exchanges adding to the tasks' complexity. Extensive language production tasks such as story narration and both formal and informal presentations were classified as the most complex and demanding. Brown's (2004) taxonomy included planning in the design of the extensive production tasks, a practice also reflected in L2 speaking assessments, such as the speaking tests in the LanguageCert International English for Speakers of Other Languages (IESOL) suite of exams.

Ellis (2009) and Foster and Skehan (1996) explored the influence of task planning on L2 oral production, focusing on syntactical *complexity*, *grammatical accuracy*, and *fluency* (CAF). Linguists have also researched the connection between planning and the length of planning time available (e.g., Li *et al.* 2014). Pang and Skehan (2014) examined task planning in a task-based language teaching (TBLT) setting, where pre-task planning is contrasted with task repetition and rehearsal.

In a meta-analysis of task planning and oral L2 production, Johnson and Abdi Tabari (2022) observed that two theoretical models have been used predominantly – Skehan's (1998) trade-off hypothesis and Robinson's (2011) cognition hypothesis. Briefly, the trade-off hypothesis suggests that attentional capacity to the three CAF elements is limited and that attending to increase performance in one area may take attention away from the other two and result in a weaker performance in those areas. The cognition hypothesis suggests that task complexity will raise both language complexity and accuracy, at the expense of fluency. Johnson and Abdi Tabari (2022) stated, nonetheless, that no research findings have consistently and unambiguously confirmed a positive relationship between planning and oral L2 production. Considering the absence of a firm conclusion of a seemingly fair assessment method and intuitively sensible good practice, the current study set out to examine the relationship between pre-task planning time and L2 oral production in the IESOL Speaking Test. Specifically, the study investigates the impact of making notes during pre-task planning time on the test-takers' performance in an extensive speaking task – a monologue – as part of a LanguageCert IESOL speaking test at B2 level. In particular, it looks at the effect that can be observed on test-taker scores – depending on whether test-takers use a note-making strategy to prepare their monologue on the assigned topic. Conducted in a formal assessment setting, scores awarded reflect test-takers' performance under the criteria of task fulfilment (TF), grammatical range and accuracy (GRA), lexical range and accuracy (VOC), and pronunciation intonation and fluency (PIF). The study also then briefly explores the key areas on which test-takers chose to direct their focus during that designated time.

Planning is “a problem-solving activity” (Ellis, 2005) through which learners select and determine a strategy to express a speech act, either in an automatic or a controlled manner depending on the learner's proficiency level, with a higher automatised ability available at higher L2 proficiency levels (De Bot, 1992). De Bot (1992) extended Levelt's (1989) system of first language (L1) production to also

apply it to L2 speech. In both systems, oral production comprises three gradual psycholinguistic processes: conceptualisation; formulation; articulation. Investigations into planning have mainly focused on its impact on the speech produced. Ellis (2005) classified planning into pre-task and within-task planning, with four issues generally identified:

- planning time length and availability
- learners' language proficiency level
- task type and task complexity
- the lack or presence of structured guidance

In L2 teaching, task-based language teaching (TBLT) has been used to operationalise strategic planning and to explore its impact on the complexity, accuracy, and fluency (CAF) of learners' performance. Crookes (1989) and Yuan and Ellis (2003) found significant improvements in all three CAF measures when pre-task planning time was allowed. Most studies have, however, only been able to report similar findings for aspects of fluency and complexity (e.g., Foster & Skehan 1996; Ortega, 1999). Geng and Ferguson (2013) found that planning time before a complex task positively affected the syntactic complexity of test-taker responses although at the expense of fluency.

An overview of research on planning time and its impact on the CAF of the spoken language produced has revealed a relative consensus on the value of providing L2 learners with an opportunity to plan their response in a teaching and learning context even if the benefits are more apparent for the complexity and fluency indicators and not always observable in the errors that learners make.

Table 1 presents an overview of research on planning time and its impact on the complexity, accuracy, and fluency of the speech produced. Out of the 19 studies reviewed, all but five observed a positive effect on complexity. Similarly, in fourteen of the studies a positive effect was observed on fluency, while only one saw a negative effect. The most inconsistent findings were observed for accuracy. Six studies found a positive effect, while another six concluded in mixed results.

A literature review of the research conducted to investigate the presence of a similar positive impact on the scores achieved by planners in an assessment setting, nonetheless, suggests a somewhat dissimilar picture. Overall, the findings of these studies do not consistently align with the conclusions reached by researchers investigating planning time within a learning context. Even within the assessment context, findings vary. An overview of the studies on the use of pre-task planning time in oral exams is presented in Table 2. Out of the thirteen studies reviewed, a positive effect on scoring was observed on only five of them, but that was significant only in three of them.

The stark contrast between the two contexts, that is, learning and assessment, is not as surprising as it may initially seem, as there are certain foundational differences between the two settings.

The amount of time provided to examination test-takers to plan their responses was, with very few exceptions, considerably shorter than what was offered to classroom learners. In TBLT, ten minutes of planning time has been used in most studies. In contrast, assessments that developed research-informed specifications and included planning time needed to also adhere to the principle of practicality (Bachman & Palmer, 1996) and limited the time offered to test-takers to one minute only in most of the test-based studies. Consequently, the time allowance for planning in a speaking test may be insufficient for an improvement in scoring to be observed.

