Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks

Anchana Rukthong
Prince of Songkla University, Thailand

Tineke Brunfaut
Lancaster University, UK

Abstract
Integrated test tasks, such as listening-to-speak or reading-to-write, are increasingly used in second language assessment despite relatively limited empirical insights into what they assess. Most research on integrated tasks has primarily focused on the productive skills involved; studies exploring the receptive skills mostly investigated tasks with reading input. Little is known about the nature of listening comprehension in integrated listening-to-write or listening-to-speak tasks. This study therefore investigates the listening construct underlying integrated tasks with oral input and its effect on summary accuracy. Eight listening-to-summarize tasks (four listening-to-speak, four listening-to-write) were administered to 72 Thai-L1, English-L2 students. Sixty participants provided their views on sources of listening difficulty through post-task questionnaires. Twelve participants produced stimulated recalls on their listening comprehension processing. The analyses of the recalls, combined with participants’ listening notes and oral/written summaries, revealed participants’ use of several cognitive listening processes and their monitoring through (meta)cognitive strategies, functioning interactively and interdependently in complex ways. The use of listening processes and strategies varied between tasks with different listening inputs, partly owing to differences in the passages’ linguistic difficulty (as perceived by the participants). However, the successful application of these processes and strategies (and their combinations) proved to be a prerequisite for producing accurate summaries.

Keywords
Construct validity, integrated tasks, language processing, second language listening, testing listening

Corresponding author:
Anchana Rukthong, Department of Languages and Linguistics, Faculty of Liberal Arts, Prince of Songkla University, Hat Yai, Songkhla 90112, Thailand.
Email: anchana.r@psu.ac.th
Integrated test tasks, which require test-takers to employ at least two language skills (receptive and productive) for task completion, are increasingly adopted in second language (L2) assessment. For example, the TOEFL iBT contains integrated reading-listening-speaking and reading-listening-writing tasks, and the Cambridge Assessment English’s C1 Advanced and C2 Proficiency tests include reading-to-write tasks. It has been suggested that integrated tasks have a number of advantages, including the following: being more authentic than item types such as multiple-choice and matching items, by assessing abilities corresponding to those performed outside testing situations (Asención, 2004; Cumming, Grant, Mulcahy-Ernt, & Powers, 2004); reducing the impact of background knowledge, as test-takers can (partly) generate the content of their response from the input (Cumming et al., 2004; but see Huang, Hung & Plakans, 2018 for conflicting findings); and increasing positive washback in classroom settings (Weigle, 2004).

At the same time, there is discussion concerning which abilities are really being tested by integrated tasks and which ones contribute either to success or failure in integrated task performances; previous studies have led to conflicting views on these issues. Gebril (2010) and Lee (2006), for example, found that performance results on integrated tasks (e.g., reading-to-write tasks) correlate highly with those on independent tasks (e.g., writing tasks), and thus they suggested that both task types assess a similar construct. As such, integrated task formats have been included in tests in order to measure productive skills (writing and speaking). However, researchers who examined test-takers’ processing behaviours (e.g., Asención, 2004; Plakans, 2008), concluded that integrated and independent tasks differ in terms of what they assess. These scholars argued that test-takers need to have a certain amount of text comprehension ability in order to understand the written or oral source text in order to complete successfully the integrated task (Brown, Iwashita, & McNamara, 2005; Cumming, Kantor, Baba, Eouanzoui, Erdosy, & James, 2006). They concluded that comprehension ability plays an essential role in task performance: The extent of understanding the input leads to differences between performances on integrated tasks versus independent tasks (see Brown et al., 2005; Cumming et al., 2006), and, in addition, it differentiates between performance levels on integrated tasks themselves (see Cumming et al., 2006; Gebril & Plakans, 2013).

Clarity on the construct, however, is vital from the perspective of test validation. If the construct underlying tests is not defined clearly, test designers will struggle to provide explicit and defensible links between test performances and interpretations of test scores (what test-takers can do), and thus will struggle to justify decisions made on the basis of the test scores (Bachman & Palmer, 2010; Messick, 1995). This highlights the need to gain more empirical insights into what it is that integrated tasks measure (Frost, Elder, & Wigglesworth, 2011). Overall, however, the number of studies on integrated tasks is relatively limited, and several are primarily focused on the productive skills involved (see, e.g., Brown et al., 2005; Cumming et al., 2006; Lee, 2006; Plakans, Gebril, & Bilki, 2016; Sawaki, Stricker, & Oranjie, 2009). Studies that explored the receptive skills in integrated testing were mostly concerned with tasks containing written input and thus with the role of reading in such tasks (see, e.g., Asención Delaney, 2008; Gebril & Plakans, 2013; Plakans, 2008). To date, few studies have focused on integrated tasks with oral input texts. Consequently, the role of listening comprehension in integrated tasks, such as listening-to-speak or listening-to-write tasks, is still
largely unexamined and thus unclear. As a first step towards addressing the above research gaps, this article sets out to explore in specific the nature of listening in integrated test tasks with oral input and integrated listening’s relationship with successful productive output in terms of content accuracy.

**Processing of oral texts**

A fundamental source of data for understanding the construct underlying tests is the processes, strategies, and knowledge employed by test-takers to complete these tests (Messick, 1989; Weir, 2005). From a psycholinguistic perspective, text comprehension is the result of cognitive and metacognitive processes which work in parallel and interactive ways (Anderson, 1985; Færch & Kasper, 1986; Garrod, 1986). The processing of oral texts, in particular, is considered to be conditioned by two main factors. The first factor is the knowledge that listeners possess, both linguistic and non-linguistic (Field, 2013; Rost, 2016). Linguistic knowledge concerns language-related information stored in the listener's memory and used during comprehension processing to create and interpret discourse in language use. It comprises phonological, lexical, syntactic, semantic, and pragmatic knowledge (Bachman & Palmer, 1996; Field, 2013). In L2 listeners, such knowledge is often limited, depending on their level of L2 acquisition. Non-linguistic knowledge refers to the world/cultural knowledge that helps language users with interpreting the meaning of words and sentences (Anderson, 1985). Listeners employ this type of knowledge to assist their comprehension processing (Field, 2013). In particular, this may play a role in L2 contexts in which people’s cultural background and experience may diverge from that of the target language context.

