Quality Beyond Measure.

TOEFL® Research Report
TOEFL–RR-93
ETS Research Report No. RR–21-09

The Effects of Extended Planning Time on Candidates' Performance, Processes, and Strategy Use in the Lecture Listening-Into-Speaking Tasks of the *TOEFL iBT*® Test

Chihiro Inoue

Daniel M. K. Lam

December 2021

The *TOEFL*® test is the world's most widely respected English language assessment, used for admissions purposes in more than 130 countries including Australia, Canada, New Zealand, the United Kingdom, and the United States. Since its initial launch in 1964, the TOEFL test has undergone several major revisions motivated by advances in theories of language ability and changes in English teaching practices. The most recent revision, the *TOEFL iBT*® test, contains a number of innovative design features, including integrated tasks that engage multiple skills to simulate language use in academic settings and test materials that reflect the reading, listening, speaking, and writing demands of real-world academic environments. In addition to the TOEFL iBT, the TOEFL Family of Assessments has expanded to provide high-quality English proficiency assessments for a variety of academic uses and contexts. The TOEFL Young Students Series (YSS) features the *TOEFL® Primary*™ and *TOEFL Junior*® tests, designed to help teachers and learners of English in school settings. The *TOEFL ITP*® Assessment Series offers colleges, universities, and others an affordable test for placement and progress monitoring within English programs.

Since the 1970s, the TOEFL tests have had a rigorous, productive, and far-ranging research program. ETS has made the establishment of a strong research base a consistent feature of the development and evolution of the TOEFL tests, because only through a rigorous program of research can a testing company demonstrate its forward-looking vision and substantiate claims about what test takers know or can do based on their test scores. In addition to the 20-30 TOEFL-related research projects conducted by ETS Research & Development staff each year, the TOEFL Committee of Examiners (COE), composed of distinguished language-learning and testing experts from the academic community, funds an annual program of research supporting the TOEFL family of assessments, including projects carried out by external researchers from all over the world.

To date, hundreds of studies on the TOEFL tests have been published in refereed academic journals and books. In addition, more than 300 peer-reviewed reports about TOEFL research have been published by ETS. These publications have appeared in several different series historically: TOEFL Monographs, TOEFL Technical Reports, TOEFL iBT Research Reports, and TOEFL Junior Research Reports. It is the purpose of the current TOEFL Research Report Series to serve as the primary venue for all ETS publications on research conducted in relation to all members of the TOEFL Family of Assessments.

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org **Web site:** www.ets.org/toefl

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

TOEFL Research Report Series and ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

The Effects of Extended Planning Time on Candidates' Performance, Processes, and Strategy Use in the Lecture Listening-Into-Speaking Tasks of the *TOEFL iBT*® Test

Chihiro Inoue, & Daniel M. K. Lam

Centre for Research in English Language Learning and Assessment, University of Bedfordshire, Bedfordshire, UK

This study investigated the effects of two different planning time conditions (i.e., operational [20 s] and extended length [90 s]) for the lecture listening-into-speaking tasks of the *TOEFL iBT*® test for candidates at different proficiency levels. Seventy international students based in universities and language schools in the United Kingdom (35 at a lower level; 35 at a higher level) participated in the study. The effects of different lengths of planning time were examined in terms of (a) the scores given by ETS-certified raters; (b) the quality of the speaking performances characterized by accurately reproduced idea units and the measures of complexity, accuracy, and fluency; and (c) self-reported use of cognitive and metacognitive processes and strategies during listening, planning, and speaking. The results found neither a statistically significant main effect of the length of planning time nor an interaction between planning time and proficiency on the scores or on the quality of the speaking performance. There were several cognitive and metacognitive processes and strategies where significantly more engagement was reported under the extended planning time, which suggests enhanced cognitive validity of the task. However, the increased engagement in planning did not lead to any measurable improvement in the score. Therefore, in the interest of practicality, the results of this study provide justifications for the operational length of planning time for the lecture listening-into-speaking tasks in the speaking section of the TOEFL iBT test.

Keywords Planning time; integrated tasks; listening-into-speaking; CAF; cognitive processes; cognitive and metacognitive strategies

doi:10.1002/ets2.12322

Integrated listening-into-speaking tasks are increasingly employed in the speaking component of language proficiency tests (e.g., the *TOEFL iBT*® test and the PTE Academic test). These are considered by researchers to tap into a different construct from independent speaking tasks (Brown et al., 2005; Lee, 2006). Whereas independent speaking tasks require little integration of the information from the input into candidates' output, integrated speaking tasks have been argued to reflect real-life communicative acts better because they not only require listening and speaking skills but also engage other cognitive skills, such as those for selecting, organizing, and transforming source information for the production of output (Barkaoui et al., 2013; Brown et al., 2005; Frost et al., 2011). Such advantages of integrated tasks over independent tasks might be particularly relevant to proficiency tests for academic purposes.

A question often surrounding the design of language test tasks is whether they resemble the corresponding real-life task in the target domain. This idea is captured by Bachman's (1990) notion of situational authenticity. Taking this perspective, a test task such as the one focused on in this study—consisting of a 2-min lecture listening input, 20 s of planning time, and 1 min of speaking time to summarize the input—could be easily dismissed by commentators as lacking resemblance to the corresponding real-life academic task. However, Galaczi and Taylor (2018) argued that "a test by definition would have limited situational authenticity" (p. 222) and that a more realistic and generally accepted view in test development and validation research is to consider "the degree of correspondence between the cognitive processes triggered by a test task and a non-test task from the target language use domain" (p. 222), such as interactional authenticity in Bachman's terms. Notably, this approach has been taken in constructing the TOEFL iBT validity argument (Chapelle, 2008) whereby "observations of performance on the TOEFL reveal relevant knowledge, skills and abilities in situations representative of those in the target domain of language use in English-medium institutions of higher education" (p. 347). Building on Weir's (2005) socio-cognitive framework for test development and validation, Field (2011) argued for the criticality

Corresponding author: C. Inoue, E-mail: chihiro.inoue@beds.ac.uk

of investigating the cognitive processes engaged by candidates during test taking. He explains the value of examining candidates' cognitive processes as validity evidence in the following words:

> This is not a simple matter of ecological validity. The goal is to establish whether the tasks proposed by a test designer elicit mental processes resembling those which a language user would actually employ when undertaking similar tasks in the world beyond the test. The processes in question might relate to the way in which the user assembles or interprets input; or they might reflect the cognitive demands imposed upon the user by facets of the task. (Field, 2011, p. 67).

Echoing Weir (2005), Field cautioned against validation approaches that rely exclusively on backtracking from the product (e.g., speaking performance) or outcome (e.g., scores) to the processes that have generated them.

Consider the cognitive processes students engage in when summarizing lecture contents in speech. In real-life academic settings, the time lag between listening to the lecture and summarizing or discussing the lecture content may vary, but there are usually some opportunities for planning prior to delivery (Elder et al., 2002). Students may get to plan the content of their speech (e.g., reviewing their notes, consolidating and refreshing their mental representation of the lecture content) as well as the language (how they will say what they want to say). They might even engage in mental or subarticulatory rehearsal (e.g., just before articulating their contribution in a class discussion) or verbal rehearsal of their speech (e.g., prior to an oral presentation). A pertinent question on the interactional authenticity or cognitive validity of the lecture listening-into-speaking tasks used in the speaking section of TOEFL iBT is therefore whether the operational 20 s planning time allows for candidates' engagement in these processes of planning for content, planning for language, and mental rehearsal.

An earlier study on the TOEFL iBT lecture listening-into-speaking tasks by Swain et al. (2009) found that the candidates engaged in planning for content during the 20 s planning time but rarely planned for language or mentally rehearsing their speech. Thus, this study aimed to explore whether an extended planning time would allow candidates to engage in the cognitive processes of planning for language and mental rehearsal to greater extents and, if so, whether the extended planning time afforded any measurable benefits to candidates' speaking performances, specifically, the scores on the task, the linguistic quality of the speech, and the reproduction of the lecture content. If the extended planning time was shown to provide stronger support for the cognitive validity of the task than the 20 s planning time, then extending the operational planning time length might be justified.

Literature Review

This section starts by reviewing key relevant studies on how learners actually use the planning time before speaking and on the effects of planning time on spoken performance and test scores. Because these studies were conducted with independent speaking tasks, the review then considers vital aspects for investigating integrated speaking tasks: the role of listening and the incorporation of idea units (IUs) as a measure of content reproduction.

Processes and Strategy Use in Planning Time

Although there is a dearth of research into how second language (L2) learners use planning time preceding spoken production within testing contexts (Wigglesworth & Elder, 2010), some evidence is found from a few studies in the field of L2 acquisition that the use of planning time interacts with proficiency levels. Ortega (2005) examined the process of planning through a retrospective method and found that low-intermediate level participants often reported using retrieval strategies, whereas advanced learners used more rehearsal strategies while planning. These results confirmed the speculations by Wigglesworth (1997) and in Ortega's earlier studies (Ortega, 1995, 1999) that planning would be beneficial for higher level learners in improving the formal aspects of performance, whereas lower level learners may need to allocate more attention and time to prioritize planning the content and the relevant lexis. To explain these differences among higher and lower level learners with reference to the modular model of L2 speech production (Kormos, 2006, as noted in Levelt, 1989), lower level learners may take more time at the stages of conceptualization and the lexico-grammatical encoding stage in formulation, whereas more advanced learners may go through these stages more quickly, coming up with more sophisticated language because of their highly automatized processes. Thus, more advanced learners are able to utilize the planning time for rehearsal operations before starting to speak, which may contribute to higher fluency and accuracy of the elicited performance.

Table 1 Summary of Findings of the Effects of Relatively Short Periods of Planning on Scores and Linguistic Variables

Study by	Planning time length	Proficiency level	Key findings
Wigglesworth (1997)	None 1 min	High/low[a]	Planning helped only the more highly proficient learners on syntactic complexity and accuracy on more cognitively demanding tasks
Mehnert (1998)	None 1 min 5 min 10 min	Intermediate[a]	• Fluency and lexical density improved with each increase of planning time • Accuracy increased only with 1 min • Syntactic complexity was higher only with 10 min
Elder and Iwashita (2005)	1.25 min 4.25 min	Advanced/intermediate (*TOEFL PBT*® 427–670)	No statistically significant effects on linguistic variables or candidate perceptions of task
Wigglesworth and Elder (2010)	1 min 2 min	Advanced/intermediate (TOEFL PBT 427–670)	No statistically significant impact on scores (of IELTS Part 2 long-turn task) or on linguistic variables
Li et al. (2015)	None 30 s 1 min 2 min 3 min 5 min	Intermediate (CET-4)	• 30 s was insufficient for any improvement; even detrimental to fluency compared with no planning • 1 min was the threshold that led to statistically significant improvement on accuracy • Longer planning time produced gradually more accurate but not steadily more fluent or complex utterances; the degrees of improvement decreased • 5 min had diminishing effects • Candidate preferred 1–3 min planning time
O'Grady (2018)	30 s 1 min 5 min 10 min	Beginner (CEFR A1, A2)/ intermediate (B1)	• 10 min planning time led to a small increase in scores compared to 30 s in beginner group • No statistically significant effects on scores found in intermediate group • No statistically significant effects on linguistic variables except the number of idea units with 5 min compared to 1 min

[a] No further details were found.

Effects of Relatively Short Planning Time

A number of studies have examined the effects of planning time on performance using independent speaking tasks in the field of task-based language learning (e.g., Crookes, 1989; Foster & Skehan, 1996; Ortega, 1999; Tavakoli & Skehan, 2005; Yuan & Ellis, 2003). In these studies, the dimensions of spoken performance in question often included syntactic and lexical complexity, accuracy, and fluency (CAF). Some studies were classroom based and therefore employed a rather long planning time, such as 5 min and over, but such planning time lengths are unrealistic in testing situations. Therefore, in this section, only key studies that included relatively short periods of planning time are reviewed; the lengths of planning time and the key results of such studies are summarized in Table 1.