Several test-based studies have employed CAF to evaluate test-taker performances (Li *et al.*, 2014; Nitta & Nakatsuhara, 2014). The most common practice remains, nonetheless, that of trained markers

Table 1 Overview of Research on Pre-task Planning Time in a TBLT Setting and its Effect on Performance

Researcher(s)	Planning time	Positive effect on		
		Complexity	Accuracy	Fluency
Crookes (1989)	10 min.	✓	✓	✓
Foster & Skehan (1996)	10 min.	✓	mixed	✓
Skehan & Foster (1997)	10 min.	mixed, with a trade-off effect		
Menhart (1998)	1 min., 5 min., 10 min.	✓	✓	✓
Ortega (1999)	10 min.	✓	mixed	✓
Foster & Skehan (1999)	10 min.	✓	✓	✓
Yuan & Ellis (2003)	10 min. and within task	✓	mixed	×
Kawauchi (2005)	10 min.	✓		✓
Sangarun (2005)	15 min.	✓		✓
Skehan & Foster (2005)	10 min.	✓		✓
Gilabert (2007)	10 min.	✓	×	✓
Mochizuki & Ortega (2008)	5 min.	✓	✓	✓
Guara-Tavares (2009)	10 min.	×	✓	✓
Ahangari & Abdi (2011)	10 min.	✓	×	
Sasayama & Izumi (2012)	5 min.	✓		×*
Genc (2012)	10 min.		×	
Geng & Ferguson (2013)	10 min.	✓	✓	✓
Nielson (2013)	10 min.	✓	mixed	✓
Khorami & Khorasani (2018)	10 min.		mixed	

Note: A significant positive effect in the specific area is symbolised with a tick (✓) in the respective column, whereas the absence of a statistically significant effect is symbolised with a cross (×). Inconsistent findings are described as mixed. Blank cells are in place when researchers did not research that area or did not report a finding for that skill. * This research found a negative effect on fluency.

Table 2 Overview of Research on Pre-task Planning Time in a Testing Setting and its Effect on Scores

Researcher(s)	Planning time	Effect on scoring
Wigglesworth (1997)	1 min.	no
Iwashita <i>et al.</i> (2001)	3 min.	no
Elder <i>et al.</i> (2002)	3 min.	no
Tavakoli & Skehan (2005)	5 min.	no
Xi (2010)	1 min.	yes, minimal
Elder & Iwashita (2005)	3 min.	no
Weir <i>et al.</i> (2006)	1 min.	yes
Elder & Wigglesworth (2006)	1 min., 2 min.	no
Nitta & Nakatsuhara (2014)	3 min.	yes
Li <i>et al.</i> (2014)	30 sec., 1 min.	no
Li <i>et al.</i> (2014)	2 min., 3 min., 5 min.	yes
O'Grady (2019)	30 sec., 1 min., 5 min., 10 min.	yes, minimal
Innué & Lam (2021)	1 min. and 30 sec.	no

assessing test-takers' responses using the respective test's rating scales or an adaptation of these. The resulting scores are then used for the research project's calculations. The absence of similar positive effects on the test-takers' spoken performance may therefore be attributed to the nature of the scale not being sufficiently sensitive for the improvements to be measured or that the CAF gains were too minor to be apparent (Wigglesworth, 1997; O'Grady, 2019).

Despite the relatively conflicted findings regarding the significance of the impact, a conclusion unanimously reached by test-based researchers was that pre-task planning time is justified and should be granted to test-takers. Swain's (1985) principle that tests ought to be biased towards eliciting a test-taker's best possible performance is widely regarded as good practice (Elder & Wigglesworth, 2006; O'Grady, 2019). Additional reasons in favour of providing test-takers with planning time in speaking tests include arguments on construct validity¹, authenticity, and fairness (O'Grady, 2019; Wigglesworth, 1997).

What appears to have been less researched in this area is "the test-takers' perspectives and their insights" into what occurs during planning time. Since inconclusive findings reported may suggest that test-takers use ineffective strategies to plan their responses, further investigation is needed into the impact of planning time as well as into test-takers' perceptions, to better comprehend the value of making notes as a planning strategy.

Research Questions

The research questions addressed in the study were:

- RQ 1:** To what extent are test-takers who make notes during planning time in the monologue task of an L2-English B2 Speaking Test awarded higher scores on the different rating scales on their spoken performances than test-takers who do not make notes?
- RQ 2:** What are test-takers' perceptions of the strategic planning time offered prior to the English B2 Speaking Test monologue task?

Methodology, Participants, Data

Data collection was completed in two phases. Test-takers' scores in the live exams provided the data for the quantitative analysis, whereas their responses to the questionnaire formed the data to explore their perspectives.