The second factor that is thought to contribute to success in listening is the automaticity of language processing. The L2 listening literature offers several listening comprehension processing models, with the following being particularly influential: Rost (2016), Field (2013), and Vandergrift and Goh (2012). All three models recognize that listening comprehension involves cognitive processing which occurs in an interactive rather than a linear manner, and which relies not only on linguistic knowledge but also on topical and world knowledge. Different terms are used in these models, however, to describe the cognitive listening processes. Rost (2016) characterized listening processing through four categories, according to the knowledge used, that is, neurological, linguistic, semantic, and pragmatic processing. Based on the functions of the processes, Field (2013) classified cognitive listening processes into two main categories, namely, lower-level and higher-level processes. Nevertheless, Field’s (2013) and Rost’s (2016) cognitive processing categories overlap. In particular, Field’s lower-level processes, which are activated to understand the literal meaning of a text, correspond with the neurological and linguistic processing of Rost’s model. The higher-level processes, which according to Field are activated to understand a text’s discourse and implied meaning, are in line with Rost’s semantic and pragmatic processing. In addition, the cognitive listening processes in these two models also seem congruent with those in Vandergrift and Goh’s (2012) cognitive model of L2 listening comprehension, which in turn was adapted from Anderson’s (1985) three-stage model of cognitive processing for language comprehension (conceptual processing, parsing, and utilization). Based on this literature, the
processing of listening texts in our study is characterized as involving the following cognitive processes (which interact in complex manners):

**Lower-level processing:**

1. **Acoustic-phonetic decoding:** accessing acoustic sounds, registering the sounds, and converting these into the representations of the phonological system of the relevant language.
2. **Word recognition:** identifying words or phrases in a speech stream.
3. **Parsing:** mapping the recognized words onto the syntactic or semantic structures of the language, or segmenting chunks of information.

**Higher-level processing:**

1. **Semantic processing:** combining the textual information and interpreting it with reference to one’s world knowledge, to make the processed information meaningful. Kintsch and van Dijk (1978) explain that semantic processing occurs at two different levels, that is, the local and global levels. At the local level, listeners connect individual propositions to each other to understand their semantic relationship and establish the text’s idea units. Semantic processing at the global level involves combining different idea units and relating them to the theme of the text in order to understand its main point.
2. **Pragmatic processing:** relying on texts’ linguistic information and communicative contexts to identify speakers’ intentions, since the ‘true’ meaning of a text is often implied rather than explicitly stated. Listeners thereby draw on their social and cultural knowledge to complement the available linguistic information.

The automaticity with which listeners are able to employ the above processes is thought to impact on success in their listening comprehension.

When listening takes place in an L2 context, however, research has shown that limitations in linguistic and non-linguistic knowledge mean that L2 listeners also often use strategies to facilitate their comprehension processing (see, e.g., O’Malley, Chamot, & Kupper, 1989; Rubin, 1981; Vandergrift & Goh, 2012). In contrast to the highly automated nature of cognitive processes, strategies are used to facilitate the listening process and fill gaps in understanding, and they are employed with a certain degree of consciousness (Vandergrift & Goh, 2012). Two types of strategies used during text processing are proposed in the literature: cognitive and metacognitive.

**Cognitive strategies** are used when listeners encounter comprehension problems. The following strategies have been put forward in the listening literature (Goh, 2002; O’Malley et al., 1989; Vandergrift, 2003; Vandergrift & Goh, 2012):

1. **Inferencing:** relying on information from the text or immediate context to identify the meaning of unknown language items, fill gaps in one’s listening, or create links between pieces of information to enable a more coherent interpretation of the text.
2. **Elaboration**: using knowledge from outside the text or from the broader context to interpret the text’s meaning.
3. **Prediction**: anticipating upcoming information.
4. **Translation**: finding equivalent L1 words, phrases or sentences before interpretation.
5. **Fixation**: pausing to try to understand a particular textual fragment.
6. **Reconstruction**: relying on words/phrases decoded from the listening text to recreate parts of the text missed while listening, in order to make comprehension complete.

**Metacognitive strategies** are employed to manage and supervise processes for listening (including cognitive processing and cognitive strategy use). The use of these metacognitive strategies enables optimal functioning of cognitive processes and cognitive strategies. Metacognitive strategies, as put forward by O’Malley et al. (1989), Goh (2002), Vandergrift (2003), and Vandergrift and Goh (2012), comprise the following:

1. **Preparing for listening**: planning or preparing for and analysing the task instructions.
2. **Selective attention**: focusing selectively on information the listener anticipated to hear.
3. **Directed attention**: consciously steering one’s attention back to the incoming text when one has lost focus.
4. **Monitoring comprehension**: checking and confirming the accuracy of one’s textual interpretations.
5. **Real-time assessment of input**: foregrounding and backgrounding information depending on the task’s requirements.
6. **Comprehension evaluation**: evaluating whether one has understood the listening input well enough to complete the task.

**Listeners’ perceptions**

Apart from gaining insights into the assessed construct through test-takers’ process and strategy usage, scholars such as Messick (1989) and Weir (2005) argued that data on the interaction between test-takers and the context of performance also help to shed light on test constructs. Information on the latter can be gathered from test-takers, since their physical, psychological, and experiential characteristics may directly affect the way in which they complete test tasks (Weir, 2005; O’Sullivan & Weir, 2011). Previous research, for example, found a relationship between test-takers’ views on the difficulty of listening input and how difficult the tasks actually were (based on performance). For instance, for different types of listening test tasks, Révész and Brunfaut (2013) and Brunfaut and Révész (2015) found that listening was effectively difficult for test-takers when they perceived the texts to contain unclear pronunciation, difficult lexis, complex grammatical structures, and/or unclear organization. To our knowledge, however, more direct links between listeners’ perceptions, strategy use and task success have not been explored. In the context of vocabulary testing, however, Ertürk and Mumford (2017) found that the task which was conditioned to be
difficult by the complexity of its linguistic characteristics and the abstractness of its ideas was not as difficult as expected. They explained that this was because when the test-takers perceived a task as difficult, they adjusted their test-taking strategies. Additionally, because of the adjustments, the test-takers tended to be more successful in their task completion.

Given the above-mentioned research gap on links between perceptions, strategy use and performances in L2 listening, it is meaningful to explore whether Ertürk and Mumford (2017)’s findings can be replicated for (integrated) listening assessments.