The aspects of performance and the extent to which the longer planning conditions affected them vary among the studies. The recent study by O'Grady (2018), which involved beginner and intermediate levels of learners, found small improvements in scores (at beginner level) and in the number of IUs with 5-min or 10-min planning. However, as noted earlier, such lengths of planning time are unrealistically long for a large-scale, standardized testing context. Among learners at the intermediate or advanced levels, there appears to be a trend for statistically significant improvements from no planning time to 1-min planning (Li et al., 2015; Mehnert, 1998; Wigglesworth, 1997). Contrastively, when the shorter planning time starts at or exceeds 1 min, the longer planning time does not lead to statistically significant improvements (Elder & Iwashita, 2005; Wigglesworth & Elder, 2010). This contrast is possibly because, according to Li et al. (2015), 1 min is a "threshold" for measurable improvements, and planning time of longer than 1 min results in further improvements but at decreased rates. Li et al. also reported that 5-min planning time had detrimental effects and speculated that it might have led to overplanning and overrehearsal by candidates, which could in turn lead to stilted spoken production and lower quality performance (Lam, 2015; Li et al., 2015).

Considerations for Integrated Speaking Tasks

Processes and Strategy Use in Listening

The crucial difference between the studies presented in the previous sections and this study lies in the type of tasks: The previous studies were conducted with independent speaking tasks, whereas this study focuses on integrated speaking tasks. Unlike independent tasks (e.g., picture-based narrative tasks and opinion-giving tasks), which provide little language and require learners to do most of the message generation, the listening texts in lecture listening-into-speaking tasks provide the basis for the content and language of the task performance (Cumming et al., 2004). Therefore, the listening stage forms part of the planning, which warrants empirical investigations into the processing of the input text and strategy use not only during planning and speaking but also during listening. However, this very intertwined relationship between comprehension of the input text and planning for output in integrated speaking tasks makes it difficult to fully separate and investigate the effects of planning time alone.

In an earlier study on the TOEFL iBT integrated speaking tasks, Swain et al. (2009) investigated the relationships between the scores, candidates' proficiency levels, and reported strategy use during listening, planning, and speaking time. They found no statistically significant differences in reported strategy use across proficiency levels or relationship between the number of reported strategic behaviors and test score. Their nonsignificant results may be attributable to using rather crude coding categories, the majority of which seem to have focused on strategies used during planning and speaking. Of the 49 strategies reported, only three appear to be relevant for listening: guessing, anticipating the structure of talk, and monitoring. Furthermore, the coding category for monitoring combined its use during listening, planning, and speaking stages, which may have masked potential differences at each stage of task completion.

In the recent work by Rukthong and Brunfaut (2020), based on the findings of Rukthong (2016), more fine-grained, theoretically based coding categories were used for the analysis of the candidates' self-reported processes and strategy use during the listening stage in integrated tasks. In contrast to Swain et al. (2009), Rukthong and Brunfaut identified differences in cognitive processing and strategy use among learners of different proficiency levels. Specifically, they found that the high scorers engaged in higher-level cognitive processes (i.e., semantic processing at the global level and pragmatic processing), which led to a successful extraction of the main points in the listening texts. In addition, high scorers monitored their understanding of the text and used a wider variety of metacognitive strategies. In contrast, lower scorers used a narrower range of metacognitive strategies and were less successful in applying higher-level cognitive strategies (i.e., inferencing and elaboration).

Measuring Content in Speaking Performance

The planning times reviewed by those studies did not include measures of content. The effect of planning time on production and reproduction of content in speaking task performance therefore remains an area to be further explored. In addition to the general CAF variables, Ellis and Barkhuizen (2005) suggested exploring IUs as a measure of propositional complexity in the elicited performance. In integrated speaking tasks, the reproduction of ideas from the input text forms a crucial part in the overall quality of the output text. Pioneering this line of research on integrated speaking test tasks, Frost et al. (2011) investigated performances in the listening-into-speaking tasks in the Oxford English Proficiency Test, where the types and accuracy of reproduced IUs were employed as performance measures. The study found that higher scoring candidates gave speaking performances with more accurate reproduction of individual IUs and less use of verbatim phrases from the listening input. However, the researchers did not find evidence of summarization and condensing or combining IUs even with higher scoring candidates. This finding may not be surprising, considering that the listening time and speaking time were of the same length (i.e., 2 min); there was no real need for summarizing the listening input. In contrast, the need to summarize and condense IUs is more likely in the 1-min speaking time for summarizing a 2-min lecture input in the listening-into-speaking task in this study. We postulate that an extended planning time might offer more opportunities for learners to plan for language not only in the sense of CAF but also potentially in paraphrasing ideas more in their own words,[1] especially at higher proficiency levels. More summarization as learners may do in real life when reporting and discussing lecture content might also be observed.

Research Questions

The review of literature pointed to a need for more research on the effect of planning time on the content of speaking performance (in terms of reproduction of input in integrated tasks) and on the cognitive and metacognitive strategies candidates engaged in pretask planning (i.e., what candidates actually do in their planning time; Wigglesworth & Elder, 2010). We hypothesized that with an extended planning time candidates would engage to greater extents in planning for content and language as well as mental rehearsals of their speech, especially among higher proficiency candidates. Such increased planning might in turn allow candidates to produce better linguistic performances in terms of CAF and reproduction of the lecture input, which would lead to higher scores on the lecture listening-into-speaking tasks in the speaking section of the TOEFL iBT test. Accordingly, the following three research questions were formulated:

RQ1. How does the extended planning time affect the overall and analytic scores on the lecture listening-into-speaking tasks at higher and lower proficiency levels?

RQ2. How does the extended planning time affect the linguistic performance at higher and lower proficiency levels in terms of (a) CAF variables and (b) reproduction of ideas from the listening text on the lecture listening-into-speaking tasks?

RQ3. What do candidates at higher and lower proficiency levels report to have done during the listening, planning, and speaking time for the lecture listening-into-speaking tasks under the operational and extended planning time conditions?

Methods

Participants

A total of 70 participants[2] were recruited, comprising 35 in the higher proficiency group (mean age: 26.9, *SD* = 2.9) and 35 in the lower proficiency group (mean age: 24.4, *SD* = 7.6).

For the purpose of exploring any differences in the nature and extent of the effects of extended planning time on performance among higher proficiency and lower proficiency learners, we recruited participants with IELTS Speaking scores 7.0 or above (i.e., TOEFL Speaking score 23 or above[3]), assigning them into the higher proficiency group, and participants with IELTS Speaking scores 5.0–6.0[4] (i.e., TOEFL Speaking score 14–19), assigning them into the lower proficiency group. All participants were required to supply a copy of their test score certificate.

Some of the participants had not taken an official IELTS Speaking test at the time of this study (*n* = 1 in the higher group; *n* = 14 in the lower group). Many of the lower group participants, in particular, were on presessional English programs studying toward meeting the minimum requirement (IELTS 6.0) for admission into academic programs. Because recruiting participants with official IELTS Speaking scores of 5.0–5.5 (i.e., those failing to meet the minimum requirement for university admission) presented difficulty with access, we accepted speaking scores by the language center at the researchers' university, where scores were reported using IELTS score bands and students were assessed by staff who were certified IELTS Speaking examiners.

The IELTS Speaking scores among the high group participants ranged between 7.0 and 8.5 (*M* = 7.56, *SD* = 0.56), whereas the low group's scores ranged between 5.0 and 6.0 (*M* = 5.53, *SD* = 0.34). The higher group participants had, on average, a 1.21-year gap (*SD* = 0.72) between the time they took their IELTS test and when the data collection session took place, whereas the lower group participants had a shorter gap on average (*M* = 0.83 years, *SD* = 0.58).

As the study was carried out in the United Kingdom, the overwhelming majority of the participants had not taken the TOEFL iBT test and thus were unable to supply a score for the speaking section for proficiency benchmarking. Only two participants in the higher group reported having experience with the TOEFL iBT test. All participants underwent a training and practice session (as detailed in the Practice Task section), which was aimed at familiarizing them with the lecture listening-into-speaking tasks.

Both the higher and lower group participants represented a range of nationalities and primary language (L1) background. For the higher proficiency group, just above half were Chinese (*n* = 19), and the nationalities of the other participants included Chilean, Indian, Indonesian, Kenyan, Macedonian, Mexican, Rwandan, Syrian, and Tajik. For the lower proficiency group, the majority were Chinese (*n* = 18) and Romanian (*n* = 10), and the rest were Greek, Italian, Japanese, and Polish.

Table 2 Combination of Planning Time and Tasks for Each Task Set

Set	Planning time for Task 1	Topic for Task 1	Planning time for Task 2	Topic for Task 2
A	20 s	Reciprocity	90 s	Fungus
B	90 s	Fungus	20 s	Reciprocity
C	20 s	Fungus	90 s	Reciprocity
D	90 s	Reciprocity	20 s	Fungus

Instruments

Pilot Study

Based on the literature review, we conducted a pilot study to test five planning time lengths (20 s, 1 min, 1.5 min, 2 min, and 3 min) with 10 participants (five in each proficiency group). The assignment of these participants into the higher or lower proficiency group was based on their official IELTS Speaking scores[5] or the language center assessment (reported in IELTS score bands). With copyright permission from ETS, we used five task versions from the *Official TOEFL iBT® Tests, Volume 2*. Given the small sample size, the task versions were not counterbalanced with planning time lengths to cancel out the topic effect.

Extended Planning Time

The pilot study suggested that participants' performance might slightly improve with 1-min or 1.5-min planning time lengths in variables of fluency, accuracy, and reproduction of IUs (see below for details about the variables used in this study), and participants started reporting planning for language and mental rehearsals with 1-min and 1.5-min planning time). Considering the higher cognitive demand of integrated tasks than independent tasks (Barkaoui et al., 2013; Brown et al., 2005; Frost et al., 2011), we decided to use 1.5 min (i.e., 90 s) in this study.

Tasks

Three task versions were supplied by ETS, one as a practice task (Product Quality), and two as the research tasks (Reciprocity and Fungus). For the two main tasks, four task sets (A, B, C, D) were created using the TOEFL iBT Form Creator Software,[6] where the two tasks, Planning Time and Order of Presentation, were counterbalanced (see Table 2). The participants met with one of the researchers on a one-to-one basis for data collection and were randomly assigned to one of the forms.

Questionnaire

A questionnaire was developed for collecting data on the self-reported use of cognitive processes and strategies, as well as metacognitive strategies (Appendix A), based on the relevant works of Huang (2016), Ortega (2005), Pang and Skehan (2014), Rukthong (2016), and Swain et al. (2009), as well as the results from the pilot study. Participants indicated the level of agreement to each of the 42 statements in the questionnaire on a 5-point scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *neither disagree nor agree*, 4 = *agree*, 5 = *strongly agree*). For ensuring the ease of responding, the questionnaire focused largely on participants' behavior relevant to the use of cognitive strategies (e.g., prediction) and metacognitive strategies (e.g., selective attention).

All participants who did not take part in stimulated recalls ($n = 54$) completed the questionnaire on paper immediately after performing each task. The participants who provided stimulated recall data ($n = 16$) answered the questionnaire after they had finished the stimulated recall session for each task.

Stimulated Recall

Sixteen (eight from each proficiency group) were recruited for a longer data collection session with a stimulated recall interview (Gass & Mackey, 2000), which immediately followed the completion of each task. For these 16 participants, two types of behavior were video-recorded during their task performance to be used as stimuli for the stimulated recall sessions: one was of their note-taking behavior during listening and planning time, which was recorded with a webcam set up near their writing hand and the note sheet, and the other was of their behavior during speaking time, which was recorded with a webcam set up to include the test screen, notes, and their face and upper body. In the stimulated recall sessions, each participant was shown both video recordings in chronological order. The researcher paused the videos at regular intervals, asking questions such as, "What was going on in your mind here?" and "So I see you were writing '[read out parts of the notes in the video].' What were you trying to do at this point?" The participants were also encouraged to make comments whenever they wanted to at any moment.

Practice Task

A practice task was used to familiarize participants with performing the listening-into-speaking tasks as well as responding to the questionnaire (and stimulated recall for 16 of the 70 participants). The familiarization/training started with the researcher giving the participant a brief description of the integrated speaking task (listen to a 2-min lecture consisting of two main ideas and examples for each, then summarize it in speaking within 1 min) and showing them a summary of the scoring criteria. The participant then performed the practice task and completed the questionnaire (and stimulated recall) with the opportunity to ask the researcher any clarification questions.