Participants

Participants who agreed to participate in the study were test-takers registered to take the LanguageCert IESOL Speaking test at B2 level of the Common European Framework of Reference for Languages (CEFR). Of the 50 participants who consented to take part in the study, 31 were students at an English language school (a LanguageCert test centre) while the remaining 19 had registered for the exams that LanguageCert conducts in Greece in its premises in Athens. There were 31 female and 19 male participants. The majority of the participants were from Greece ($n=45$), while four were Albanian and one was Indian. On average, the participants had been studying English as a foreign language for six years.

¹ Construct validity refers to the capacity of the generated test scores to be generalised and interpreted meaningfully and legitimately into the intended real-world use or target language use (TLU) domain (Bachman & Palmer, 1996).

Live Speaking Tests

Each test-taker in the study sat a face-to-face IESOL B2 Speaking exam. The exams were conducted as per the normal process i.e., a live interview with one interlocutor and one test-taker per session. A recording device was used to record audio only, based on which the test-taker's performance is assessed, at a later stage. A LanguageCert IESOL B2 Speaking test exam paper was used. The study focused on the final task: the monologue. The framework and a sample task are shown in Figure 1 below, with (I) referring to the interlocutor and (C) referring to the test-taker.

LanguageCert Communicator B2

PART 4 (4 minutes including follow-up questions)

I: In Part Four you are going to talk about something for two minutes. Your topic is *(choose topic for candidate)*.

Topics

A The countryside nearest to where you live

B Some recent news which has interested you

C What people should do to improve their language skills

I: *(Hand over piece of paper and pen/pencil.)* You now have thirty seconds to write some notes to help you. So your topic is *(repeat topic)*. *(Withdraw eye contact for thirty seconds. Leave recorder running.)*

I: *(Candidate's name)*, please start.

C: *(Talks.)*

I: *(When candidate has talked for a maximum of two minutes, say, 'Thank you', and then ask some follow-up questions.)*

Figure 1 Speaking Test Part 4 Rubric – Interlocutor Framework.

The format for the B2 Speaking test involves the interlocutor choosing a topic from a selection of three equivalent topics. For equivalence and reliability purposes, interlocutors were instructed to use the same topic for all test-takers and to ignore the other two options. The selected topic was: “A time when your family helped you.” Interlocutors announced the topic to the test-takers orally only and informed them that they were given 30 seconds as planning time to “write some notes to help [them].” Test-takers were given a pencil and a piece of paper, and the interlocutors then repeated the task topic. Planning time began at that point and lasted 30 seconds. If the test-taker insisted on starting their monologue early, however, this was permitted.

Test-takers kept their notes and could consult them during their talk. The interlocutor retrieved the notes at the end of the test, with test-takers aware that these would not be assessed.

Of the 55 sheets that were returned, five were unnamed and excluded from the study. For the remaining, 44 had at least one word noted while the remaining six were blank.

Post-Test Questionnaire

A questionnaire was used for two reasons. Firstly, previous researchers had identified the need for further insight into the test-takers' perceptions of pre-task planning. Secondly, since the creation of very brief notes does not allow significant mapping via analytic discourse analysis, a post-test questionnaire was administered that enquired into the use of planning time for the monologue part of the Speaking Test.

The questionnaire items comprised short questions, no negative constructions, and were written in language below the CEFR B2 level at which test-takers were being assessed.

A first draft was piloted on two volunteers from within LanguageCert's research team, following which, a second draft was then piloted with ten mock test-takers. The final version of the questionnaire was then developed. It included only four items and did not enquire into test-taker demographics. The first item asked test-takers whether they had made notes. Based on their initial response, they were asked what they focussed on during planning time, regardless of note-making. The last item asked test-takers whether the planning time allowed had been sufficient.

Out of the fifty test-takers who had consented to participate in the study only one did not wish to complete the questionnaire. The remaining forty-nine agreed and answered all items, with questionnaire completion taking place as close as possible to the Speaking Test itself, while test-takers exited the examination room. The questionnaire items can be found in Appendix A.

Scoring the Speaking Performances

Five trained interlocutors followed a scripted framework, interacting with the test-taker throughout the test. To avoid contamination, the interlocutors were not briefed on the details of the study. Interlocutors in face-to-face exams do not assess the test-takers but record the audio of the session to be assessed by a different marker.

An experienced marker rated all the Speaking test performances using the standard IESOL mark scheme. For the study, the marker was asked to first listen to and rate the monologue task individually, and to rate the test-taker's performance on the rest of the test (i.e., Parts 1–3) at a later stage. Marking was done over five days to minimise any halo effect and to avoid marker fatigue. A recording or parts of each test were available to ensure confident rating. The marker was not given access to the test-takers' notes. The test-takers' performances were then all second marked and inter-rater reliability was good ($\alpha=0.81$). For a better understanding of the marks available to be awarded, the rating scale at Communicator – CEFR B2 level with the analytic mark scheme and descriptors per mark can be found in Appendix B.

Quantitative Data Analysis

For each test-taker, five different raw marks were generated as per table 3 below.

Test-takers were divided into two groups. Group A ($n=44$) comprised test-takers who used their planning time to make notes writing down anything apart from the topic. Group B ($n=6$) consisted of those who only noted down the title of the topic or produced no notes. Samples of notes produced by the test-takers can be found in Appendix C.