Therefore, to provide a more comprehensive understanding of the nature of listening in performing integrated tasks with oral input, we also investigated test-takers’ perceptions of listening difficulty. Following Skehan’s (1998) conceptualization of task difficulty, we define listening difficulty as the linguistic complexity\(^1\) and, more specifically, the syntactic and lexical difficulty and range, density of the language and redundancy in the listening input. This selection of listening text characteristics (for the exploration of test-takers’ perceptions of listening difficulty) is motivated by earlier findings on their association with listening difficulty (for a review, see Bloomfield et al., 2011).

**Research questions**

As shown in the literature review, insights into how test-takers process the oral input in integrated listening tasks, and thus the construct underlying such tasks, are relatively limited. Therefore, we aimed to answer the following research questions:

- **RQ1.** What forms of cognitive processing and of strategy use do English-L2 test-takers activate to comprehend the listening input integrated in listening-to-summarize tasks?
- **RQ2.** How do test-takers’ cognitive processing and their strategy use relate to summary quality?
- **RQ3.** What are English-L2 test-takers’ perceptions of the linguistic complexity of the listening input in listening-to-summarize tasks?
- **RQ4.** How do test-takers’ perceptions relate to their cognitive processing and strategy use, and to their summary content performances?

**Methodology**

**Overall research design and rationale**

To answer the research questions, we designed a study in which English-L2 speakers completed a set of listening-to-summarize test tasks that required them to listen to an oral input text and then provide an oral summary for some of the tasks (listening-to-speak) and a written summary for other tasks (listening-to-write). While listening, they were allowed to take notes, and immediately after completing each task, they either did a stimulated recall to provide insight into processing and strategy use or completed a perception questionnaire to provide further insight into sources of processing difficulty.
Verbal reports have long been argued to provide meaningful data on cognitive processes and strategies (Ericsson & Simon, 1993), including in language testing (e.g., Green, 1998) and L2 listening (e.g., Vandergrift, 2003). In this study, we adopted the stimulated recall method because verbalization after task completion means that it does not interfere with hearing the listening input or with natural task performance. In addition, the use of stimuli such as replays of listening input, test papers, and notes can help to jog the memory of participants. Several studies have successfully employed this method to investigate constructs, processes, and strategies in testing listening and integrated test tasks, for example, Barkaoui et al. (2013), Field (2012), Goh (2002), Holzknecht et al. (2017), Révész and Brunfaut (2013), and Yi’an (1998). Barkaoui et al. (2013) warned, however, that protocol data do not usually reveal the effectiveness of processing and strategy use, and thus advised to analyse performance output in addition. Therefore, in this research, we also studied test-takers’ notes and oral/written summaries.

To explore perceptions of processing difficulty sources, we opted for questionnaires, as this method has been shown to work well for investigating perceptions of factors that impact on task difficulty (Tavakoli, 2009). Examples of studies that have used questionnaires for these purposes in the context of testing L2 listening include Révész and Brunfaut (2013), and Brunfaut and Révész (2015). For a more detailed rationale for the study’s methodology, see Rukthong (2016).

Participants

Seventy-two Thai-L1, English L2-speakers from four UK universities participated in the study. All participants completed a set of listening-to-summarize tasks. With 12 of these students, we also conducted stimulated recalls. They were recruited through purposive sampling, to ensure that they had different levels of English proficiency to be able to explore potential differences in processing and strategy use depending on performance scores. Specifically, six participants whose English test scores were at or near the minimum entry requirement for their program of study (IELTS overall band score of 6.0–6.5) and six with higher scores (IELTS overall band score of 7.0–8.5) were recruited. Although these self-reported IELTS scores constituted only a rough indication of participants’ English proficiency, it was hoped that they would ensure at least some variation in abilities within the participant group. In addition, recruiting only Thai-L1 speakers allowed for conducting the recalls in participants’ L1, as the researcher gathering the recall data was also a Thai-L1 speaker. These 12 participants were distributed equally between male and female genders and were 23–40 years old ($M = 29$, $SD = 5.56$).

The 60 other participants completed the perception questionnaires. Sixty-two percent of these participants were female and 38% were male. Their ages ranged between 20 and 40 years old ($M = 27$, $SD = 4.29$). They reported IELTS overall scores ranging between 4.5 and 8.0, and with a range of 4.5–8.0 in listening, 4.5–7.5 in writing, and 5.5–8.5 in speaking.

Instruments

Listening-to-summarize tasks. For the purposes of the study, Pearson provided us with four listening-to-summarize tasks from the Pearson Test of English Academic (PTE
These were selected from the PTE Academic database for having similar statistical difficulty coefficients, as based on large-scale test administration. Given that we had four tasks, there were four different listening inputs, which we refer to in this article as Listening A, Listening B, Listening C, and Listening D. These inputs (audio-only) consisted of short lecture excerpts (1:00–1:29 mins) on different topics: Listening A was about corruption, B about talent, C about the functions that Vitamin D performs, and D discussed the life stories of a successful individual. In addition, these texts presented information in different ways: Listening A featured a description and comparison of two charts, B was a lecture comparing various points of views, C a critique of various false beliefs, and D a series of narratives leading to a final overall point.2

The output modalities of these PTE Academic listening-to-summarize tasks differed, that is, two of the four tasks provided by Pearson were listening-to-write tasks and two were listening-to-speak tasks. For the purposes of our study, we created a second, parallel version for each task in the alternative output mode (speak or write). This was done so that we had both a listening-to-speak and a listening-to-write version for each of the four listening inputs. Thus, in total, we had eight listening-to-summarize tasks, of which four were listening-to-speak and four listening-to-write tasks.3 Each participant, however, completed only two listening-to-speak and two listening-to-write tasks; in this manner, the same participant did not listen to the same input twice.

Prior to data collection, participants were familiarized with the task types (and stimulated recall methodology, where relevant) by means of two sample tasks. For all tasks, and adhering to the regular PTE Academic procedures, the participants listened to the input once and were allowed to take notes on paper while listening.4 Then, in the listening-to-speak tasks, they were given 10 seconds to prepare their oral response and then 40 seconds to summarize orally the listening. In the listening-to-write tasks, the participants had 10 minutes to write a 50–70 word summary of the listening passage. To control for order effects, a counterbalanced design was used for task administration: half of the participants started with two listening-to-write tasks and half with two listening-to-speak tasks. In addition, different sequences of listening input texts were randomly assigned.