For the stimulated recall participants, the researcher informed them about the kinds of questions they would be asked before the practice task and provided feedback following the stimulated recall of the practice task, such as "We are interested in what you were thinking/doing rather than explaining why you did certain things" and "In asking what you were thinking/doing at a particular time, we are not questioning how you approach the test task." Again, participants were given opportunities to ask questions before moving on to the research tasks.

Data Analysis

Scores

The 140 task responses (i.e., two responses per participant) were audio-recorded and sent to ETS-certified raters. All task responses were single scored, as per operational scoring in the speaking section of the TOEFL iBT test at the time of data collection,[7] by a team of eight raters. Each task response was awarded a holistic score (0–4) by one rater and analytic scores (0–4) by another rater for the four criteria: Pronunciation, Pace, Language Use, and Topic Development. A mixed two-way ANOVA was carried out on the overall and analytic scores given by the raters.

It should be noted that the criteria of Language Use and Topic Development were identical to the operational TOEFL iBT integrated speaking rubrics (see Appendix B). However, the criterion of Delivery was separated into Pronunciation and Pace to align the descriptors in the rubrics to what is measured by the CAF variables (for RQ2) more closely. Specifically, fluency would at least partially correspond with the descriptors in Pace, and intrusive pronunciation errors would be reflected in the descriptors in Pronunciation and in Language Use (regarding accuracy). The descriptors for Pronunciation and Pace were adapted from the operational descriptors for Delivery by the researchers (see Appendix B). The rater training for using these two criteria separately was conducted by ETS. Sample transcripts with scores of a participant from each proficiency group are presented in Appendix C.

CAF Variables

The recorded performances were transcribed, segmented into AS-units (Foster et al., 2000), and coded for the variables listed in Table 3. A mixed two-way ANOVA was conducted on the resultant values for each variable.

Among a huge number of variables for each aspect of performance that have been used in previous studies, the variables in Table 3 were chosen on the basis of the research evidence of their validity within specific contexts and studies as well as their degrees of match with the performance descriptors in the TOEFL iBT integrated speaking rubrics.

For fluency, *speech rate* (syllables per minute) was used because it was shown to correlate very highly with perceived fluency, whereas the variables related to hesitation phenomena did not (Kormos & Dénes, 2004). There was also a practical advantage in using speech rate due to its ease of calculation because the number of syllables could be identified quickly and reliably using Text Inspector (https://textinspector.com/), an online text analysis tool.

Table 3 Complexity, Accuracy, and Fluency Variables Used in This Study

Dimension	Variable
Fluency	Syllables per min
Accuracy	Errors per 100 words
	Weighted errors per clause ratio
Syntactic complexity	Words per AS-unit
	Words per clause
	Subordinate clauses per AS-unit
Lexical complexity	Vocd-D

Note. AS-unit = analysis of speech unit.

For accuracy, two variables were selected. Mehnert (1998) argued for *errors per 100 words* because it does not rely on the identification of clause-based units, which can be problematic. Later, Inoue (2016) demonstrated that this variable correlated more highly with perceived accuracy than the more commonly used variable of accuracy, namely, the percentage of error-free clauses (e.g., Skehan & Foster, 1999; Yuan & Ellis, 2003). The other variable of accuracy selected in the present study was *weighted errors per clause ratio* (Foster & Wigglesworth, 2016). It takes into account both the error gravity (i.e., the degree to which an error interferes with meaning) and clause-based complexity, which seems to correspond well with the Language Use descriptors of the TOEFL iBT integrated speaking rubrics. The errors were identified and tallied into three categories defined by Foster and Wigglesworth (2016) according to the degree of interference with meaning.

For lexical complexity, the vocd-D value was selected and calculated using Text Inspector. Jarvis (2002) demonstrated that this variable was most reliable among different variables of lexical variety. Lexical sophistication was not investigated in the present study because it has been shown to be largely affected and restricted by the source texts in integrated tasks (Kyle & Crossley, 2016).

For syntactic complexity, based on the comprehensive review of numerous variables by Norris and Ortega (2009), three variables were chosen to capture the different ways in which the participants might syntactically complexify their performance: *words per AS-unit* (e.g., Tavakoli & Foster, 2008), s*ubordinate clauses per AS-unit* (e.g., Crookes, 1989; Mehnert, 1998), and *words per clause* (Norris & Ortega, 2009).

Reproduced Idea Units

All 140 transcribed performances (i.e., two per participant) were segmented for IUs using the slightly modified definitions and criteria based on Frost et al. (2011) and Frost et al. (2020; see Appendix D). The IUs in the participants' performances were then coded according to the types of correspondence with the IUs in the lectures: unit-for-unit correspondence (individual IU [IIU]), combined unit correspondence (combined IU [CIU]), or further integration in the form of summarizing, generalizing, or inferencing (macroproposition [MP]). The IUs were then coded as either *accurate* or *inaccurate*[8] reproduction in order to calculate the proportion of accurately reproduced IUs for each IU type as these examples show:

1. Accurate reproduction
    a. IIU
        i. Lecture: And the first type I wanna talk about is generalized reciprocity.
        ii. Participant: The first one is generalized reciprocity.
    b. CIU
        i. Lecture: Now there's also fungus inside an old tree. / Fungus feeds on that dead wood, / it literally eats it up / (Note: Non-adjacent IUs).
        ii. Participant: The fungus inside the trunk of a tree eat up the dead wood.
    c. MP
        i. Lecture: Generalized reciprocity only works among people who are close / … / Here, [for balanced reciprocity,] the social distance between the giver and the receiver is greater than with generalized reciprocity.

        ii. Participant: The differences between them lies between the social distance between the givers and also the receivers.

2. Inaccurate reproduction
    a. IIU
        i. Lecture: This is when people give each other goods or gifts, without expecting anything in return immediately.
        ii. Participant: The generalized one is about goods or gifts that people give to other people.
    b. CIU
        i. Lecture: See, the trunk of an old tree is full of dead wood. / … / Fungus feeds on that dead wood, / … / and the trunk becomes empty inside, hollow.
        ii. Participant: Inside the dead woods, there are hollows that they feed it.
    c. MP
        i. Lecture: See, the trunk of an old tree is full of dead wood. / … / Fungus feeds on that dead wood/
        ii. Participant: The first way is the fungus can protect the trees through death woodders.
3. Others: Points not mentioned in the lecture input, such as participants' own views (e.g., And actually I do not quite familiar with it), IUs carrying a structuring discourse or rhetorical function (e.g., That's all.), IUs repeating what has already been mentioned in previous IUs, or incomplete utterances (e.g., They give something …).

In order to examine the effects of the two planning time conditions on the proportion of accurately reproduced IUs, we used R (R Core Team, 2020) with lme4 (Bates et al., 2015) to perform mixed effects logistic regression with the glmer function. It was used to model binary outcome variables (i.e., accurate or inaccurate reproduction of IUs), which regressed the log odds of the outcome probability on the predictor variables (i.e., planning time conditions) with the fixed effects (i.e., proficiency level) and random effects (i.e., repeated design). Mixed effects logistic regression does not assume any particular distributions of variables because it models with binary[9] and categorical data.

Questionnaire Responses

The responses to each statement in the questionnaire were compared between the two planning conditions using Wilcoxon signed rank tests. The Wilcoxon signed rank test was chosen because it allowed comparisons of the ordinal dependent variable for matched pairs (i.e., candidates indicated degrees of agreement to statements twice under the operational and extended planning time conditions). This test's symmetry assumption was examined through the skewness value of the distribution of the differences between the responses on each questionnaire statement (see Appendix E). As can be seen in Appendix E, the skewness values to all the statements for both proficiency groups were within the acceptable range for a reasonably symmetrical distribution of −2 to +2 (George & Mallery, 2010), except for four statements (i.e., Statement 20 for the low group and Statements 3, 10, and 30 for the high group). For these four statements, related-sample sign tests were used instead.

Conducting a Wilcoxon signed rank test or a sign test for each statement for each proficiency group meant that multiple tests were performed on the same set of data. However, a decision was made to not adjust the alpha level, as doing so would drastically increase the probability of committing Type II errors (i.e., accepting the null hypothesis when it is wrong). Avoiding Type II errors was deemed as more important than Type I errors (i.e., rejecting the null hypothesis when it is correct) because the analysis of the questionnaire responses was exploratory and would feed into the subsequent analysis of the stimulated recall data. The nonadjustment of alpha levels is justified in Armstrong (2014) where multiple tests are used as pointers for further analysis.

Stimulated Recall Data

The transcribed stimulated recall data were thematically analyzed using NVivo 12. The statements in the questionnaire were used as a coding scheme, and emerging new themes and subthemes were added as the coding progressed. When a participant reported on more than one process and/or strategy at the same time, the relevant parts of the recall data were coded for multiple categories (cf. Rukthong & Brunfaut, 2020). The coded data were used to gain more in-depth insights into the results of the questionnaire responses in terms of the thinking and behavior by the participants during task performance.

Coding Reliability

In order to warrant coding reliability, 10% (14 of 140) of the transcribed performances (for RQ2: AS-units, errors, subordinate clauses, and IUs) and 12.5% (four of 32) of stimulated recall verbal reports (for RQ3: cognitive processes and cognitive and metacognitive strategies) were double coded by the two researchers. For RQ2, the percentages of exact agreement between the two sets of independent coding reached 93.6% for the number and levels of errors, 96.9% for AS-units, 86.6% for subordinate clauses, and 86.0% for the types and accuracy of reproduced IUs. For RQ3, the exact agreement reached 87.1% for the types of processes and strategies in the stimulated recall data. All discrepancies were discussed and resolved, and then one of the researchers continued to code the remaining data.

Results

Scores (RQ1)

The descriptive statistics of the holistic and analytic scores[10] for the task responses across the planning time conditions and proficiency groups are shown in Table 4. The Levene's tests confirmed the equality of variances for all the holistic and analytic scores for both planning time conditions except for Pronunciation for 90 s planning time ($F[1,68] = 4.67, p = .03$), Pace for 90 s planning time ($F[1,68] = 8.32$, $p = .01$), and Language Use for 20 s planning time ($F[1,68] = 5.29$, $p = .03$). Despite these violations of the normality assumption, a decision was made to use the parametric tests because ANOVA is said to be sufficiently robust to use even on nonnormally distributed data (Tabachnick & Fidell, 2013). Additionally, there is no nonparametric equivalent to two-way mixed ANOVA that tests the interaction effects of the two proficiency levels and two planning time conditions with repeated samples.

The results of a two-way mixed ANOVA indicated that the effect of planning time on participants' holistic scores was not statistically significant ($F[1,69] = .00$, $p = 1.00$). Similarly, there was no statistically significant main effect of planning time on any of the analytic scores (Pronunciation: $F(1,69) = .03$, $p = .87$; Pace: $F(1,69) = .00$, $p = 1.00$; Language Use: $F(1,69) = .24$, $p = .63$; Topic Development: $F(1,69) = .02$, $p = .89$). Furthermore, no statistically significant interaction was found between planning time and proficiency group in terms of holistic ($F[1, 69] = 0.89$, $p = .35$) or analytic scores (Pronunciation: $F(1,69) = .70$, $p = .41$; Pace: $F(1,69) = .00$, $p = 1.00$; Language Use: $F(1,69) = .67$, $p = .42$; Topic Development: $F(1,69) = .95$, $p = .33$).