Since the marker used a rating scale to award specific marks, there was already some indication that the data, being ordinal, should be analysed as not normally distributed. To confirm this, a descriptive statistics analysis of the test-takers' scores per criterion was performed using the program IBM SPSS Statistics

Table 3 *Marks Available*

TF	GRA	VRA	PIF	Total
0 – 3	0 – 3	0 – 3	0 – 3	0 – 12

Table 4 *Shaphiro-Wilk Tests for Data Normality*

Rating scale	Statistic
Task Fulfilment	(W = .747, p < .001)
Grammatical Resources	(W = .657, p < .001)
Lexical Resources	(W = .704, p < .001)
Pronunciation, Intonation, Fluency	(W = .519, p < .001)

(Version 27) to run a test of normality. A Shaphiro-Wilk test is suggested for a sample size of up to 50. The null hypothesis is that data are normally distributed. The results of the tests are shown in Table 4. For all four criteria, significance $P < .001$, which means the data should be handled as non-parametric.

As can be seen, on all rating scales, the results of Shaphiro-Wilk tests showed that the data were not normally distributed, so non-parametric tests were adopted to analyse the data of this study.

The effect of note-making during planning time on the test-takers' performance was then investigated by comparing the scores of Group A (note makers) and Group B (non-note makers) in the monologue. The nonparametric Mann-Whitney U test was used to compare the performance of the two groups.

Results

Monologue Rating Scale Scores

The main research question examined whether test-takers who make notes during planning time in the monologue task are awarded higher scores on the four criteria than test-takers who do not make notes. Mann-Whitney U test results are reported in turn below for the rating scales. Table 5 reports the results for Task Fulfilment.

The Mann-Whitney U Test results did not reach significance ($p = .85$), indicating that note makers do not score higher than non-note makers for topic development.

Table 6 reports the results for Grammatical Range and Accuracy.

The Mann-Whitney U Test revealed no significant difference in the GRA score of the two groups. The mean rank analysis of the GRA scores shows a much larger difference than the one reported for the TF criterion. However, the difference in test-takers' scores failed to reach statistical significance, suggesting that note makers do not use grammatical structures which are awarded higher scores than non-note makers.

Table 7 reports the results for vocabulary range and accuracy between note-makers and non-note-makers.

The Mann-Whitney U Test results did not reach significance indicating no substantial difference in the VRA scores of test-takers who made notes versus those who did not.

Table 8 presents the results of the Mann-Whitney U test for pronunciation intonation and fluency.

The Mann-Whitney U Test revealed no significant difference in the PIF score of the two groups – mirroring the null findings in the other three criteria and suggesting that there is no significant difference

Table 5 Mann-Whitney U Test on TF Scores

Mann-Whitney U	U = 125, p = .803
Group A	Md = 2, n = 44, mean rank = 25.34
Group B	Md = 2, n = 6, mean rank = 26.67

Table 6 Mann-Whitney U Test on GRA Scores

Mann-Whitney U	U = 81, p = .067
Group A	Md = 1, n = 44, mean rank = 26.66
Group B	Md = 1, n = 6, mean rank = 17.00

Table 7 Mann-Whitney U Test on VRA Scores

Mann-Whitney U	U = 114, p = .538
Group A	Md = 1, n = 44, mean rank = 25.91
Group B	Md = 1, n = 6, mean rank = 22.50

Table 8 Mann-Whitney U Test on PIF Scores

Mann-Whitney U	U = 111, p = .310
Group A	Md = 2, n = 44, mean rank = 25.03
Group B	Md = 2, n = 6, mean rank = 28.92

in the performances between note-making test-takers and those who did not make notes in pronunciation intonation and fluency aspects.

In summary, none of the tests conducted to examine the impact of making notes during planning time on the scores awarded for the four criteria of the rating scale produced statistically significant results, indicating that note-making test-takers were not awarded significantly different scores from non-note-makers.

Post-Test Questionnaire

The second research question explored test-takers' perceptions of the strategic planning time offered prior to the Speaking Test monologue task. Some items on the questionnaire (see Appendix A) allowed multiple responses: for example, what test-takers spent their planning time on, regardless of whether they had made notes.

Table 9 provides detail on questionnaire item 1, whether test-takers made notes in the planning time prior to the monologue.

Table 9 *Questionnaire Item 1 – Making Notes During Planning Time*

Did you make notes during planning time?	Responses
Yes, I made a lot of notes.	7 (14%)
Yes, but just some words.	36 (74%)
No, I didn't make any notes.	6 (12%)

Table 10 *Questionnaire Item 2 and Item 3 Responses (Combined Groups) – Planning Time Usage*

During planning time, I focussed on...	Responses
Generating ideas	38 (78%)
Structuring my monologue	10 (20%)
Planning my grammatical structures	4 (8%)
Selecting useful vocabulary items	8 (16%)
Calming down	5 (10%)
Nothing in particular	2 (4%)

Most of the test-takers (74%) stated they had made notes, but just a few words. The ratio of note sheets only containing just a few words is also an accurate representation of the collected sheets. There were a few test-takers (12%) who stated they had made no notes at all, and this number coincides with the number of sheets that were returned blank.