Stimulated recalls. For those participants taking part in the stimulated recalls, a video-recording was made while they were completing the listening-to-summarize tasks. This video was then used as a stimulus in the recall interview, together with the notes the participant had taken and the summary they had produced in response to the task. In order to maximize the collection of data about the participant’s processing and strategy use, the videos were shown twice. The first time the participant watched the recording, he or she was in charge and could pause the recording and say whatever he or she wanted at any moment. The second time, the researcher stopped the recording at specific points in the passage (the same for all participants), probing the participant to report his or her thought processes from when he or she had been completing the task. Although the participants could choose either Thai or English, all verbalized their thoughts in Thai.

Perception questionnaire. The questionnaire consisted of five Likert-scale statements for which participants indicated their level of agreement on a five-point scale (5 – strongly
agree, 4 – agree, 3 – neutral, 2 – disagree, and 1 – strongly disagree). Participants completed this set of statements after doing each of the listening-to-summarize tasks. Based on Skehan (1998), the statements focused on five linguistic characteristics of listening texts shown to contribute to task difficulty: (1) lexical complexity, (2) syntactic complexity, (3) information density, (4) information redundancy, and (5) discourse complexity. We administered the questionnaire in Thai, but English translations of the statements can be found in Table 1.

**Data analysis**

We analysed various sources of data to gain insight into the cognitive processing and strategy use on which the participants relied in order to understand the listening input of the listening-to-summarize tasks (RQ1), and the association between their processing/strategy use and the production of successful summary content (RQ2). This comprised the following: (a) the stimulated recall data; (b) participants’ notes made while listening; (c) their task responses (the oral and written summaries they had produced as part of the task completion requirements); and (d) their scores on the criterion content for the oral/ written output they had produced. To analyse these data (a–c), we adopted a coding scheme based on the literature on listening comprehension processing. That is, we coded the data for cognitive processes, cognitive strategies, metacognitive strategies, and their respective sub-processes/strategies (see the list of processes and strategies in the section “Processing of oral texts” above). As listening is an interactive process, we note that multiple codes were often allocated to individual recall utterances/notes/summary statements.

**Analysis of stimulated recalls.** After having transcribed the stimulated recalls, we analysed the data in NVivo. Following Gass and Mackey (2000), we chunked the data into units that plausibly corresponded to processes and strategies. Next, we assigned process and strategy codes to these units, using the coding framework, and translated extracts quoted in this article from Thai into English.

Although the stimulated recalls were expected to reveal important information on participants’ processing, we anticipated that automated processes would not be reported as much. We predicted that this would particularly be the case for lower-level processes. However, the participants’ handwritten notes and their oral and written summaries could potentially give direct evidence of lower-level processing. For example, written-out words and sentences could indicate that a participant had phonologically and lexically processed (parts of) the listening input. Therefore, in order both to supplement and triangulate the stimulated recall data, we also analysed any notes the participants had taken while listening, together with their summary contents.

**Analysis of participants’ notes and summary content.** For the analyses of the notes, we used the same coding scheme as for the stimulated recalls, to make inferences about participants’ use of cognitive processes and of strategies. Then, we used the stimulated recall data and participants’ summaries to verify and confirm any such inferences. For example, Figure 1 shows notes from Participant 1 (P1), which indicate at least two cognitive
processes (i.e., acoustic-phonetic decoding and word recognition), as he had written down a number of words from the passage. In addition, the visual organization with lines and arrows suggests that P1 had also parsed and semantically processed the information, since he pieced different bits of information together. We thus coded the notes with these processes. Next, we verified our interpretations with P1’s recall data. Indeed, P1 reported that he had drawn lines to link groups of words together, and explained how exactly the connections should be understood. Furthermore, P1’s written summary content reflected those lines and contained a successfully connected content summary of the listening passage. P1’s employment of semantic processes, as we hypothesized from his notes, was therefore confirmed.

**Coding reliability.** To warrant the reliability of the coding of the recalls and the notes, an external coder (a Thai-L1 university lecturer with a Master’s degree in TESOL and expertise in verbal data analysis) double-coded 25% of the entire dataset. The inter-coder agreement, established using Cohen’s Kappa, was .87 for the cognitive processes, .86 for the cognitive strategies, and .85 for the metacognitive strategies. Overall, therefore, there was strong inter-coder agreement (Green, 2013). Where the double codings differed, the coders discussed these until they reached agreement. The remainder of the coded dataset was then revisited in light of the inter-coder discussion and revised where relevant.

**Performance scores.** Participants’ performances on the listening-to-summarize tasks (i.e., their written or oral summaries) were assessed by two regular PTE Academic markers using the rating scales for human marking of the listening-to-summarize tasks. The oral summaries were marked for content, pronunciation, and fluency, and the written summaries for content, vocabulary, and grammar. For the purpose of this article, only the content scores have been considered as they can be regarded as indicators of test-takers’ input comprehension ability in integrated test tasks.

Finally, to establish participants’ perceptions of the linguistic complexity of the listening input in listening-to-summarize tasks (RQ3), we analysed the perception questionnaire responses. Namely, we calculated descriptive statistics (frequencies) on the Likert-scale data to establish participants’ level of agreement with the listening linguistic complexity statements. We also conducted an analysis of variance (ANOVA) to investigate whether
participants’ perceptions of linguistic complexity were significantly different across the four listening inputs. In addition, correlational analyses (Spearman’s rho) were run to explore the relationship between perceptions of task difficulty and task performance, and qualitative analyses were conducted to explore patterns between the use of particular listening processes/strategies and test-takers’ perceptions and performances (RQ4).

**Results**

The analyses of participants’ stimulated recalls, handwritten notes, and the content of their oral/written summaries revealed that the participants relied on both cognitive processing and strategy use to comprehend the listening passages of the listening-to-summarize tasks. However, differences in the accuracy of the content of the oral/written summaries were associated with some differences in ways of processing and strategy use. In addition, the questionnaires showed that participants perceived there to be differences in the linguistic complexity of the four different listening inputs, and this was further reflected in their perceptions of differences in the difficulty of the listening-to-summarize tasks as a whole. Below, we report the findings per research question (for additional examples, see Rukthong, 2016).

**RQ1: What forms of cognitive processing and of strategy use do English-L2 test-takers activate to comprehend the listening input integrated in listening-to-summarize tasks?**

The stimulated recall data indicated that in order to complete the tasks, the participants appeared to activate both cognitive processes and cognitive and metacognitive strategies, some of which were task-specific.
Cognitive processes

Both lower- and higher-level processes were used by the participants while listening. Acoustic-phonetic processing and word recognition (the two lowest-level processes) were engaged in across the four listening topics (see Figure 2). All participants used parsing processes while they listened to two of the passages (Listening A and Listening C). The number of participants who engaged in higher-level processes differed somewhat between the listening passages. Overall, however, higher-level processes appeared to be relatively less used than lower-level ones.

