

L2 listening comprehension

Theory and research

Elvis Wagner

Temple University

This chapter examines how L2 listening ability has been modeled and operationalized in the research literature, and provides a critical overview of the dominant models. It also describes how researchers have used both taxonomies of listening skills as well as data-driven approaches to creating models of listening ability. The chapter then provides a critical discussion of how the constructs of L2 listening ability have been operationalized and measured by empirical researchers. The chapter concludes with an analysis of why models and operationalizations of L2 listening ability have often neglected to include or focus on those aspects of language that are unique to listening ability.

1. **Background: A conceptual introduction to the key constructs and models of L2 listening**

Second language (L2) listening ability has been the focus of a robust amount of research in the last thirty years. This chapter will examine and critique how the key constructs of L2 listening ability have been modeled and operationalized in the research literature.

When trying to formulate a model of L2 listening ability, most researchers have based their model at least in part on models of L1 listening ability. This certainly makes sense, as the processes are the same physiologically, and are very similar in many other ways as well. Therefore, before describing the different models of L2 listening ability, it is necessary to briefly describe some models of L1 listening ability that have been influential for L2 models.

Anderson's cognitive information processing model (1985, 1990, 1995) has been hugely influential for L2 listening researchers, and continues to be the cognitive model used as the basis for most models of L2 listening ability. The model has evolved over time, but his 1995 model conceptualizes human information processing as a series of internal states that are constantly being revised and updated

as new information (the input) enters and is acted upon. In the model, there are three distinct yet interdependent stages: the selective perception stage, the parsing stage, and the utilization stage. In the selective perception stage, the input (in this case, spoken input) is registered in sensory memory, with the listener attending to at least some of the input, and this attended-to input (sometimes referred to as intake) is transferred to working memory. In the parsing stage, the attended-to input stored in working memory is acted upon, integrating this new information with previously known information, and moved to long-term memory. In the third stage, the information is retrieved from long-term memory and is used for a particular purpose. These three stages are partially ordered in the manner in which the input is transmitted and received, yet each stage overlaps with the other stages and influences the processing of the information at each stage. Anderson's model has evolved over time, but this conceptualization seems to be the dominant one for most models of L2 listening ability.

Another important and relevant model of memory and cognition that has been applied to (some) models of L2 listening ability is Paivio's dual coding theory (1971, 1986, 1991). In this theory, verbal systems represent the properties of language, and nonverbal systems represent the nonlinguistic world. Both of these classes exist together in different modalities. Visual linguistic input is prototypically printed words, while visual nonlinguistic input would include visual objects; auditory linguistic input is prototypically spoken words, while auditory nonlinguistic input would include environmental sounds. An important component of the theory is the idea of functional independence of subsystems. That is, the verbal and nonverbal systems are obviously interrelated, yet still independent, as evidenced by the fact that one system can be active while the other is inactive, or they can be active in parallel (Paivio, 1991). In a listening context at a coffee shop, for example, a listener might be simultaneously processing aural linguistic information (e.g., spoken words) spoken by a conversation partner, as well as aural non-linguistic information (e.g., background coffee shop noises). That listener might also be processing visual linguistic information (e.g., signs in the coffee shop), and visual non-linguistic information (e.g., the gestures and appearance of the speaker, objects and people in the coffee shop).

1.1 Taxonomies and data-driven examinations of L2 listening ability

Instead of creating formal models of L2 listening ability, a number of researchers elected to create taxonomies of the skills or sub-skills that are involved in the L2 listening process (e.g., Aitken, 1978; Lund, 1991; Petersen, 1991; Richards, 1983; Weir, 1993). Perhaps the most influential of these taxonomies was Richards (1983). Richards created a taxonomy of micro-skills for both academic listening and

conversational listening. For academic listening, Richards listed 18 micro-skills, including the “ability to identify the purpose and scope of lecture”, the “ability to identify topic of lecture and follow topic development”, and the “ability to identify relationships among units within discourse” (p. 229). Similarly, Richards listed 33 micro-skills that make up conversational listening ability, including the “ability to retain chunks of language of different lengths for short periods”, “the ability to discriminate among the distinctive sounds of the target language”, and “the ability to recognize the functions of stress patterns of words” (p. 228).

In almost direct contrast to the taxonomy listing/creation are data-driven examinations of listening ability. These data-driven examinations have focused on L2 listening test performance, and how the characteristics of the texts and tasks used on those test affect test-taker performance. Two examples are Freedle and Kostin (1999) and Nissan et al. (1996). These two studies are similar, in that they examined test-taker performance on the TOEFL listening section in the hopes of identifying variables that would account for the variance in item difficulty on the tests. Freedle and Kostin examined three different item types: main idea, inference-application (in which test-takers must use their background knowledge to make the appropriate inferences), and inferencing (the test-taker must make appropriate inferences based on information that is available in the spoken text). Using a regression analysis, they found that main idea items and inference-application items were significantly easier than inferencing items, and identified a number of text variables (e.g., negatives, referentials, rhetorical organizers, fronted structures, concrete texts, multisyllabic words, lexical overlap, subject matter, pauses and fillers) that affected the difficulty of each of these item types. Similarly, using a regression analysis, Nissan et al. (1996) found five variables that contributed to the variance in item difficulty, although the three best predictor variables (i.e., lower frequency vocabulary, implicit information in the text, dialogue finishing with a statement and not a question) only accounted for 12% of the variance.