For the next set of questions, Question 2 (Q2) and Question 3 (Q3), test-takers were asked to specify how they had chosen to spend the thirty seconds of time they had at their disposal, or what their notes' purpose was, if they had made any.

Table 10 presents the available options and the responses for each one.

As can be seen from table 10, the most prevalent response was noting down or thinking of ideas to talk about, with 78% of the respondents reporting this as their main focus. 20% reported planning the monologue's structure. 16% reported focusing on vocabulary. Other options accounted for 10% of respondents or less.

To distinguish between planning strategies test-takers used and examine the relationship within note makers and non-note makers a crosstabulation of the focus areas within the two groups was undertaken. Table 11 shows the results of the crosstabulation of the questionnaire responses.

As can be seen from Table 11, 79% of note-makers focussed on thinking of and writing down the ideas to talk about.

The final question on the questionnaire enquired into the adequacy of the provided planning time. Table 12 depicts test-takers' responses to that regardless of whether they had made notes during that time or not.

The majority of test-takers responded that the time they were provided with sufficed to plan their response. Respondents from both groups agreed by approximately the same percentages – 63% and

Table 11 *Planning Time Usage by Test-Taker Group*

Planning Time Usage	Yes Notes	No Notes	Total
Ideas	34 (79.1%)	4 (66.7%)	38
Structure	10 (23.3%)	0 (0.0%)	10
Grammar	4 (9.3%)	0 (0.0%)	4
Vocabulary	8 (18.6%)	0 (0.0%)	8
Calm down	2 (4.7%)	3 (50.0%)	5
Nothing	0 (0.0%)	2 (33.3%)	2
Total	43	6	49

Table 12 *Questionnaire Q4 Responses: 30'' to Plan Your Monologue. Was This Time Enough?*

	Planning time was enough	Responses
Note makers	Yes	27 (63 %)
	No	16 (37 %)
Non note makers	Yes	4 (67 %)
	No	2 (33 %)

Table 13 *Questionnaire Frequency Statistics (n = 49)*

	Made Notes	Ideas	Structure	Grammar	Vocabulary	Calm down	Nothing	Enough time
Mean	.88	.78	.20	.08	.16	.10	.04	.61
SD	.33	.42	.41	.28	.37	.306	.2	.49

67% respectively – that they did not need more time to plan better. A follow-up question was asked about what they would use the extra time for. Most test-takers responded that they would have used it to think of more ideas to talk about while a few others mentioned they would have used it to relax. The frequencies analysis for all items is presented in Table 13.

Discussion

This study investigated the potential effects of planning one's monologue through making notes on test-taker performance in a B2 speaking test setting using LanguageCert IESOL speaking test-takers. Assessed performances of note makers were compared on four rating scales with the performances of test-takers who did not use a note-making strategy.

The research found that note-makers were not awarded statistically significant higher marks on any of the criteria. This suggests that notes did not help test-takers fulfil the given task more fully or more coherently, nor did they demonstrate a consistently higher level of GRA, a superior VOC, or a more natural and effective PIF. In other words, the analyses outcomes seem to suggest that there is no difference on the test-taker's performance, regardless of whether they prepare a response through making notes or not.

These results may appear counter-intuitive and in conflict with the TBLT literature which suggests that pre-task planning time offered in a classroom setting can substantially improve test-takers' task performance when speech production is measured against CAF indices (Foster & Skehan, 1996; Geng & Ferguson, 2013). Fluency and complexity, the two areas where gains from planning are most frequently observed in TBLT studies, did not appear to behave any differently from the rest of the criteria examined in the current study. A relatively uncomplicated explanation is that TBLT studies in most cases offered learners much more planning time (ten minutes in TBLT compared to one to two minutes for most test tasks). However, studies that allowed test-takers five to ten minutes to prepare results were still not able to confirm a meaningful effect on test-takers' performances (Tavakoli & Skehan, 2005).

Nevertheless, and perhaps more importantly, the current study's null results are consistent with the body of research conducted under exam conditions. These suggest that strategic planning did not have a meaningful or substantial effect on test-takers' spoken production (Inoue & Lam, 2021; Wigglesworth & Elder, 2010). Thus, the conclusion is that using note-making as a strategy for optimal performance in the monologue section of the LanguageCert IESOL speaking test does not produce an observable improvement on spoken performance and may be of limited effectiveness, as currently used by test-takers.

By way of reinforcing the fairness argument and adding to it an element of face validity, responses to the appropriacy of the offered length of planning time revealed that test-takers were relatively split between those who were happy with the time provided (approximately 60%) and those who would have wished for more (approximately 40%). Consequently, although extending planning time may be against test practicality and unsupported by the study's findings, reducing or removing it altogether might jeopardise score acceptance by test-takers and stakeholders (Inoue & Lam, 2021; Wigglesworth & Elder, 2010).