Linguistic Performance (RQ2)

CAF Variables

The Levene's tests confirmed the equality of variances for all the variables for both planning time conditions except for those of accuracy (errors per 100 words for 20 s planning: $F(1,68) = 19.91$, $p = .00$ and for 90 s planning: $F(1,68) = 15.04$,

Table 4 Descriptive Statistics of Scores

	Low ($n = 35$)				High ($n = 35$)			
	20 s		90 s		20 s		90 s	
Score	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Overall	2.14	0.77	2.06	0.76	3.11	0.63	3.20	0.72
Pronunciation	2.46	0.56	2.51	0.82	3.43	0.56	3.34	0.59
Pace	2.23	0.77	2.17	0.79	3.20	0.63	3.14	0.55
Language use	2.26	0.74	2.14	0.81	3.14	0.55	3.17	0.62
Topic development	2.29	0.79	2.17	0.95	3.26	0.66	3.34	0.73

Table 5 Statistics for Complexity, Accuracy, and Fluency Variables

		Low ($n = 35$)				High ($n = 35$)				Main effect of planning time			Interaction between planning time and proficiency group		
		20 s		90 s		20 s		90 s							
Dimension	Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*	*df*	*p*	*F*	*df*	*p*
Fluency	Syllables per minute	117.03	37.78	113.34	37.16	182.00	39.93	178.34	39.60	.82	69	.37	.0	69	1.00
Accuracy	Errors per 100 words	7.22	4.78	6.96	4.67	2.92	1.90	2.60	2.29	.54	69	.47	.0	69	.94
	Weighted errors per clause ratio	0.89	0.07	0.88	0.11	0.95	0.05	0.96	0.05	.251	69	.62	1.39	69	.24
Syntactic complexity	Words per AS-unit	11.12	2.52	11.72	3.17	14.05	3.33	14.04	3.23	.45	69	.5	.51	69	.48
	Subordinate clauses per AS-unit	0.60	0.42	0.65	0.48	0.73	0.42	0.74	0.36	.28	69	.59	.14	69	.71
	Words per clause	7.07	0.99	7.18	1.23	8.22	1.43	8.15	1.55	.01	69	.94	.18	69	.67
Lexical complexity	vocd-D	46.53	15.56	45.29	17.47	50.06	14.28	52.61	17.27	.04	64	.84	.36	64	.55

Note. AS-unit = analysis of speech unit.

Table 6 Proportion of Reproduced Idea Units

	Low ($n = 35$)				High ($n = 35$)			
	20 s		90 s		20 s		90 s	
Idea unit	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
IIU	.63	.20	.71	.17	.67	.15	.61	.22
CIU	.21	.19	.12	.13	.18	.11	.19	.19
MP	.17	.15	.17	.12	.15	.10	.20	.21

Note. IIU = individual idea unit; CIU = combined idea unit; MP = macroproposition.

Table 7 Proportions of Accurately Reproduced Idea Units

	Low ($n = 35$)				High ($n = 35$)			
	20 s		90 s		20 s		90 s	
Idea units	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
IIU	.66	.28	.59	.25	.86	.15	.84	.23
CIU	.37	.43	.33	.45	.65	.40	.67	.42
MP	.21	.41	.34	.42	.57	.43	.48	.42

Note. IIU = individual idea unit; CIU = combined idea unit; MP = macroproposition.

$p = .00$) as well as weighted errors per clause ratio for 20 s planning ($F(1,68) = 9.08$, $p = .00$) and for 90 s planning ($F(1,68) = 22.72$, $p = .00$). Similar to the score analysis (RQ1), despite these violations of the normality assumption, a decision was made to use two-way mixed ANOVA due to its robustness with nonnormally distributed data (Tabachnick & Fidell, 2013) and the lack of nonparametric equivalent.

The results of a two-way mixed ANOVA revealed that there was no significant main effect of planning time or interaction between planning time and proficiency level on any of the CAF variables. The descriptive and inferential statistics for each of the CAF variables across the planning time conditions and proficiency groups are shown in Table 5.

Reproduced Idea Units

The descriptive statistics for each type of IUs across the planning time conditions and proficiency groups are shown in Table 6, and those for the accurately produced IUs are summarized in Table 7. As can be observed in Table 7, the candidates in the higher proficiency group seemed to have generally reproduced more IUs accurately in their responses.

The mixed effect logistic regression models used CIU as a reference and tested for significant main effects and interaction on the other two IU types (i.e., IIU and MP). Table 8 shows that candidates were almost 5 times more likely

Table 8 Effects of Types of Idea Units, Planning Time, and Proficiency Levels

Type	Estimate	*SE*	*Z*	*p*
IIU	1.60	.30	5.35	.00***
MP	−.71	.27	−2.62	.00***
Planning	.05	.33	.16	.87
Proficiency	−1.67	.26	−6.37	.00***
Planning × proficiency	−0.00	.46	−0.01	.99

Note. IIU = individual idea unit; MP = macroproposition. ***Significant at <.001 level.

to reproduce IIUs accurately than CIUs (β = .1.60, *SE* = .30, $z[420] = 5.35$, $p = .00$; with an odds ratio of 4.95[11]) and 2 times more likely to produce CIUs accurately than MPs (β = −.71, *SE* = .27, $z[420] = -2.62$, $p = .00$; with an odds ratio of .49[12]), which indicates the likelihood of accurate production of MP was less than half that of CIU. Moreover, the candidates in the lower proficiency group were nearly one-fifth times less likely to reproduce IUs accurately than those in the higher group (β = −1.67, *SE* = .26, $z[420] = -6.37$, $p = .00$; with an odds ratio of .19[13]), which, again, confirms the appropriate grouping of the candidates according to proficiency levels. However, as shown in Table 8, the results for the planning time revealed no statistically significant main effect of the length of planning time on the proportion of accurately reproduced IUs (β = .05, *SE* = .33, $z[420] = .16$, $p = .87$) and no significant interaction was found between planning time and proficiency group (β = .002, *SE* = .46, $z[420] = -.005$, $p = 1.00$).

Self-Reported Use of Processes and Strategies (RQ3)

This section reports on the results of analysis from the questionnaire and stimulated recalls together in order to explore the extent to which participants engaged in the stated processes and strategies, as well as why and/or how exactly they did so. The following sections are organized in the chronological order of task completion—listening, planning, and speaking—with subheadings of the relevant stated processes and strategies in the questionnaire or those that emerged from the stimulated recall analysis. The descriptive statistics of all the questionnaire statements can be found in Appendix F.

Listening

No significant differences were found between the two planning time conditions in the reported processes and strategy use on the questionnaire during listening according to Wilcoxon signed rank tests or sign tests (for Statements 3 and 10 for the higher group). Nevertheless, it is worth noting that the majority of participants, regardless of proficiency levels or planning time, engaged in the stated processes and strategies (as indicated by a mean of 3.8 or higher and a median of 4 or higher, where 4 indicates *agree* and 5 indicates *strongly agree*) on six statements. Table 9 presents each of the relevant statements (with the statement number used in the questionnaire) and its mean and median for both planning time conditions and proficiency groups.

Contrary to the questionnaire results, subsequent analysis of the stimulated recall data highlighted some qualitative differences between participants in the lower and higher proficiency groups in terms of how they were engaging in the stated processes and behaviors related to listening.

Recognizing the Lecture Structure

The analysis of stimulated recall found that participants in the high proficiency group often reported successful recognition of the lecture's structure, whereas unsuccessful recognition and prediction were reported by both lower and higher groups. Relevant excerpts are shown here (Excerpt 1 shows successful recognition; Excerpt 2 shows unsuccessful recognition. R indicates researcher; P indicates participant; H before a participant ID number indicates a participant from the higher proficiency group and L indicates a lower proficiency group):

***Excerpt 1** (H715, 90 s)*

R: And then now you do the arrow [which branches down to the left from 'reciprocity' on the notes] when she said the first type.

Table 9 Statements With High Agreement Across Groups and Planning Time at Listening Stage

		Low ($n = 35$)							High ($n = 35$)						
		20 s		90 s					20 s		90 s				
Statement		*M*	Median	*M*	Median	*Z*	*p*	*r*	*M*	Median	*M*	Median	*Z*	*p*	*r*
2	I tried to identify the main points	4.3	4	4.1	4	−1.13	.26	−.19	4.3	4	4.3	4	.14	.89	.02
3	I tried to connect the examples to the main points	4.2	4	4.1	4	−.94	.35	−.16	4.4	4	4.5	4	.00[a]	1.00	
5	I tried to recognize the structure of the lecture	3.9	4	3.8	4	−.47	.64	−.08	4.2	4	4.3	4	1.51	.13	.26
9	I wrote down key words and phrases	4.1	4	4.1	4	.15	.88	.03	4.7	5	4.6	5	−.38	.71	−.06
10	I made general notes about the main points	4.0	4	4.0	4	−.35	.73	−.06	4.6	5	4.7	5	.50[a]	.63	
12	I tried to memorize important information	4.1	4	4.1	4	−.76	.45	−.13	4.1	4	4.1	4	.00	1.00	.00

Note. 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree. [a]From related-samples sign test.

P: Two types. So while writing I was listening to the lecture and I heard she was talking about two types so I already—that's why I made this arrow for the first type.

Excerpt 2 *(L527, 90 s)*

Ok, in this moment, I lost it. I lost it because I did not heard […] if she already passed to the second type of reciprocity. I was thinking that she was still talking about this one [generalized reciprocity] …

Here, L527 failed to recognize that the lecturer had already moved on from the first type of reciprocity to the second. Being "lost" and missing some key information about the second main idea of the lecture had repercussions on how he used the extended planning time. He reported using the time trying to understand the concepts that he did not manage to hear while listening.

Taking Notes

Participants from both proficiency groups reported making notes of key words, phrases, and main points in both planning time conditions in the questionnaire (Statements 9 and 10). Nevertheless, among higher proficiency group participants, a number of factors seemed to have been considered quickly before arriving at the decisions as to whether and how they noted something. Such factors included their note-taking speed, available time (Excerpt 3), predicted sequence of the lecture, and the perceived importance of information (Excerpt 4).

Excerpt 3 *(H710, 20 s)*

R: Ok so now you wrote down "tree" and "dead wood."

P: "Dead wood" was sort of highlighted that speech she made at that particular point, it was related to tree. So I decided [not] to write complete sentence, because that would take a lot of time and I would miss [what's coming], and that's why I marked that down like 'tree.'

Excerpt 4 *(H728, 90 s)*

R: Ok, so you added the word 'trust,' I guess that's what you were about to write and also you highlighted [underlined] 'you know.'

P: […] I highlighted it because I thought I do not have the word to describe this, [and] she was carrying on to explain different things, and I thought I need to listen to these other things, but, this you know are very important, this definition. So I am just gonna highlight it, and I am gonna write 'trust' which is new word that she is saying and might be a key word.

In contrast, some decisions seemed to have been made without much consideration or understanding, which were more characteristic in the stimulated recall data from the lower proficiency group participants.

Excerpt 5 *(L502, 90 s)*

R: Ok so at the start of the lecture she mentioned …

P: I do not hear that much because I was trying to write the words to find what it means 'fungus.'

R: So you heard the word 'fungus' and you were trying to write it down.

P: Yeah.

R: Although you do not understand what it means.

S: Yeah I tried to write the word. I think it was more like the on the excer [external?] of tree. When I tried to write the word [fungus], I do not hear the [word] mushroom.

In Excerpt 5, this participant did not understand the keyword "fungus." However, fixated on trying to note down the word, he did not attend to (let alone evaluate the importance of) further incoming pieces of information. He ended up missing the gloss, "little mushrooms," provided by the lecturer. Information missed during listening is unlikely to be retrievable (recalled) later during planning time and speaking performance and, therefore, may bear consequences for the usefulness of planning time and, more directly, what and how much content can be reproduced during the speaking performance.

Planning

For the planning stage, several questionnaire statements revealed a statistically significant difference between the two planning time conditions in each proficiency group according to Wilcoxon signed rank tests. Table 10 presents each of the relevant statements (with the statement number used in the questionnaire) and its mean and median for both planning time conditions and proficiency groups. The larger means for 90 s planning time for all the statements in either or both of the groups in Table 10 indicate that participants tended to engage in the stated strategic behavior when planning with 90 s of planning time.

Planning for Content

According to the questionnaire results, participants in the higher group tended to heed the question on the screen regardless of planning time duration, whereas participants in the lower group paid more attention to the question when they had 90 s planning time (Statement 13).

Moreover, both proficiency groups indicated engaging more with reviewing of the notes to organize ideas (Statement 17) and adding marks or numbers to the notes to guide speech (Statement 18) under the 90 s planning condition. In the stimulated recall sessions, participants from both groups reported using the extended planning time to write additional notes from memory, but only participants in the higher proficiency groups mentioned filtering out less important ideas from their notes during planning (Excerpt 6). This process of filtering nonessential information may be related to participants in the higher proficiency groups reporting more often that they tried remembering the lecture content (Statement 14).