Examples from the stimulated recalls of cognitive processing by the participants are as follows:

Extract 1
I heard “he is the wonderful example” and “he can overcome all kinds of obstacles” (P11: Acoustic-phonetic processing, Word recognition and Parsing)

Extract 2
I know that the main focus was about the main function of Vitamin D that helps maintain blood calcium level. It’s not about maintaining bones and teeth as generally understood. (P6: Semantic processing (global level))

Cognitive strategies

Five cognitive strategies were found to be used by the participants in order to increase their comprehension of the passages (see Figure 3). Typically, at the beginning of the listening task participants predicted or anticipated what would come up next in the listening passage. This was followed by the use of three other cognitive strategies: (a) fixation, (b) inferencing, and/or (c) elaboration. Near the end of the listening, some participants

![Figure 3. Cognitive strategies used to solve listening problems.](image-url)
employed reconstruction strategies. Of these five cognitive strategies, inferencing and prediction appeared to be more popular than the others. All 12 participants appeared to make inferences while listening to all four listening passages. Prediction was used by seven participants in Listening A, B, and C (but by different individuals), and by five participants in Listening D. The strategies of elaboration and reconstruction were applied by different numbers of participants, depending on the listening texts. None of the participants reported using reconstruction strategies for Listening B and C. Fixation strategies were used by the smallest number of participants, especially in Listening D.

Examples from the stimulated recalls of cognitive strategy use are as follows:

*Extract 3*

I stopped to think about the word I had just heard. It’s something like bri . . . power. I was not sure. I thought it was brain power. (P4: Fixation)

*Extract 4*

I heard “tetany.” I didn’t know this word, but from this context I can say that it’s a disease that occurs when we don’t have enough Vitamin D. (P1: Inferencing)

**Metacognitive strategies**

The data analyses revealed that the participants applied six different metacognitive strategies to oversee and manage their listening processing (see Figure 4). At the initial stage (i.e. after the participant was shown the task and before the listening file began to play), some participants prepared for listening, for example to determine the task requirements, reduce their anxiety, anticipate key words, and rehearse the sounds of potential key words. Once the audio had started and participants were listening, four metacognitive strategies were activated: selective attention, directed attention, comprehension

![Figure 4. Metacognitive strategies used for monitoring listening comprehension.](image-url)
monitoring, and real-time assessment of input. When the passage was close to the end, some participants reported employing the strategy *comprehension evaluation*.

The use of metacognitive strategies varied between participants and listening topics. Overall, comprehension monitoring was used by the highest number of participants, followed by selective attention and directed attention. Less than half of the participants presented evidence of engaging in the strategies *preparing for listening*, *real-time assessment of input*, or *comprehension evaluation*.

Examples from the stimulated recalls of metacognitive strategy use are as follows:

**Extract 5**

I focused on numbers. I was thinking that I should have heard more numbers because it was a graph which was usually described by numbers. (P7: Selective attention)

**Extract 6**

I was following the listening text, trying to create its outline. I linked it back to the graph and thought whether it described the graph . . . . . . I checked if I understood the graph correctly. (P3: Real-time assessment of input)

**RQ2: How do test-takers’ cognitive processing and their strategy use relate to summary quality?**

Although quite similar types of cognitive processing and of strategic use emerged for the tasks with different listening input, we observed several differences between participants depending on the scores achieved for the content of their productive output (oral/written summaries). We defined high scorers as those participants who received 2 out of 2 marks on the content criterion, average scorers obtained 1.5–1.0, and low scorers 0.5–0.

First, as would be expected, we found that the cognitive processes used by the high scorers were more effective than those by lower scorers. The high scorers engaged in semantic processing at the global level and in pragmatic processing, which enabled them to generate successfully the main point(s) of the text based on the textual information that they processed at lower levels. Second, the high and average scorers tended to use more types of metacognitive strategies, with a high frequency of comprehension monitoring. Of all the processes and strategies they activated, the high scorers reported using metacognitive strategies in 45% of the cases, the average scorers in 40%, and the low scorers in 27%. The low scorers, however, applied fewer types of metacognitive strategies and relied heavily on the directed attention strategy (66%). Third, the summaries scoring low on content were produced by test-takers who were not as successful in using cognitive strategies, in particular inferencing and elaboration, as compared to the strategy use of the high scorers. In *Listening A, B*, and *D*, the proportion of inferencing used by the high scorers was about twice as high as that used by the low scorers. In addition, as evidenced in their summary contents, their inferred information was more accurate than that of the low scorers. On the other hand, the low scorers, as they reported in their stimulated recalls, had relied mainly on fixation (stopping to figure out the meaning of unknown words). Lastly, the high-scoring test-takers did not report employing reconstruction strategies (relying on
words/phrases obtained to recreate the listening’s main points), whereas the low-scoring test-takers did. The latter used such strategies to more or less fabricate listening ideas in order to fulfill the task requirements.

**RQ3: What are English-L2 test-takers’ perceptions of the linguistic complexity of the listening input in listening-to-summarize tasks?**

The questionnaire responses, presented in Table 1, indicate that the test-takers perceived the four listening inputs as differing in difficulty, although the passages had been selected according to the same specifications and the tasks had been chosen for having similar difficulty coefficients from operational testing. Overall, tasks with *Listening D* were perceived as having the most complex input ($M = 3.26$ out of a maximum of 5), followed by tasks with *Listening A* ($M = 3.05$), *Listening B* ($M = 2.79$) and *Listening C* ($M = 2.74$), respectively. *Listening B* and *C*, which were perceived to be the easiest, were thought to be less complex in terms of their lexis, syntax, ideas, and organization and to have more redundancy. Although *Listening A* was rated as being difficult in terms of the quantity of ideas, it received comparatively low scores with regard to lexical and syntactical difficulty as well as redundancy. In contrast, *Listening D* was rated difficult across the board, and in particular with regard to lexical complexity and its perceived lack of redundancy as compared to the other listening inputs. The differences between

| Text characteristics                                  | Listening A | Listening B | Listening C | Listening D |
|-------------------------------------------------------|-------------|-------------|-------------|-------------|
| Vocabulary in the listening passage was difficult for me. | 3.05 (.93)  | 2.65 (.95)  | 2.97 (1.01) | 3.43 (.91)  |
| Sentence structures in the listening passage were complicated for me. | 3.05 (.95)  | 2.80 (.88)  | 2.58 (.89)  | 3.18 (.91)  |
| There were many important ideas to be processed during the listening passage. | 3.87 (.83)  | 3.55 (.75)  | 3.33 (1.02) | 3.85 (.88)  |
| Important ideas in the listening passage were paraphrased or repeated more than once.* | 2.72 (1.01) | 2.58 (.85)  | 2.60 (.92)  | 3.20 (.86)  |
| Ideas in the listening passage were clearly connected.* | 2.57 (.77)  | 2.38 (.64)  | 2.20 (.63)  | 2.63 (.80)  |
| Overall perceived complexity. | 3.05 (.90)  | 2.79 (.81)  | 2.74 (.89)  | 3.26 (.87)  |
| Mean summary content score (out of 2) | 1.16 1.00 | 1.45 0.58 |  |