Limitations and Implications

In the process of acknowledging the limitations of this research some additional points should be considered. Firstly, the rating scale's (in)capacity to measure the kind of gains produced by planning and note-making could perhaps account for the absence of a demonstrable effect on scores. The issue of the appropriacy and fitness of the rating scale was encountered with different types of scales and measures. O'Grady (2019) used an empirically derived binary-choice, boundary-definition (EBB) scale, and an analytic rating scale but did not find substantial differences between the generated test scores. It may therefore be useful to explore whether a different, perhaps longer than four bands, scale would guide markers to different assessment decisions that may allow gains from strategical planning to manifest in test-takers' scores.

Secondly, the quality of test-takers' notes during planning should be considered because only a very limited amount of the notes would qualify as good plans for an effective oral response. Very few test-takers used arrows or a bullet list and none sketched notes in the form of a mind map. These poorly written plans demonstrate the test-takers' overall lack of good note-making skills which otherwise could be conducive to an improved response to the oral task. The possibility for test takers' note-making skills to be ineffective and unable to assist them in producing a better oral response than they would produce without any notes entails certain pedagogical implications. Teachers preparing students for LanguageCert oral exams – but also for any speaking examination that contains similar tasks where the opportunity for making notes is offered to the test takers – might want to teach planning skills in a structured and explicit manner, to help their learners develop and sharpen them. Despite the fact that note making is a life skill which will also be useful to the learner beyond the test, it is generally

undervalued. The interlocutors conducting the speaking exams are a case in point. In discussing their views on the task after their had conducted the exams, they reflected that, as language teachers, they had never explicitly taught their students how to make notes in preparation for that speaking task. This cannot be generalized, and a study on how teachers teach planning strategies could shed light on the matter. In the meantime, teachers should perhaps consider honing such skills through practicing different planning strategies, in an attempt to explore what works for each learner, and that will be a welcome positive washback of the speaking assessments.

A significant limitation in this study was that the participants were not equally split into note makers and non-note makers as the majority of the test-takers opted for some sort of notes. In this study the aim was to investigate the note-making strategy keeping the live exam conditions untouched. A future study could also examine how test-takers who normally make notes would perform if deprived of the note-making option.

Conclusion

The present study sought to examine the use of note-making as a pre-task planning strategy and its effect on L2 oral performance at a CEFR B2 level (Council of Europe 2001) speaking test task aiming at eliciting a monologue. The view that there is a positive correlation between planning and performance has been widely endorsed in language teaching and supported in the relevant research (Nielson, 2013; Yuan and Ellis, 2003). Nevertheless, a similar effect is not consistently present in a testing setting (Inoue & Lam, 2021; Nitta & Nakatsuhara, 2014; Wigglesworth & Elder, 2010).

The current study, conducted in a face-to-face exam context, reflects the findings of other relevant studies that using note-making as a pre-task planning strategy does not have a significant effect on test-takers' performance in terms of their scores. This finding was consistent across all four criteria comprising the test's mark scheme. In this sense, this study can be used to complement the body of research on pre-task planning time usage by adding to the range of task types (a 2-minute monologue), the specific strategy (note-making), and the specific planning time provided (30 seconds). It also contributes to the less researched area of test-takers' perceptions of their own use of planning time and note-making.

These findings, however, should not be taken to imply that pre-task planning time is redundant in an oral test or to suggest that assessment developers should eliminate them from test task specifications. The current study argues that planning time of as little as thirty seconds should be included in the design of all extensive monologic tasks; the provision of a longer planning period should be considered for more demanding tasks which may require more than idea generation.

Apart from researchers and test developers interested in designing research-informed test specifications, SLA practitioners such as teachers of English as a foreign language (EFL) and materials developers may also benefit from being aware of the results and the pedagogical implications of this type of studies. The poor quality of the test takers' note making practices as observed in their returned indicates there may be a learning gap in planning strategies and note-making skills.

References

- Bachman, L.F., & Palmer, A.S. (1996). *Language testing in practice*. Oxford University Press.
Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. Longman.