Excerpt 6 *(H715, 90 s)*

> I was thinking about whether to start about 'rules and norms,' 'goods and services' in general and narrow it down, or straight away start with reciprocity and its benefits. […] I was thinking about time and I assume that it will not be possible for me to start from here, from general, so it's better to […] be up to the point

Table 10 Statements With a Statistically Significant Difference at Planning Stage

		Low ($n = 35$)							High ($n = 35$)						
		20 s		90 s					20 s		90 s				
	Statement	*M*	Median	*M*	Median	*Z*	*p*	*r*	*M*	Median	*M*	Median	*Z*	*p*	*r*
13	I looked back at the question on the screen	3.0	3	3.5	4	2.44	.02*	.41	3.3	4	3.4	4	1.03	.30	.17
14	I tried to remember what was said in the lecture	3.9	4	3.7	4	−1.01	.32	−.17	3.6	4	3.9	4	2.18	.03*	.37
17	I reviewed the notes to organize ideas	3.7	4	4.1	4	2.04	.04*	.34	4.0	4	4.5	5	2.49	.01*	.42
18	I added marks or numbers to the notes to guide my speech	3.3	4	3.8	4	2.33	.02*	.39	3.2	4	4.1	4	2.75	.01*	.46
20	I tried to think of more difficult words to use	2.5	2	2.5	3	.00[a]	1.00		2.1	2	2.9	3	2.85	.00*	.48
21	I tried to think of sentences to use in my speech	3.4	4	3.5	4	.86	.39	.15	3.0	3	3.9	4	3.82	.00*	.65
22	I tried to think of differently worded sentences to use	2.9	3	3.2	3	1.46	.14	.25	2.8	2	3.4	4	2.66	.01*	.45
23	I wrote down sentences on my notes to use in my speech	2.7	3	3.1	3	1.74	.08	.29	2.5	2	3.4	4	3.35	.00*	.57
24	I thought about how to pronounce particular words	2.9	2	3.5	4	2.30	.02*	.39	3.0	3	3.1	3	.22	.83	.04
25	I practiced some sentences in my mind	3.5	4	3.7	4	1.28	.20	.22	3.0	3	3.6	4	2.50	.01*	.42

Note. 1 = *strongly disagree*, 2 = *disagree*, 3 = *neither agree nor disagree*, 4 = *agree*, 5 = *strongly agree*. [a]From related-samples sign test. *Significant at $p < .05$.

> and talk about the main point that were in the lecture. So I started with reciprocity and I missed this part, the introductory part.

Planning for Language

The questionnaire results showed significant differences in behaviors related to planning for language (Statements 20–23) between the two planning time conditions for the higher proficiency group. In stimulated recall, participants from both groups said that 20 s of planning time was too short to "write full sentences" (L518, 20 s) or "think about the sentence structures" (H717, 20 s). For 90 s planning time, the same participant from the higher proficiency group reported having been able to do so:

t

***Excerpt 7** (H717, 90 s)*

> Back then I think I have enough time to review my notes and to build up some connections. The rest of the seconds I am trying to go over some structures and sentences [which] probably I'm going to use in my speech.

In addition, some participants in the higher proficiency group reported thinking of using words and sentences different from the lecture (Statements 20 and 22) with the 90 s planning time "for maximizing the language variety" (H709, 90 s), showcasing their proficiency in the hopes of achieving a higher score.

Furthermore, a contrast between the proficiency groups was observed in the questionnaire data regarding writing sentences during planning (Statement 23) where a significant difference was found only in the higher proficiency group. The stimulated recall data revealed that "writing down sentences on my notes," as phrased in Statement 23, does not necessarily mean constructing full sentences from scratch but adding words/phrases onto their notes. For instance, H709 (90 s) reported adding phrases to create a sentence to use in the introduction of his speech, with "something easy, common

Table 11 Statements with a statistically significant difference at speaking stage

		Low ($n = 35$)							High ($n = 35$)						
		20 s		90 s					20 s		90 s				
	Statement	*M*	Median	*M*	Median	*Z*	*p*	*r*	*M*	Median	*M*	Median	*Z*	*p*	*r*
31	I tried to use difficult words	2.5	2	2.5	2	−.03	.97	−.01	2.3	2	2.9	3	2.94	.00*	.50
32	I tried to use complex sentences	2.5	2	2.5	2	−.10	.92	−.02	2.4	2	2.8	3	2.18	.03*	.37

Note. 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree. *Significant at $p < .05$.

to keep it going to make the start." In contrast, some participants in the lower group revealed how they tried to write down full sentences in the 90 s of planning time but did not accurately estimate what could be done in time (Excerpt 8):

> ***Excerpt 8** (L501, 90 s)*
>
> I started to take notes,* and once I finished a paragraph, it [the 90 s] just finished. I did not have time to write down everything.
> *His notes show that he was writing the speech in full sentences.

Mental Rehearsal

As indicated in the questionnaire results, the lower proficiency group did more mental rehearsal at the word level (Statement 24), whereas the higher proficiency group rehearsed more at the sentence level (Statement 25). Although the lower proficiency group did not mention episodes of mental rehearsal in the stimulated recall sessions, Excerpt 9 illustrates some strategic ways in which some participants in the higher proficiency group might have engaged in mental rehearsal at the sentence level.

> ***Excerpt 9** (H709, 90 s)*
>
> I think with 90 seconds, I could do the full rehearsal. So I would do a full rehearsal, and as I go, I would see gaps in my notes where I need to probably make clearer and fill in more gaps for the purpose. […] I feel like instead of just going over it in my mind, actually making some sounds […] would help warm your tongue up … like kick-starts the engine.

H709 stated that he (quietly) mouthed some sounds with the thought of improving fluency during speaking and iteratively revised his notes. He also explicitly reported not having the time for mental rehearsal with 20 s of planning time.

Speaking

In the speaking section of the questionnaire, participants in the higher proficiency group reported trying to attend more to lexical complexity and syntactic complexity (Statements 31 and 32, respectively) under the 90 s planning condition, according to the Wilcoxon signed rank tests. Table 11 presents each of the relevant statements (with the statement number used in the questionnaire) and its mean and median for both planning time conditions and proficiency groups. In the stimulated recall, there was no mention of attending to different dimensions of performance while speaking. However, as evidenced in the earlier sections on planning for language, participants in the higher proficiency group paid attention to improving complexity (by using different words and structures) and fluency (by rehearsing speech) while planning. Perhaps they responded to the questionnaire statements remembering these behaviors during the planning time, although the statements were actually placed in the speaking time section in the questionnaire.

Discussion and Conclusion

To summarize the results, there was neither statistically significant main effect of the length of planning time nor interaction between planning time and proficiency on the holistic and analytic scores (RQ1) or on the quality of elicited performance in terms of the CAF variables and reproduction of IUs (RQ2). Although the participants' self-reported use of cognitive and metacognitive processes and strategies showed increased planning for language and mental rehearsal in the extended planning time in the higher proficiency group (RQ3), this finding did not translate to measurable outcomes in terms of scores or linguistic and content performance measures.

These results may be attributable to the fact that, first and foremost, a key to successful performance in the integrated task is understanding the listening input (Brown et al., 2005; Rukthong & Brunfaut, 2020). While the study by Choi and So (2018) suggested that candidates would only need to have the equivalent of a TOEFL iBT Listening score of 8 (IELTS Listening score of 5.5 according to ETS, 2010) to be able to understand the input, 13 of the participants in the lower group may have had listening proficiency just below this threshold (with an IELTS Listening score of 5.0[14]). The important role of understanding the listening input in completing this integrated speaking task can be gleaned from several observations. Firstly, at least two participants (who would have been included in the lower group) withdrew from the study due to failure in understanding the listening input. Secondly, there were participants even in the higher group who reported not understanding the input and perceived a negative effect on their performance (e.g., "I wrote down what the word sounded like … But I was like, oh it's not good. I didn't know what the lecture is about." (H735); "I felt very unsure about what to say because I felt like I had missed some parts [of the lecture]" (H728). Thirdly, and perhaps most importantly, the stimulated recall analysis identified specific issues in the listening phase that had a perceived influence on the subsequent planning and speaking phases of the task. For instance, there were reports from both proficiency groups on nonunderstanding of crucial terms in both tasks (such as *fungus*, *hollow*, and *reciprocity*) and related examples. There were also instances of less successful listening operations, such as missing incoming information through failing to recognize the lecture's structure (Excerpt 2) or focusing on taking notes about previous points (Excerpt 5). When it happened, participants either had to guess to try and make sense of the lecture based on partial understanding (cf. the reconstruction strategy in Rukthong & Brunfaut, 2020) or to avoid mentioning the unclear parts altogether when speaking. It follows that when crucial information is missed while listening or not recorded in sufficiently retrievable ways, there is not much a candidate can do to compensate for it in the planning time—extended or otherwise. There may be an interesting parallel to Ortega (2005) here—whereas the low complexity of the task content made pretask planning superfluous to some of her participants, the effect of nonunderstanding of the lecture input in this task may have been insurmountable through extended pretask planning. As H728 put it, "Even if I had more time, I don't know if I could have done something better because I actually missed some parts." With partial or inaccurate understanding of the lecture content, participants may have had to focus on the generation of content at the conceptualization stage (Kormos, 2006; Levelt, 1989), leaving little time for planning the language (formal) aspects of the speaking performance.

The second potential reason for little effect of extended planning time on score and performance is that higher proficiency participants might not have engaged extensively in planning for language and mental rehearsal—the two planning processes that would potentially lead to improved CAF of the elicited performance. Although participants in the higher proficiency group engaged more in planning for language (Statements 20–23 in the questionnaire) and mental rehearsal at the sentence level (Statement 25) in the extended planning time, they also did more planning for content, taking time to remember the lecture content (Statement 14) and to review and organize notes (Statements 17 and 18). Such planning for content included filtering out less relevant ideas in the notes (e.g., Excerpt 6 from the stimulated recall data) and deeper engagement with understanding the lecture content, as reflected in participants' reports (e.g., "I was writing things that had been said [while listening …] but now I was making sense of everything" (H725), and "I think with 90 seconds I felt pushed to do more, to think more, to be more critical" (H728). This echoes Ortega's (2005) findings, where advanced learners focused on organizing their thoughts and rehearsing for language during pretask planning, but they also displayed differential orientations toward communication of content versus accuracy. Thus, our higher proficiency participants may not have done as much planning for language as expected—at least not to a sufficient degree to lead to measurable improvements in CAF in their speaking performance.

The implication of these findings for the speaking section of the TOEFL iBT test is that the measurement of the lecture listening-into-speaking construct is stable regardless of the length of planning time (operational [20 s]) or extended [90 s]). One may argue that the cognitive validity of the test task could be enhanced with extended planning time through eliciting a wider variety of cognitive and metacognitive processes and strategies (RQ3), with closer approximation to the real-life processes when trying to summarize the content of a lecture. However, extended planning time did not lead to measurable changes in the elicited performance or scores—insofar as captured by performance variables used in this study and the adapted TOEFL iBT Speaking rubrics.

The current findings may invite questions around whether planning time in speaking tests is warranted at all or is rather superfluous. However, in addition to the closer similarity to real-life tasks (i.e., cognitive validity), planning time arguably has a role to play in the affective dimension of the test-taking experience. Indeed, in the study of planning time in the IELTS individual long-turn task, Wigglesworth and Elder (2010) concluded with a case for preserving some amount of planning time from this perspective, notwithstanding null results for improvements in the quality of performances. According to the authors, there is a consideration for fairness: the provision of planning time as a means to reduce candidates' anxiety in a highly stressful test context, allowing opportunities for candidates to give their best performance. There is also a face validity argument: 89% of 90 participants in Wigglesworth and Elder expressed a preference for planning time, and the authors argued that planning time may "engender greater confidence in speaking tests on the part of candidates and, accordingly, they may have greater acceptance of the scores they obtain" (p. 18). Indeed, a higher sense of confidence and preparedness was reported among some of our participants in both the lower and higher proficiency groups when given 90 s planning time. L501 reported using the extended planning time "to induct [induce] myself in the state to be more confident, like treat [trick] my brain," and H725 reported feeling more confident in summarizing the lecture: "this one was […] more difficult to obviously getting the points, but then expressing it was easier for me because I had prepared well for it than the other one." As such, there is an ethical argument for the inclusion of some amount of pretask planning time. Balancing this with practical considerations in operational testing, the current study supports the provision of the operational planning time (20 s) for the lecture listening-into-speaking tasks.