*Since these were positive statements, the values presented have been reversed so that higher figures signify a lower level of agreement and thus a higher level of difficulty.
participants’ overall perceptions of the four passages’ complexity were statistically significant, as established through ANOVA ($F = 13.822, p \leq .01$).

*RQ4: How do test-takers’ perceptions relate to their cognitive processing and strategy use, and to their summary content performances?*

Some tentative patterns could be observed when considering the participants’ perceptions in relation to their performances on the different tasks. In particular, tasks featuring *Listening B* and *C* were perceived to have a similar, medium-level difficulty of input, with an average overall perceived complexity of 2.79 and 2.74 (out of 5), respectively. This appeared to align with participants’ performance on these two tasks since test-takers’ mean content scores were at an average level (1.00 and 1.45, respectively, out of 2). The tasks with *Listening A* were perceived to have slightly more difficult input than those with *Listening B* and *C*, while the mean content score was still at an average level (1.16 out of 2). Tasks with *Listening D*, however, were consistently perceived to have more difficult listening input than the others ($M = 3.26$) and at the same time participants’ average summary content score on the *Listening D* tasks was also low ($M = .58$ out of 2). The associations between summary content scores and overall linguistic complexity perceptions (as analysed with Spearman’s rho) did not reach statistical significance at the $p < .05$ level.

When examining the use of cognitive processes and of strategies, as compared to participants’ content scores and perceptions, some patterns emerged. To process the input texts which were perceived to be more lexically and syntactically difficult, namely *Listening A* and *D*, the participants appeared to rely more on local-level processing (word recognition and parsing) to decode the text. In those cases where the test-takers were not able to recognize the words/phrases, their stimulated recall indicated that they used fixation and directed attention strategies. To try to arrive at the main point of the text, the participants furthermore applied inferencing and elaboration strategies, as evidenced in their recalls. They inferred on the basis of the words and phrases decoded and the information that they had predicted to hear in the incoming input. During the comprehension processing, these processes and strategies were found to function interactively and be closely monitored by the comprehension monitoring strategy. In tasks featuring *Listening A*, the participants managed to establish more accurate information in order to summarize the main point of the text (as evidenced in their recalls and summary contents) with the use of these processes and strategies, compared to *Listening D*. Hence, the low content scores for *Listening D* indicate that, for this particular input, such processes and strategies were not adequate.

Although the questionnaire results point to several possible areas of difficulty in *Listening B* and *C*, the participants tended to go beyond word/phrase recognition, and semantically processed the texts to obtain their idea units, as observed in the recalls and summary contents. Based on these idea units, test-takers were then able to infer and elaborate on the main point of the text. Greater familiarity with the lexis and syntactic structures of the texts, as indicated by the questionnaire responses, enabled participants to make more accurate predictions of what the listening was about than for *Listening A* and *D*, which had less familiar lexis and syntax. Also, in contrast to *Listening A* and *D*, no evidence of the use of reconstruction was found for *Listening B* and *C*, and less
elaboration usage. This is possibly because of the higher levels of comprehension that the participants had while listening to $B$ and $C$.

**Discussion: Listening processing, input complexity, and performance on listening-to-summarize tasks**

Overall, the results indicate that in order to understand listening input and produce accurate oral/written summaries of its main point(s), L2 test-takers must rely on both lower- and higher-level text processing and on strategy use. Our finding that high-scoring participants relied on more types of processes and strategies and with higher degrees of success than did the lower-scoring ones is, in fact, not a surprise, as the L2 listening literature suggests that higher-level cognitive processes have to be employed if listeners are to understand the main point of the text to which they listen (Field, 2013; Rost, 2016). However, as has also been found in the general L2 listening literature (e.g., Field, 2013; Rost, 2016), it is unlikely that test-takers would be successful at activating cognitive processes, especially at the higher level, without effectively engaging in cognitive processing at the lower level (acoustic-phonetic processing, word recognition, and parsing), which intuitively would be expected, and also not without effectively engaging in strategy use. For example, it was observed in this study that the test-takers had to activate comprehension processing (a metacognitive strategy) to help them decide what other strategies (e.g., fixation, inferencing, elaboration, and directed attention) needed to be used to assist their cognitive processing and whole-text processing. The fact that the success of several strategies reported by the participants seemed to rely on the concurrent use of other strategies and processes provides empirical evidence for the complex interaction of cognitive and strategic processing in listening as theorized in the literature (see, e.g., Buck, 2001; Goh, 2002; O’Malley et al., 1989; Rost, 2016; Vandergrift & Goh, 2012). It might also explain why previous studies such as Barkaoui et al. (2013) and Purpura (1999) found no or a weak significant relationship between the use of strategies and task performance. In particular, because they counted the use of individual strategies, the interactive nature of process and strategy use might have been underestimated in these studies. They focused primarily on developing listening comprehension taxonomies rather than on describing the interactive nature of listening processing. In addition, it should be kept in mind that test-takers’ reporting of the use of certain processes or strategies does not necessarily mean they used these successfully. As evidenced in this study, the participants reported activating inferencing to establish the main point of the listening input. The analysis of some test-takers’ summary content, however, showed that the main point they identified was inaccurate because they had made inferences on the basis of words or propositions they had decoded from the listening which were irrelevant to the main point. From a methodological perspective, it is thus also vital to crosscheck the use of processes and strategies (as reported, e.g., through verbal protocols) with the effectiveness of use (as demonstrated, e.g., in written/spoken discourse performances).