- Bui, G., & Huang, Z. (2018). L2 fluency as influenced by content familiarity and planning: Performance, measurement, and pedagogy. *Language Teaching Research*, 22(1), 94–114. <https://doi.org/10.1177/1362168816656650>
- De Bot, K. (1992). A bilingual production model: Levelt's "speaking model" adapted. *Applied Linguistics*, 13, 1–24.
- Council of Europe (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.
- Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition*, 11(4), 367–383. <https://doi.org/10.1017/S0272263100008391>
- Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer? *Language Testing*, 19(4), 347–368. <https://doi.org/10.1191/0265532202lt235oa>
- Elder, C., & Iwashita, N. (2005). Planning for test performance: Does it make a difference? In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 219–239). John Benjamins.
- Elder, C., & Wigglesworth, G. (2006). An investigation of the effectiveness and validity of planning time in part 2 of the IELTS speaking test. *IELTS Research Reports (Vol. 6)*, pp. 1–28. IELTS Australia and British Council.
- Ellis, R. (2005). Planning and task-based performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 3–34). John Benjamins.
- Foster, P., & Skehan, P. (1996). The influence of planning time on performance in task-based learning. *Studies in Second Language Acquisition*, 18, 299–234. <https://doi.org/10.1017/S0272263100015047>
- Gilbert, R. (2007). The simultaneous manipulation of task complexity along planning time and (+/- here and now): Effects on oral production. In M. Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 44–68). Multilingual Matters.
- Genc, Z. (2012). Effects of strategic planning on the accuracy of oral and written tasks in the performance of Turkish EFL learners. In A. Shehadeh & C. Coombe (Eds.), *Task-based language teaching in foreign language contexts research and implementation* (pp. 67–89). John Benjamins.
- Geng, X., & Ferguson, G. (2013). Strategic planning in task-based language teaching: The effects of participatory structure and task type. *System*, 41, 982–993.
- Guara-Tavares, M. (2009). The relationship among pre-task planning, working memory capacity, and L2 speech performance: A pilot study. *Linguagem & Ensino*, 12(1), 165–194. <https://doi.org/10.1590/S0103-18132013000100002>
- Inoue, C., & Lam, D.M.K. (2021). The effects of extended planning time on test takers' performance, processes, and strategy use in the lecture listening-into-speaking tasks of the TOEFL iBT® Test. *ETS Research Report Series, 2021*, 1–32. <https://doi.org/10.1002/ets2.12322>
- 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, 401–36. <https://doi.org/10.1111/0023-8333.00160>
- Johnson, M. D., & Abdi Tabari, M. (2022). Task planning and oral L2 production: A research synthesis and meta-analysis. *Applied Linguistics*, 43(6), 1143–1164. <https://doi.org/10.1093/applin/amac026>
- 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. 143–165). John Benjamins.
- Khorami, A., & Khorasani, R. (2018). The effects of planning time and proficiency level on accuracy of oral task performance. *Global Journal of Foreign Language Teaching*, 7(4), 155–168. <https://doi.org/10.18844/gjft.v7i4.3004>

- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge University Press.
- Li, L., Chen, J., & Sun, L. (2014). The effects of different lengths of pre-task planning time on L2 learners' oral test performance. *TESOL Quarterly*, 49(1), 38–66. <https://doi.org/10.1002/tesq.159>
- Mochizuki, N., & Ortega, L. (2008). Balancing communication and grammar in beginning-level foreign language classrooms: A study of guided planning and relativization. *Language Teaching Research*, 12(1), 11–37. <https://doi.org/10.1177/1362168807084492>
- Nielson, K. (2013). Can planning time compensate for individual differences in working memory capacity? *Language Teaching Research*, 18(3), 272–293. <https://doi.org/10.1177/1362168813510377>
- Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on oral task performance. *Language Testing*, 31(2), 147–175. <https://doi.org/10.1177/0265532213514401>
- 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. <https://doi.org/10.1177/0265532219826604>
- 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>
- Pang, F., & Skehan, P. (2014). Self-reported planning behaviour and second language reporting in narrative retelling. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 95–128). John Benjamins.
- Robinson, P. (Ed.). (2011). *Second language task complexity: Researching the cognition hypothesis of language learning and performance*. John Benjamins. <https://doi.org/10.1075/tblt.2>
- Sangarun, J. (2005). The effects of focusing on meaning and form in strategic planning. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 111–143). John Benjamins.
- Sasayama, S., & Izumi, S. (2012). Effects of task complexity and pre-task planning on Japanese EFL learners' oral production. In A. Shehadeh & C. Coombe (Eds.), *Task-based language teaching in foreign language contexts research and implementation* (pp. 23–43). John Benjamins.
- Skehan, P., & Foster, P. (1997). Task type and processing conditions as influences on foreign language performance. *Language Teaching Research*, 1(3), 185–211. <https://doi.org/10.1111/1467-9922.00071>
- Skehan, P., & Foster, P. (2005). Strategic and on-line planning: The influence of surprise information and task time on second language performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 193–219). John Benjamins.
- Swain, M. (1985). Large scale communicative testing: A case study. In Y. Lee, C. Fok, R. Lord & G. Low (Eds.), *New directions in language testing* (pp. 35–46). Pergamon Press.
- 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. 239–277). Amsterdam: John Benjamins.
- Weir, C., O'Sullivan, B., & Horai, T. (2006). Exploring difficulty in speaking tasks: An intra-task perspective. *IELTS Research Reports* (Vol. 6, pp. 1–42). IELTS Australia and British Council.
- Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, 14(1), 85–106. <https://doi.org/10.1177/026553229701400105>
- Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly*, 7(1), 1–24. <https://doi.org/10.1080/15434300903031779>
- Xi, X. (2010). Aspects of performance on line graph description tasks: Influenced by graph familiarity and different task features. *Language Testing*, 27(1), 73–100. <https://doi.org/10.1177/0265532209346454>
- Yuan, F., & Ellis, R. (2003). The effects on pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, 21(1), 1–27.

Appendix A

Questionnaire items

- 1) **During your speaking exam, you were given some preparation time to plan your monologue. Did you make any notes during your preparation time?**
 - a) Yes, I made a lot of notes.
 - b) Yes, but just some words.
 - c) No, I didn't make any notes.