For the wider field of language testing research, this study has provided some insights into the qualitative differences in the test-taking processes and strategy use by learners at different levels of proficiency. A recurrent pattern in the stimulated recall data was that the participants in the higher proficiency group were more goal directed and selective in their listening and planning and exhibited flexibility and adaptability in adjusting their processes and strategies according to the task demands and the time available. Correspondingly, the participants in the lower proficiency group seemed to show less sense of direction in listening and planning. These observations echo the findings of relevant literature that learners of higher proficiency are able to apply selective attention when listening (Ortega, 2005) and quickly evaluate the importance of incoming information (Rukthong & Brunfaut, 2020). They are also in line with the argument in the learner strategies literature that there are important qualitative differences in how learners of higher and lower proficiency engage in the same processes and strategies (Chamot, 2001; Ortega, 2005; Rukthong & Brunfaut, 2020).

Finally, two limitations of this study must be acknowledged. One is the partial representation of the speaking performances by the variables of CAF and IUs. Although the variables were carefully selected based on the previous literature and their relevance to the contents of the TOEFL iBT integrated speaking rubrics, there are other features that could have been investigated (e.g., cohesion and coherence), which might reveal different results. Another limitation of this study lies in the methods in investigating learners' pretask planning and other test-taking processes and strategies. The absence of self-reported data (in the stimulated recall but perhaps also in the questionnaire) does not necessarily mean the absence of the relevant processes or strategies. An interesting extension to the present study could involve grouping the 16 stimulated recall participants according to their understanding of the lecture listening input (as evidenced in the verbal report as well as in the speaking performances) and conducting a follow-up analysis that cross-examine the participants on their (a) note-taking behavior and notes taken, (b) stimulated recall reports, and (c) speaking performance. Such an individual-based yet in-depth exploratory approach may allow us to gain more insights into the complex interaction between the test-taking processes and strategies in the integrated listening-into-speaking tasks and uncover successful or unsuccessful transfer (Ortega, 2005) of pretask planning to subsequent online speaking performance. Based on observations in this study about individual variations in understanding the input and engagement with planning time, future studies could also examine the role of test-taker characteristics (e.g., field of study, experience/familiarity with the TOEFL iBT Speaking test) on performance in the integrated listening-into-speaking tasks, exploring any potential effects on comprehension of the listening input and the use of planning time.

Notes

[1] Use of paraphrases is relevant to the rating criterion of the topic development rubric in the TOEFL iBT Integrated Speaking Rubrics (ETS, 2019), in which "repetition from the prompt" receives a Band 1 score, while more spontaneous choice of language is expected at higher score bands.

[2] A power analysis using G*Power 3.1 (Faul et al., 2007) revealed that a repeated measure two-way ANOVA design with a large effect (0.4) and a power of 0.90 would require a total of 68 participants.

[3] ETS, 2010.

[4] For descriptions of IELTS overall band scores, please refer to https://www.ielts.org/-/media/publications/guide-for-institutions/ielts-guide-for-institutions-uk.ashx?la=en

[5] One of the 10 pilot participants was recruited based on an overall score of 70/100 in the Cambridge English Certificate in Advanced English (CAE), approximating an IELTS score between 7.0 and 7.5 (https://www.cambridgeenglish.org/exams-and-tests/advanced/results/results-before-2015/ielts-score/). Her speaking was graded as exceptional.

[6] This software, developed and distributed for research purposes by ETS, creates a stand-alone application that administers TOEFL iBT test items, sections, and/or forms as required.

[7] Operational scoring of the speaking section is now completed by both a human rater and *SpeechRater®* (since August 2019). As the data collection for this study took place before August 2019, only the human rater scores were used.

[8] Frost et al. (2011) used the term "distorted" for inaccurate reproduction.

[9] The predictor variables (i.e., proportions of accurately reproduced IUs), might appear continuous but are bounded by a binary decision of 0 (inaccurate) and 1 (accurate).

[10] The grouping of participants into higher and lower levels was proven to be appropriate by two-way mixed ANOVA, which revealed a statistically significant main effect of proficiency levels on holistic scores ($F(1,69) = 51.58$, $p = .00$, $\eta p^2 = .43$) and analytic scores (Pronunciation: $F(1,69) = 49.98$, $p = .00$, $\eta p^2 = .42$; Pace: $F(1,69) = 43.07$, $p = .00$, $\eta p^2 = .39$; Language Use: $F(1,69) = 47.43$, $p = .00$, $\eta p^2 = .41$; Topic Development: $F(1,69) = 45.89$, $p = .00$, $\eta p^2 = .40$).

[11] The odds ratio was calculated as an exponential value of the estimate ($\beta = 1.60$).

[12] The odds ratio was calculated as an exponential value of the estimate ($\beta = -.71$).

[13] The odds ratio was calculated as an exponential value of the estimate ($\beta = -1.67$).

[14] Note that participants in this study have been sampled primarily on their IELTS Speaking score.

[15] For 2a to 2d, the exceptions are paraphrases that immediately follow the first mention (e.g., *fungus indirectly helps the tree, brings benefits to it* or *the trunk becomes empty inside, hollow*). In addition, for the Reciprocity task, certain coordinated clauses or verb phrases were not coded as separate IUs because they communicate a single idea about the nature of reciprocity, for example: *One gives something, and knows when to expect that something of similar value will be returned*.

References

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics*, *34*(5), 502–508. https://doi.org/10.1111/opo.12131

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

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

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-Academic-Purposes speaking tasks* (TOEFL Monograph Series MS-29). ETS. https://doi.org/10.1002/j.2333-8504.2005.tb01982.x

Chamot, A. U. (2001). The role of learning strategies in second language acquisition. In M. P. Breen (Ed.), *Learner contributions to language learning: New directions in research* (pp. 25–43). Longman.

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 319–352). Routledge.

Choi, I., & So, Y. (2018). A measurement model for listen-speak tasks. In G. J. Ockey & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp. 228–245). John Benjamins. https://doi.org/10.1075/lllt.50.15cho

Crookes, G. (1989). Planning and interlanguage variability. *Studies in Second Language Acquisition*, *11*(4), 367–383. https://doi.org/10.1017/S0272263100008391

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL®. *Language Testing*, *21*(2), 107–145. https://doi.org/10.1191/0265532204lt278oa

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. 219–238). John Benjamins. https://doi.org/10.1075/lllt.11.14eld

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

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

ETS. (2010). *Linking TOEFL iBT® scores to IELTS scores – A research report*. ETS. https://www.ets.org/s/toefl/pdf/linking_toefl_ibt_scores_to_ielts_scores.pdf

ETS. (2019). TOEFL iBT(c) Test integrated speaking rubrics. ETS. https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175 – 191. https://doi.org/10.3758/BF03193146

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (pp. 65 – 111). Cambridge ESOL/Cambridge University Press.

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

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

Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics*, *36*, 98 – 116. https://doi.org/10.1017/S0267190515000082

Frost, K., Clothier, J., Huisman, A., & Wigglesworth, G. (2020). Responding to a TOEFL iBT integrated speaking task: Mapping task demands and test takers' use of stimulus content. *Language Testing*, *37*(1), 133 – 155. https://doi.org/10.1177/0265532219860750

Frost, K., Elder, C., & Wigglesworth, G. (2011). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing*, *29*(3), 345 – 369. https://doi.org/10.1177/0265532211424479

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

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Laurence Erlbaum Associates.

George, D., & Mallery, M. (2010). *SPSS for windows step by step: A simple guide and reference* (10th ed.). Pearson.

Huang, H.-T. D. (2016). Exploring strategy use in L2 speaking assessment. *System*, *63*, 13 – 27. https://doi.org/10.1016/j.system.2016.08.009

Inoue, C. (2016). A comparative study of the variables used to measure syntactic complexity and accuracy in task-based research. *Language Learning Journal*, *44*(4), 487 – 505. https://doi.org/10.1080/09571736.2015.1130079

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*(1), 57 – 84. https://doi.org/10.1191/0265532202lt220oa

Kormos, J. (2006). *Speech production and second language acquisition*. Laurence Erlbaum.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*, 145 – 164. https://doi.org/10.1016/j.system.2004.01.001

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12 – 24. https://doi.org/10.1016/j.jslw.2016.10.003

Lam, D. M. K. (2015). Contriving authentic interaction: Task implementation and engagement in school-based speaking assessment in Hong Kong. In G. Yu & Y. Jin (Eds.), *Assessing Chinese learners of English: Language constructs, consequences and conundrums* (pp. 38 – 60). Palgrave Macmillan. https://doi.org/10.1057/9781137449788_3

Lee, Y.-W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, *23*(2), 131 – 166. https://doi.org/10.1191/0265532206lt325oa

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

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. https://doi.org/10.1002/tesq.159

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. https://doi.org/10.1017/S0272263198001041

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

O'Grady, S. (2018). *The impact of pre-task planning on speaking test performance for English-medium university study* (Doctoral dissertation). University of Bedfordshire Repository, Bedfordshire, UK. https://uobrep.openrepository.com/handle/10547/623295

Ortega, L. (1995). *The effects of planning in L2 Spanish oral narratives* (Research Note No. 15). Second Language Teaching and Curriculum Centre.

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

Ortega, L. (2005). What do learners plan? Learner-driven attention to form during pre-task planning. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 77 – 109). John Benjamins. https://doi.org/10.1075/lllt.11.07ort

Pang, F., & Skehan, P. (2014). Self-reported planning behaviour and L2 performance in narrative retelling. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 95 – 128). John Benjamins. https://doi.org/10.1075/tblt.5.04pan

R Core Team. (2020). R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rukthong, A. (2016). *Investigating the listening construct underlying listening-to-summarize tasks* (Doctoral dissertation). Lancaster University Research Directory. http://eprints.lancs.ac.uk/78054/1/2016ARukthongPhD

Rukthong, A., & Brunfaut, T. (2020). Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks. *Language Testing*, *37*(1), 31 – 53. https://doi.org/10.1177/0265532219871470

Skehan P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, *49*(1), 93 – 120. https://doi.org/10.1111/1467-9922.00071

Swain, M., Huang, L.-S., Barkaoui, K., Brooks, L., & Lapkin, S. (2009). *The speaking section of the TOEFL iBT (SSiBT): Test takers' reported strategic behaviors* (TOEFL Research Report RR-09-30). https://doi.org/10.1002/j.2333-8504.2009.tb02187.x

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

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. https://doi.org/10.1111/j.1467-9922.2008.00446.x

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. https://doi.org/10.1075/lllt.11.15tav

Weir, C. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan. https://doi.org/10.1057/9780230514577