These findings may also not be in line with Field’s (2013) model of cognitive processing for listening, which is widely referenced in L2 listening research, as this study also specifically explored strategic processing (cognitive and metacognitive strategies) and found that these strategies play a significant role in task performance. Although Field’s model includes some strategies such as monitoring and inferencing as part of cognitive
processing, it does not integrate other strategies which were found to assist comprehension processing in this study. This may suggest that successful listening, which in this study means that the listener can correctly summarize the main points of the oral text, is not only associated with inferencing and comprehension monitoring but also with real-time assessment of input, prediction, and directed attention strategy uses. One factor that might contribute to this difference between the studies could be the source of research data. Although Field (2013) developed his model based on L1- and expert L2-listeners’ text comprehension processes, this study relied on the processing data of L2 listeners who may have more limited language knowledge. This study, nevertheless, provides empirical support for Rost (2016) and Bachman and Palmer (2010), who recognize strategic processing as part of the construct underlying L2 abilities.

Another finding of our study is that test-takers scored differently when different listening inputs were given; high scorers on one task were in some cases low scorers on another. This variation in the summary content scores within individuals suggests that input text characteristics may play a crucial role in performance on listening-to-summarize tasks. To gain a clearer understanding of why most participants performed differently on tasks with different listening inputs, we also explored task difficulty via test-takers’ perceptions.

Although the listening inputs used in this study were selected by the item writers to fall within the same range of difficulty level, and the test developer shared tasks of similar statistical difficulty with the researchers, our findings show that the test-takers perceived the texts’ linguistic information as differing in difficulty. The perceived difficulty appeared to be linked to the activation of specific strategies on which participants relied to succeed in the tasks. That is, in the task where the test-takers thought that the listening input was lexically and structurally complex, successful test-takers tended to rely heavily on lower-level processing, fixation, and inferencing of the text’s main points. In contrast, the texts that were perceived as less lexically complex, but containing an unclear organization, unclear importance markers, and unpredictable content for the test-takers were likely to trigger more interactive and interdependent uses of cognitive processes and (meta)cognitive strategies. This suggests that although tasks are constructed to target a specific set of abilities, an additional ability that plays a role in task performance (and, by extension, affects performance scores) is adaptability, as success seems to depend not only on test-takers’ language skills in the narrow sense, but also on their ability to adapt skills to the purpose of the task. This supports Iwashita, McNamara, and Elder’s (2001) and Ertürk and Mumford’s (2017) conclusion that task difficulty cannot be predicted without reservation unless interactions between test-takers and the task requirements/purposes are taken into account. The study also seems to lend empirical support to the view that the interaction between test-takers and the test situation or test context is an essential source of data in test validation (see, e.g., Bachman & Palmer, 2010; Weir, 2005).

Conclusion

Key insights

Our investigation of listening-to-summarize test tasks shows that a wide range of cognitive processes and (meta)cognitive strategies are employed by test-takers to process the
listening input. Our findings also suggest that listeners of different performance levels employ these processes and strategies in partly different ways, and that these processing differences are associated with different levels of task completion success as demonstrated in the written/oral summaries. Relying on these findings, our study, in accordance with Gebril and Plakans (2013) and Sawaki, Quinlan, and Lee (2013), shows that the abilities to comprehend (listening) texts play a crucial role in this type of integrated task and, importantly, impact on the quality of the productive output, at least on its content. Consequently, we argue that, together with the productive skills involved in and represented in most rating approaches to integrated test tasks, listening comprehension abilities should also be described as a part of the construct or abilities assessed by integrated test tasks containing a listening component, in particular listening-to-summarize tasks, and should be firmly represented in any rating approach used for this task type.

The listening-to-summarize tasks were found to activate and require higher-level listening processes, which is the aim in many test contexts (Taylor & Geranpayeh, 2011). Although lower-level processes proved to be crucial for oral text comprehension, the sole activation of these processes was found to be insufficient if test-takers are to summarize accurately the listening input in their written/spoken performance: what the listeners can establish through such processes may be the text’s literal meaning, but not necessarily the main point. What is essential is the successful activation of high-level processes (semantic processing at the global (discourse) level and pragmatic processing) and the ability to activate useful strategies (cognitive and metacognitive) at the right time.

Although integrated-listening tasks have the potential to tap into higher-level task processing, it is not easy to predict what specific abilities the tasks assess, because during task completion, test-takers will adjust their processes and strategies to best suit the task, resulting in different degrees of success in task completion. The ability to adapt to task requirements and purposes may thus need to be recognized more explicitly as part of the construct of listening-to-summarize tasks and as part of the scoring of these tasks, and further research on these issues seems warranted.

Limitations

Although we designed this study carefully, it has some limitations. The findings reported in this article were generated from participants with the same L1 background, and from a limited number of people in the study’s qualitative component as a result of its labour intensiveness. This may affect the generalizability of the study. In addition, the study investigated listening-to-summarize tasks, which involve the production of oral and written summaries of the listening input. Integrated listening test tasks requiring other forms of output (e.g., discussing or retelling the listening input) may require different (proportions of) processing. Another limitation concerns the fact that the study focused primarily on the processes used to understand the listening input and their effect on the productive output in terms of content; the study did not investigate in detail the speaking and writing processing involved in task production or the effect on other aspects of the productive output such as its linguistic or speech complexity. We therefore recommend further research that combines the processing of receptive and productive skills, in order
to obtain a comprehensive picture of all abilities measured by integrated test tasks. Finally, although we gained rich insights into test-takers’ process and strategy use by means of verbal protocols (complemented with the analyses of notes and performance content), there is always the possibility of incomplete capturing because of the lack of reporting or awareness by the test-takers.

Despite these limitations, this study has shed light on the largely under-researched role of receptive skills (in this case, listening) in integrated forms of L2 testing, and evidenced the extensive and intricate employment of listening processes and strategies as part of the task completion process.

Acknowledgements

We wish to thank Pearson for their support with materials from the PTE Academic, and we are grateful to the students who participated in this study for their time and effort.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was part-funded by a Prince of Songkla University PhD scholarship.

ORCID iD

Tineke Brunfaut https://orcid.org/0000-0001-8018-8004

Notes

1. Note that Skehan (1998) used the term code complexity to refer to linguistic complexity.
2. The PTE Academic tasks also contained a still image of the topic or a graph. However, these simply seemed to fulfil the role of broad topic activation; reference to these visuals was not necessary in order to complete the task successfully.
3. Note that the investigation of the effect of summary mode (oral/written) formed part of the larger study on integrated listening tasks of which this study is a part, but that those analyses are not the focus of this article.
4. Although the participants could choose the language in which they took notes, they all did so in English.
5. There is also an automated scoring system.
6. Bachman and Palmer (2010) and Weir (2005) defined the terms ‘test situation’ and ‘test context’ as relating to the interactions between test-takers and task requirements/purposes/demands.