- 2) **If you made notes, what was the purpose behind them? (You may choose more than one response, if necessary.)**
 - a) To note down the ideas to speak about.
 - b) To structure my monologue.
 - c) To plan what grammar I will use.
 - d) To note down useful vocabulary.
 - e) Other:

- 3) **If you didn't make notes, what did you use your preparation time for? (You may choose more than one response, if necessary.)**
 - a) To think of the ideas to speak about.
 - b) To think of how to structure my talk.
 - c) To think about the grammar I will use.
 - d) To think about useful vocabulary.
 - e) To calm myself down before I start talking.
 - f) I wasn't thinking of anything.
 - g) Other:

- 4) **You were given 30" to prepare. Was the time enough?**
 - a) Yes.
 - b) No.

- 5) **Level of the exam you took (circle):** A1 / A2 / B1 / B2 / C1 / C2

Appendix B

LanguageCert IESOL Speaking B2 – Markscheme and analytic descriptors

	Task Fulfilment and Coherence	Accuracy and Range of Grammar	Accuracy and Range of Vocabulary	Pronunciation, Intonation and Fluency
3	<ul style="list-style-type: none"> – Tasks are fulfilled with ease and confidence – Turn taking is spontaneous and natural – Contributions are fully relevant and detailed – Significant points are appropriately highlighted with supporting detail – Discourse is clear and coherent and produced in an appropriate style with a wide range of B2 level cohesive devices 	<ul style="list-style-type: none"> – A wide range of B2 level grammar is used – There is a consistently high level of accuracy and control – Occasional errors may occur, but are often corrected 	<ul style="list-style-type: none"> – A wide range of B2 level vocabulary is used to deal with the tasks – Choice of vocabulary is generally appropriate and effective 	<ul style="list-style-type: none"> – Pronunciation is clear and natural – Intonation is used to convey meaning effectively – The flow of language is maintained effectively – No evident hesitations
2	<ul style="list-style-type: none"> – Tasks are fulfilled with relative ease – Turn taking is naturally handled – Contributions are mostly relevant – Intended message is clearly communicated. Misunderstandings are rare – Discourse is mostly clear and coherent with use of B2 level cohesive devices 	<ul style="list-style-type: none"> – A good range of B2 level grammar is used – There is a good level of accuracy and control – Some errors may occur, but the message is always communicated 	<ul style="list-style-type: none"> – A sufficient range of vocabulary is used to deal with the B2 tasks – Choice of vocabulary is generally appropriate and effective – Some vocabulary errors occur, but do not impede communication 	<ul style="list-style-type: none"> – Pronunciation is reasonably clear and easily understood – Stress and intonation patterns are appropriately used to help convey meaning – The flow of language is generally maintained despite some hesitation – No undue strain on the listener
1	<ul style="list-style-type: none"> – Tasks remain largely unfulfilled – Interaction is only maintained with the support of the interlocutor – Little natural turn taking takes place – Contributions lack relevance – Intended message is only communicated with difficulty – Ideas are linked together simply and may be difficult to follow 	<ul style="list-style-type: none"> – Range of grammar is too limited to deal with the B2 level tasks – Frequent errors are noticeable, and may impede communication 	<ul style="list-style-type: none"> – Range of vocabulary is too limited to deal with the B2 level tasks – Vocabulary errors may make the message difficult to follow 	<ul style="list-style-type: none"> – Unclear pronunciation leads to undue strain on the listener – Inappropriate stress and intonation patterns impede communication – Frequent hesitations are evident, with repetition and attempts to repair language
0	<ul style="list-style-type: none"> – The tasks are unfulfilled and intended message is not successfully communicated – Ideas are difficult to follow and not linked together into connected speech – OR insufficient sample of language to assess 	<ul style="list-style-type: none"> – Inadequate range of grammar – Frequent errors impede communication – OR insufficient sample of language to assess 	<ul style="list-style-type: none"> – Lacks the vocabulary to deal with the B2 level tasks – The message is obscured by vocabulary errors – OR insufficient sample of language to assess 	<ul style="list-style-type: none"> – Unclear pronunciation and/or intonation prevents clear understanding – Frequent hesitation places strain on the listener – OR insufficient sample of language to assess

Appendix C

Samples of test-takers' note making sheets

a friend of mine / dissatisfied
 school bullying
 B2/ 26

family help
 B2/ 23
 to me
 ↳ Brother
 ↳ Sister
 ↳ Parents
 usually and d.c. a time

- ~~advice~~
 - ~~support~~
 - Help to decide the right
 and advice
 B2/ 4

• family helped
 • support
 • coach
 B2/ 8

A time that my family
 B2/ 7
 family is support + love
 they were here by my side
 about public

A time when my family helped me
 B2/ 38
 when I am sad

few months ago
 normal health
 always support me
 always by my side
 B2/ 10

when I have problems with school
 and exercises and about friends
 be it's difficult for me to
 finish
 B2/ 33

who,
 when
 B2/ 34

2 years ago
 I was very accident
 with my family
 B2/ 3

a time when my family help me.
 B2/ 16

a time at 6th grade.
 B2/ 14

a time when your family helped
 chores
 study
 driving
 learn
 swim
 B2/ 22