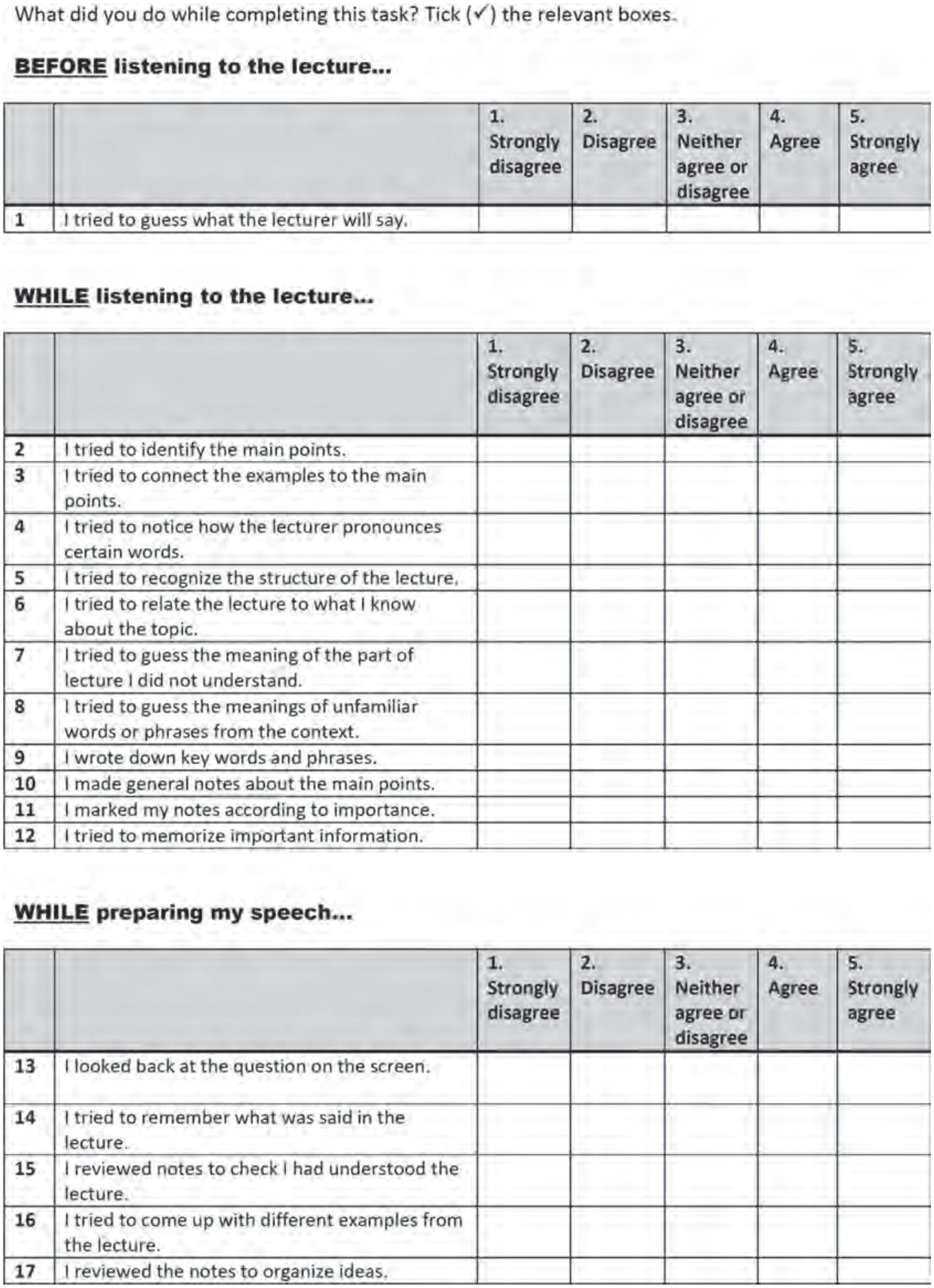
Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, *14*(1), 85 – 106. https://doi.org/10.1177/026553229701400105

Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly*, *7*, 1 – 24. https://doi.org/10.1080/15434300903031779

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. https://doi.org/10.1093/applin/24.1.1

Appendix A

Questionnaire on the Use of Cognitive and Metacognitive Processes and Strategies



What did you do while completing this task? Tick (✓) the relevant boxes.

BEFORE listening to the lecture...

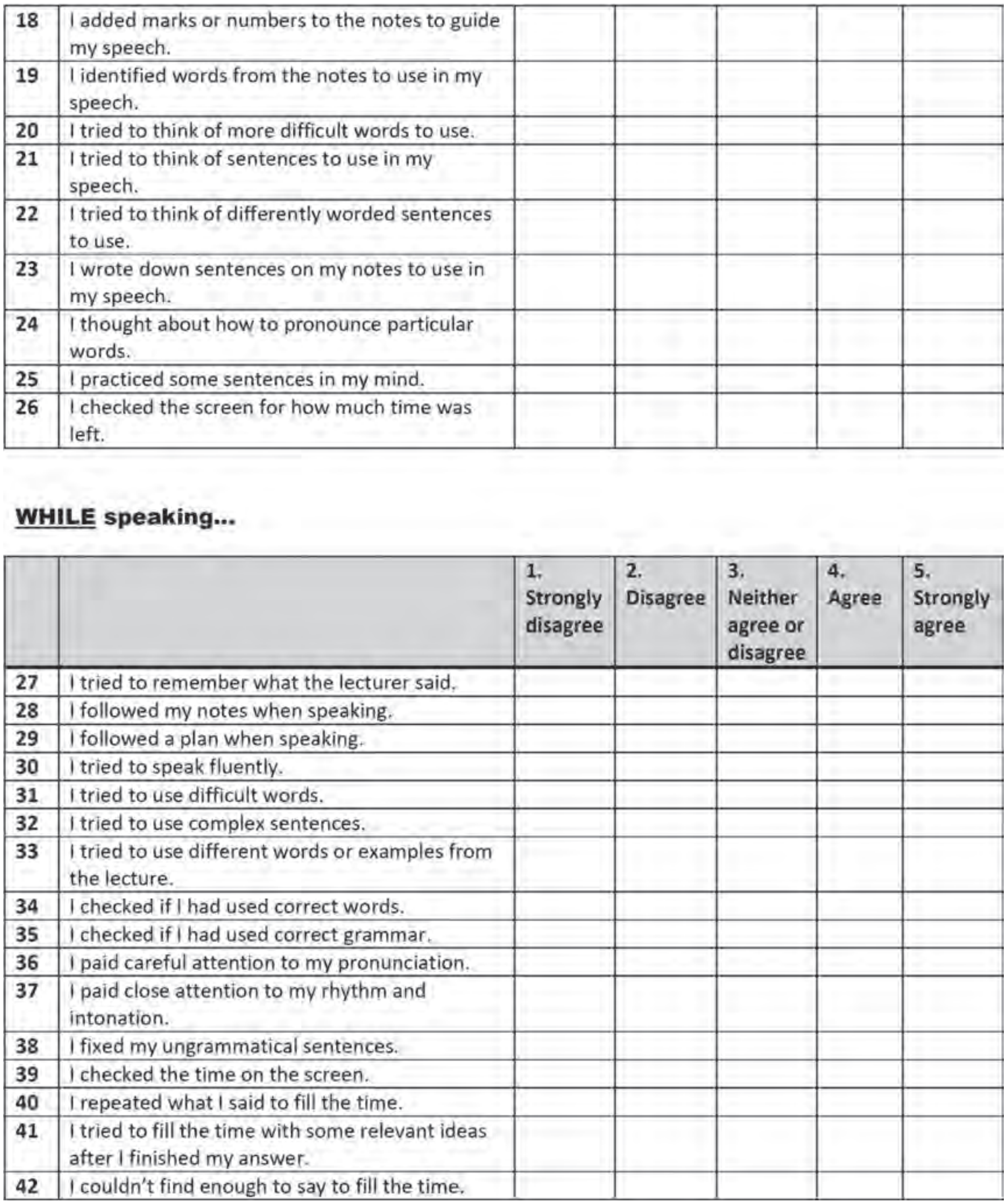
		1. Strongly disagree	2. Disagree	3. Neither agree or disagree	4. Agree	5. Strongly agree
1	I tried to guess what the lecturer will say.					

WHILE listening to the lecture...

		1. Strongly disagree	2. Disagree	3. Neither agree or disagree	4. Agree	5. Strongly agree
2	I tried to identify the main points.					
3	I tried to connect the examples to the main points.					
4	I tried to notice how the lecturer pronounces certain words.					
5	I tried to recognize the structure of the lecture.					
6	I tried to relate the lecture to what I know about the topic.					
7	I tried to guess the meaning of the part of lecture I did not understand.					
8	I tried to guess the meanings of unfamiliar words or phrases from the context.					
9	I wrote down key words and phrases.					
10	I made general notes about the main points.					
11	I marked my notes according to importance.					
12	I tried to memorize important information.					

WHILE preparing my speech...

		1. Strongly disagree	2. Disagree	3. Neither agree or disagree	4. Agree	5. Strongly agree
13	I looked back at the question on the screen.					
14	I tried to remember what was said in the lecture.					
15	I reviewed notes to check I had understood the lecture.					
16	I tried to come up with different examples from the lecture.					
17	I reviewed the notes to organize ideas.					



18	I added marks or numbers to the notes to guide my speech.					
19	I identified words from the notes to use in my speech.					
20	I tried to think of more difficult words to use.					
21	I tried to think of sentences to use in my speech.					
22	I tried to think of differently worded sentences to use.					
23	I wrote down sentences on my notes to use in my speech.					
24	I thought about how to pronounce particular words.					
25	I practiced some sentences in my mind.					
26	I checked the screen for how much time was left.					

WHILE speaking...

		1. Strongly disagree	2. Disagree	3. Neither agree or disagree	4. Agree	5. Strongly agree
27	I tried to remember what the lecturer said.					
28	I followed my notes when speaking.					
29	I followed a plan when speaking.					
30	I tried to speak fluently.					
31	I tried to use difficult words.					
32	I tried to use complex sentences.					
33	I tried to use different words or examples from the lecture.					
34	I checked if I had used correct words.					
35	I checked if I had used correct grammar.					
36	I paid careful attention to my pronunciation.					
37	I paid close attention to my rhythm and intonation.					
38	I fixed my ungrammatical sentences.					
39	I checked the time on the screen.					
40	I repeated what I said to fill the time.					
41	I tried to fill the time with some relevant ideas after I finished my answer.					
42	I couldn't find enough to say to fill the time.					

Appendix B

Adapted Integrated Speaking Rubrics

Score	General description	Pronunciation	Pace	Language use	Topic development
4	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. A response at this level is characterized by all of the following	Speech is generally clear. It may include minor lapses or minor difficulties with pronunciation or intonation. Overall intelligibility remains high	Speech is generally fluid and sustained. Pace may vary at times as the speaker attempts to recall information	The response demonstrates good control of basic and complex grammatical structures that allow for coherent, efficient (automatic) expression of relevant ideas. Contains generally effective word choice. Though some minor (or systematic) errors or imprecise use may be noticeable, they do not require listener effort (or obscure meaning)	The response presents a clear progression of ideas and conveys the relevant information required by the task. It includes appropriate detail, though it may have minor errors or minor omissions
3	The response addresses the task appropriately, but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas. A response at this level is characterized by at least two of the following	Speech is generally clear, but it exhibits minor difficulties with pronunciation or intonation and may require some listener effort at times. Overall intelligibility remains good, however	Speech generally shows some fluidity of expression, but it exhibits minor difficulties with pacing	The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. Such limitations do not seriously interfere with the communication of the message	The response is sustained and conveys relevant information required by the task. However, it exhibits some incompleteness, inaccuracy, lack of specificity with respect to content, or choppiness in the progression of ideas

Appendix B: Continued

Score	General description	Pronunciation	Pace	Language use	Topic development
2	The response is connected to the task, though it may be missing some relevant Information or contain inaccuracies. It contains some intelligible speech, but at times problems with intelligibility and/or overall coherence may obscure meaning. A response at this level is characterized by at least two of the following	Speech is clear at times, though it exhibits problems with pronunciation or intonation, and so may require significant listener effort. Problems with intelligibility may obscure meaning in places (but not throughout)	Speech exhibits problems with pacing. Speech may not be sustained at a consistent level throughout	The response is limited in the range and control of vocabulary and grammar demonstrated (some complex structures may be used, but typically contain errors). This results in limited or vague expression of relevant ideas and imprecise or inaccurate connections. Automaticity of expression may only be evident at the phrasal level	The response conveys some relevant information but is clearly incomplete or inaccurate. It is incomplete if it omits key ideas, makes vague reference to key ideas, or demonstrates limited development of important information. An inaccurate response demonstrates misunderstanding of key ideas from the stimulus. Typically, ideas expressed may not be well connected or cohesive so that familiarity with the stimulus is necessary to follow what is being discussed
1	The response is very limited in content or coherence or is only minimally connected to the task. Speech may be largely unintelligible. A response at this level is characterized by at least two of the following	Consistent pronunciation and intonation problems cause considerable listener effort and frequently obscure meaning	Delivery is choppy, fragmented, or telegraphic. Speech contains frequent pauses and hesitations	Range and control of grammar and vocabulary severely limit (or prevent) expression of ideas and connections among ideas. Some very low-level responses may rely on isolated words or short utterances to communicate ideas	The response fails to provide much relevant content. Ideas that are expressed are often inaccurate, limited to vague utterances, or repetitions (including repetition of prompt)
0	Speaker makes no attempt to respond OR response is unrelated to the topic				

Appendix C

Sample Transcripts With Scores

Scores are presented in the following order: holistic, pronunciation, pace, language use, topic development.