References

Anderson, J. R. (1985). Cognitive psychology and its implications. New York: Worth.
Asención, Y. (2004). Validation of reading-to-write assessment tasks performed by second language learners (Unpublished doctoral dissertation). Northern Arizona University, USA.
Asención Delaney, Y. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes, 7*(3), 140–150. https://doi.org/10.1016/j.jeap.2008.04.001

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

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

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

Bloomfield, A., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2011). *What makes listening difficult? Factors affecting second language listening comprehension*. University of Maryland Center for Advanced Study of Language.

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 MS29). Princeton, NJ: Educational Testing Service.

Brunfaut, T., & Révész, A. (2015). The role of task and listener characteristics in second language listening. *TESOL Quarterly, 49*(1), 141–168. https://doi.org/10.1002/tesq.168

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.

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

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL (TOEFL Monograph Series MS30). Princeton, NJ: Educational Testing Service.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Ertürk, N., & Mumford, S. E. (2017). Understanding test-takers’ perceptions of difficulty in EAP vocabulary tests: The role of experiential factors. *Language Testing, 34*(3), 413–433. https://doi.org/10.1177/0265532216673399

Færch, C., & Kasper, G. (1986). The role of comprehension in second language learning. *Applied Linguistics, 7*(3), 257–274. https://doi.org/10.1093/applin/7.3.257

Field, J. (2012). The cognitive validity of the lecture-based question in the IELTS Listening paper. In L. Taylor & C. Weir (Eds.), *IELTS Collected Paper 2: Research in reading and listening assessment* (pp. 391–453). Cambridge: Cambridge University Press.

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 77–151). Cambridge: Cambridge University Press.

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), 1–25. https://doi.org/10.1177/0265532211424479

Garrod, S. (1986). Language comprehension in context: a psychological perspective. *Applied Linguistics, 7*(3), 226–238. https://doi.org/10.1093/applin/7.3.226

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum.

Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing, 15*(2), 100–117. https://doi.org/10.1016/j.asw.2010.05.002
Gebril, A., & Plakans, L. (2013). Toward a transparent construct of reading-to-write tasks: The interface between discourse features and proficiency. *Language Assessment Quarterly, 10*(1), 9–27. https://doi.org/10.1080/15434303.2011.642040

Goh, C. C. M. (2002). Exploring listening comprehension tactics and their interaction patterns. *System, 30*, 185–206. https://doi.org/10.1016/S0346–251X(02)00004–0

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge: Cambridge University Press.

Green, R. (2013). *Statistical analyses for language testers*. Basingstoke, UK: Palgrave Macmillan.

Holzknecht, F., Eberharter, K., Kremmel, B., Zehentner, M., McCray, G., Konrad, E., & Spöttl, C. (2017). Looking into listening: Using eye-tracking to establish the cognitive validity of the Aptis listening test. ARAGs Research Reports Online AR-G/2017/3. London: British Council.

Huang, H. D., Hung, S. A., & Plakans, L. (2018). Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks. *Language Testing, 35*(1), 27–49. https://doi.org/10.1177/0265532216677106

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

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363–394.

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

Messick, S. (1989). Meaning and value in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11. https://doi.org/10.3102/0013189X018002005

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1002/j.2333–8504.1994.tb01618.x

O’Malley, J. M., Chamot, A. U., & Kupper, L. (1989). Listening comprehension strategies in second language acquisition. *Applied Linguistics, 10*(4), 418–437. https://doi.org/10.1093/aplin/10.4.418

O’Sullivan, B., & Weir, C. J. (2011). Test development and validation. In B. O’Sullivan (Ed.), *Language testing: Theories and practices* (pp. 13–32). London: Palgrave Macmillan.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing, 13*(2), 111–129. https://doi.org/10.1016/j.asw.2008.07.001

Plakans, L., Gebril, A., & Bilki, Z. (2016). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing, 36*(2), 161–179. https://doi.org/10.1177/0265532216669537

Purpura, J. E. (1999). *Learner strategy use and performance on language tests: A structural equation modeling approach*. Cambridge: Cambridge University Press.

Révész, A., & Brunflaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition 35*(1), 31–65. https://doi.org/10.1017/S0272263112000678

Rost, M. (2016). *Teaching and researching listening* (3rd ed.). New York and Oxon: Routledge.

Rubin, J. (1981). Study of cognitive processes in second language learning. *Applied Linguistics, 11*(2), 117–131. https://doi.org/10.1093/aplin/II.2.117

Rukthong, A. (2016). *Investigating the listening construct underlying listening-to-summarize tasks* (Unpublished doctoral dissertation). Lancaster University, UK.
Sawaki, Y., Quinlan, T., & Lee, Y.-W. (2013). Understanding learner strengths and weaknesses: Assessing performance on an integrated writing task. *Language Assessment Quarterly, 10*(1), 73–95. https://doi.org/10.1080/15434303.2011.633305

Sawaki, Y., Stricker, L. J., & Oranjie, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing, 26*(1), 5–30. https://doi.org/10.1177/0265532208097335

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Tavakoli, P. (2009). Investigating task difficulty: Learners’ and teachers’ perceptions. *International Journal of Applied Linguistics, 19*(1), 1–25. https://doi.org/10.1111/j.1473–4192.2009.00216.x

Taylor, L., & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalising the test construct. *Journal of English for Academic Purposes, 10*(2), 89–101. https://doi.org/10.1016/j.jeap.2011.03.002

Vandergrift, L. (2003). Orchestrating strategy use: Toward a models of the skilled second language listener. *Language Learning, 53*(3), 463–496. https://doi.org/10.1111/1467–9922.00232

Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening*. New York: Routledge.

Weigle, S. C. (2004). Integrated reading and writing in a competency test for non-native speakers of English. *Assessment Writing, 9*, 27–55. https://doi.org/10.1016/j.asw.2004.01.002

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke, UK: Palgrave Macmillan.

Yi’an, W. (1998). What do tests of listening comprehension test? A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing, 15*(1), 21–44. https://doi.org/10.1177/026553229801500102