L520 (20 s) [2, 2, 1, 2, 2]

in this task uh we talk about uh it's in a botany class / uh the meaning is uh the tree- the tree fungus how can uh this fungus help uh indirectly grow up the tree / they uh eat the dea- uh dead uh wood and uhm help tree be- be very stable / uh in- special for younger tree they have a lot benefit / uh it's- and uh another uhm good things for grow this tree from the animals' products / they uhm have a lot of waste waste uh

L520 (90 s) [2, 2, 1, 1, 1]

uh in this uh mainly it's about reciprocity, general reciprocity / uh they suggest two groups uh give of uh *reciprocity / one is general reciprocity and second balanced reciprocity / uh general reciprocity is uhm like you do a good thing- uh things for uhm a person close to you like a brother somebody from family or a friends / uhm for example i- if you- my brother need uhm uh he buy a house and he need furniture like a bed uh I can help with some money and uh

H707 (20 s) [3, 3, 4, 3, 4]

uh the fungus will help the tree actually in two ways because uh there are many dead- uh dead woods in the tree trunk / so the fungus can help to eat it up and uh make the uh tree empty / the emptiness can help the tree to become much more lighter have a sta- much uhm stabler status especially during in a heavy wind compared to the young trees / and the second one is uh the fungu- uh the hollow tree also can provide home for uh- and shelter for animals / and animals will produce some waste / and this waste can help the trees to become more fertilized / and the- uh the trees can absorb the uh nutrite- nutritions from the uh animal waste uh to grow much more healthier /

H707 (90 s) [3, 4, 4, 3, 4]

the first type is generalized reciprocity / it means uh we do not- uh people do not get uh uh- people uh gave something to somebody without expectation of the immediate return / and it is usually for the uh socially close relationships for example the family members / for example, if your brothers move into a new house and want a new bed and you give it to him / so you do not need him to give it back immediately / but uh you may want him to help you when you are in the same situation in your- in the near future / so it shows a high level of trust / the balanced re- uh reciprocity uh is much more uh straight-forward / you expect immediate return or specification uh of the time in- of return in the near future / so uh the social distance is much more greater uh compare with the generalized reciprocity / for example the neighbor- uh you gave a neighbor a new bed, and- and you uh expect him to pay you uh the money or something in the equal value /.

Appendix D

Definitions and Criteria for Segmenting Idea Units

1 All clauses, including subordinate and relative clauses are separate IUs, except defining clauses (e.g., *This is one of the ways | that fungus benefits trees*) and complement clauses (e.g., *you know | that he'll help you someday*) coded as within the same IU as the main clause.

2 Sub-clause units are counted as separate IUs, according to the following parameters and exceptions[15]:

  a coordinated verb phrases are counted as separate IUs; but double verb constructions (e.g., you *expect the receiver to return a gift*) are not;

b phrases acting as discourse markers, typically set off from related clauses by commas, are considered to combine with related clauses as a single IU;
c coordinated nouns or noun phrases connected to a common verb phrase are counted as separate IUs; and.
d coordinated independent adjectives connected to a common verb phrase are counted as separate IUs.

3. Illustrating examples are separate IUs, even where included in a clause (e.g., *it works for family members, or close friends*).

Appendix E

Skewness of Differences Between Questionnaire Responses

Question	Low ($n = 35$)	High ($n = 35$)
Q1	−.22	−1.18
Q2	−1.41	1.31
Q3	−.68	−2.15
Q4	.81	.11
Q5	−.63	−1.22
Q6	1.10	−1.53
Q7	.10	−.15
Q8	.12	−.23
Q9	−1.32	.13
Q10	−1.15	−2.51
Q11	.10	−.23
Q12	−1.04	.00
Q13	−.94	.17
Q14	1.55	−.96
Q15	−.41	−.94
Q16	−.17	−.99
Q17	−1.14	−1.04
Q18	−1.32	.01
Q19	−.63	.36
Q20	−2.15	−.20
Q21	−.88	−.26
Q22	−.02	−.04
Q23	−.30	−.51
Q24	.30	−.21
Q25	.08	.46
Q26	.36	.33
Q27	.95	.23
Q28	−.59	−1.91
Q29	−.33	.83
Q30	.18	−2.74
Q31	.98	−.82
Q32	−.30	−.38
Q33	.56	.41
Q34	−.22	−.98
Q35	.40	−.79
Q36	−1.24	−.26
Q37	.59	−.32
Q38	−.51	.07
Q39	.68	.90
Q40	.42	.15
Q41	−.33	.15
Q42	−.05	−.28

Appendix F

Descriptive Statistics for Questionnaire Responses

		Low ($n = 35$)								High ($n = 35$)							
		20 s				90 s				20 s				90 s			
No.	Statement	*M*	*SD*	Median	Range	*M*	*SD*	Median	Range	*M*	*SD*	Median	Range	*M*	*SD*	Median	Range
1	I tried to guess what the lecturer will say	3.49	1.20	4	4	3.74	1.11	4	4	3.49	1.25	4	4	3.63	1.31	4	4
2	I tried to identify the main points	4.26	.85	4	3	4.11	.68	4	3	4.34	.84	4	4	4.34	.84	4	4
3	I tried to connect the examples to the main points	4.20	.72	4	2	4.09	.45	4	2	4.40	.69	4	3	4.49	.51	4	1
4	I tried to notice how the lecturer pronounces certain words	3.51	1.09	4	4	3.49	1.22	4	4	3.66	1.24	4	4	3.26	1.17	3	4
5	I tried to recognize the structure of the lecture	3.89	.93	4	4	3.82	.72	4	3	4.20	.76	4	3	4.34	.73	4	3
6	I tried to relate the lecture to what I know about the topic	3.74	1.01	4	4	3.60	1.12	4	4	3.49	1.15	4	4	3.71	1.25	4	4
7	I tried to guess the meaning of the part of lecture I did not understand	3.83	.95	4	3	3.69	1.11	4	4	3.54	1.09	4	4	3.69	1.02	4	3
8	I tried to guess the meanings of unfamiliar words or phrases from the context	3.60	1.12	4	4	3.66	.91	4	3	3.49	1.12	4	4	3.62	.95	4	3
9	I wrote down key words and phrases	4.09	1.07	4	4	4.14	.69	4	3	4.66	.48	5	1	4.63	.49	5	1
10	I made general notes about the main points	4.03	1.18	4	4	3.97	.87	4	3	4.62	.55	5	2	4.71	.46	5	1
11	I marked my notes according to importance	3.40	1.33	4	4	3.54	1.24	4	4	3.80	1.23	4	4	4.00	1.28	5	4
12	I tried to memorize important information	4.14	.81	4	4	4.06	.84	4	3	4.14	.91	4	4	4.14	1.06	4	4
13	I looked back at the question on the screen	3.00	1.28	3	4	3.51	1.22	4	4	3.26	1.34	4	4	3.40	1.38	4	4
14	I tried to remember what was said in the lecture	3.91	.89	4	4	3.69	.96	4	4	3.63	1.00	4	3	3.94	.94	4	4

Appendix F: Continued

		Low ($n = 35$)								High ($n = 35$)							
		20 s				90 s				20 s				90 s			
No.	Statement	*M*	*SD*	Median	Range	*M*	*SD*	Median	Range	*M*	*SD*	Median	Range	*M*	*SD*	Median	Range
15	I reviewed notes to check I had understood the lecture	3.89	1.02	4	4	3.97	.82	4	3	4.31	.63	4	3	4.43	.50	4	1
16	I tried to come up with different examples from the lecture	2.74	1.12	3	4	3.00	1.16	3	4	2.63	1.35	2	4	3.00	1.48	3	4
17	I reviewed the notes to organize ideas	3.66	1.14	4	4	4.06	.73	4	3	4.03	1.01	4	3	4.51	.56	5	2
18	I added marks or numbers to the notes to guide my speech	3.29	1.41	4	4	3.80	1.02	4	3	3.23	1.42	4	4	4.06	1.16	4	4
19	I identified words from the notes to use in my speech	3.66	1.06	4	4	3.94	.78	4	3	4.26	.74	4	3	4.26	.92	4	3
20	I tried to think of more difficult words to use	2.50	1.13	2	4	2.54	1.09	3	4	2.14	.94	2	3	2.86	1.40	3	4
21	I tried to think of sentences to use in my speech	3.37	1.21	4	4	3.54	1.04	4	4	2.97	1.18	3	4	3.91	1.01	4	4
22	I tried to think of differently worded sentences to use	2.91	1.22	3	4	3.23	.97	3	4	2.77	1.19	2	4	3.43	1.31	4	4
23	I wrote down sentences on my notes to use in my speech	2.71	1.05	3	4	3.09	1.15	3	4	2.49	1.15	2	4	3.43	1.36	4	4
24	I thought about how to pronounce particular words	2.89	1.18	3	4	3.49	1.09	4	4	3.03	1.44	3	4	3.11	1.32	3	4
25	I practiced some sentences in my mind	3.51	1.17	4	3	3.71	1.13	4	4	3.00	1.24	3	4	3.57	1.27	4	4
26	I checked the screen for how much time was left	3.57	1.31	4	4	3.26	1.38	4	4	3.34	1.49	4	4	3.49	1.48	4	4
27	I tried to remember what the lecturer said	3.83	1.01	4	4	3.86	.94	4	3	3.63	1.17	4	4	3.74	1.15	4	4
28	I followed my notes when speaking	3.97	.98	4	4	4.12	.77	4	3	4.46	.66	5	2	4.60	.50	5	1
29	I followed a plan when speaking	3.60	1.17	4	4	3.83	1.12	4	4	3.89	1.05	4	4	4.20	.76	4	3
30	I tried to speak fluently	3.86	1.00	4	3	3.74	.89	4	3	4.06	.91	4	4	4.34	.54	4	2
31	I tried to use difficult words	2.54	1.15	2	4	2.49	1.12	2	4	2.26	1.02	2	4	2.89	1.30	3	4
32	I tried to use complex sentences	2.54	1.12	2	4	2.51	1.15	3	3	2.43	1.07	2	4	2.83	1.29	3	4

Appendix F: Continued

		Low (*n* = 35)								High (*n* = 35)							
		20 s				90 s				20 s				90 s			
No.	Statement	*M*	*SD*	Median	Range	*M*	*SD*	Median	Range	*M*	*SD*	Median	Range	*M*	*SD*	Median	Range
33	I tried to use different words or examples from the lecture	2.94	1.30	3	4	2.85	1.23	3	4	2.86	1.35	3	4	3.14	1.14	3	4
34	I checked if I had used correct words	3.09	1.17	3	4	2.97	1.12	3	4	3.46	1.12	4	4	3.71	.89	4	3
35	I checked if I had used correct grammar	2.83	1.15	3	4	2.80	.99	3	4	3.23	1.21	3	4	3.51	1.07	4	3
36	I paid careful attention to my pronunciation	2.86	1.24	3	4	3.17	1.18	3	4	3.06	1.16	3	4	3.29	1.05	3	4
37	I paid close attention to my rhythm and intonation	2.86	1.00	3	4	2.83	.98	3	4	2.80	1.11	3	4	2.89	1.11	3	4
38	I fixed my ungrammatical sentences	2.77	1.03	3	4	2.77	.94	3	3	3.00	1.14	3	4	3.20	1.02	3	3
39	I checked the time on the screen	3.54	1.20	4	4	3.17	1.34	4	4	3.71	1.41	4	4	3.57	1.40	4	4
40	I repeated what I said to fill the time	2.40	1.14	2	4	2.49	1.01	2	4	2.14	.94	2	3	1.91	.85	2	3
41	I tried to fill the time with some relevant ideas after I finished my answer	2.40	1.14	2	4	2.51	1.12	2	4	2.31	1.21	2	4	2.18	1.14	2	4
42	I could not find enough to say to fill the time	2.91	1.42	3	4	2.89	1.30	3	4	1.91	.98	2	4	2.20	1.26	2	4

[a] *Note*. 1 = strongly disagree; 2 = disagree; 3 = neither agree nor disagree, 4 = agree; 5 = strongly agree.

Suggested citation:

Inoue, C., & Lam, D. M. K.. (2021). *The effects of extended planning time on candidates' performance, processes, and strategy use in the lecture listening-into-speaking tasks of the TOEFL iBT® test* (TOEFL Research Report No. RR-93). ETS. https://doi.org/10.1002/ets2.12322

Action Editor: John Norris

Reviewers: This report was reviewed by the Research Subcommittee of the TOEFL Committee of Examiners.