While these theory-based taxonomies and data-driven lists of variables that affect listening performance are useful in many ways (i.e., for curriculum developers, test writers, and teachers), they are less useful for research purposes. The taxonomies do not “provide clear definitions or non-redundant orderings of components in any systematic graded hierarchy” (Dunkel, Henning, & Chaudron, 1993, p. 182), nor do they usually provide a hierarchy of which of the sub-skills are more important or fundamental to the listening process. Indeed, these taxonomies are generally hypothetical listings of things involved in listening ability, and they usually have little or no empirical research validating them. In contrast, while the data-driven variables are useful in identifying what aspects of listening are difficult, they tend to be focused on the lower-order components of listening, and are less able to provide a bigger picture definition or overall model of listening ability.

Indeed, it is this identification of the shortcomings of previous taxonomic papers and data-driven listings of listening variables that motivated Buck and Tatsuoka's (1998) empirical exploration of the knowledge, skills and abilities that are part of L2 listening ability. They used a rule-space procedure to identify the knowledge and cognitive processing skills required on their particular L2 listening test, and then examined how each of these attributes affected the listeners' performance on that test (rule-space methodology is a type of statistical pattern recognition technique that provides information about how individual test-takers perform on each of the attributes). 412 Japanese test-takers completed an open-ended, short-answer listening comprehension test, and rule-space methodology was used to provide information about the cognitive attributes ("anything that affects performance on a task: either a task characteristic, or any of the knowledge, skills or abilities necessary to complete the task" [p. 121]) that contributed to test performance. They identified 15 primary attributes and 14 interaction attributes that explained 96% of the variance for 96% of the test-takers. The 15 primary attributes included things like "the ability to scan fast spoken text, automatically and in real time", "The ability to use previous items to help information location", and "The ability to understand and utilize heavy stress", and "The ability to make text-based inferences" (pp. 141–142). Their list of 15 primary attributes mirrors skills or sub-skills identified in many of the listening taxonomies described above. The authors acknowledge that test-taker attributes interact with the attributes of the test itself (the spoken text and items), and thus they avoid some of the pitfalls of the taxonomies that focused solely on the listeners' ability, and the data-driven lists, that focused solely on the characteristics of the test. While Buck and Tatsuoka's study was informative, it was based on a single set of test-takers taking one particular test, and the "model" that resulted from it was never really developed beyond the results of the study.

1.2 Models of second language (L2) listening ability

Many of the models of L2 listening ability described here are focused on assessment. This is not surprising, because principled assessment entails choosing or creating a construct definition of the ability to be assessed, and then an operationalization of that model.

A common theme in many models of L2 listening is the idea that listening involves both bottom-up processing and top-down processing. Bottom-up processing occurs when the listener perceives the aural input and tries to interpret and build up the meaning bit by bit, word by word (Kelly, 1991). It involves speech perception and word recognition, which provide the data for the listener in attempting to decode the utterance in trying to comprehend it (Rost, 2002). Top-down processing requires the listener to use their background, contextual, and world knowledge to

set up expectations and create a conceptualization of what the utterance means. Kelly (1991) describes it as “...the application of cognitive faculties in the attempt to give the sound input meaning. The mind sets up the expectations and the sound provides confirmation” (p. 135). The two processes work in tandem and are constantly influencing each other. At times one process might be dominant, and at other times the other process. This idea of the two different types of processing occurring simultaneously and integratively is appealing both logically and intuitively, and it is also helpful pedagogically, but in practice it is difficult to operationalize them separately, since there are so tightly interrelated.

Perhaps the most influential and useful model of L2 listening ability was presented by Buck (2001) in his book, *Assessing listening*. Based on his previous empirical research (e.g., Buck, 1991, 1994; Buck & Tatsuoka, 1998; Buck, Tatsuoka, Kostin, & Phelps, 1997), Buck provided what he called a “default listening construct” (p. 113). In presenting this construct definition, he provided a number of recommendations to consider when creating listening tests, including focusing on those aspects of language that are unique to listening; requiring the listener to understand basic linguistic information on different topics and texts; going beyond the literal meaning and also assessing inferred meanings; considering how to deal with knowledge-dependent interpretations of the text; and avoid assessments that assess general cognitive abilities (Buck, 2001). His formal definition of the default listening construct is the ability:

- to process extended samples of realistic spoken language, automatically and in real time,
- to understand the linguistic information that is unequivocally included in the text, and
- to make whatever inferences are unambiguously implicated by the content of the passage. (Buck, 2001, p. 114)

The first bulleted point addresses a number of issues that other models bypass or miss altogether. According to Buck, L2 listening ability involves the ability to process “extended” samples of spoken language. It involves the ability to process “realistic” spoken language (and in fact, many assessments purporting to assess a listener’s communicative competence often do not use spoken texts with “realistic” spoken language, a point that will be expanded on below). Listening requires the listener to process the spoken language “automatically and in real time”, again, addressing how listening actually occurs in real-world situations. The second bulleted point is relatively straightforward, involving the ability to understand explicitly stated information. The third bulleted point describes the ability to make appropriate inferences from the spoken input.

This “default listening construct” is useful because it is simple and straightforward, with just three bulleted points, yet these three points address the construct broadly and thoroughly, and it can be adapted to different listening situations to create a more contextualized definition of L2 listening ability.

Wagner (2002) reviewed the literature and posited a model of L2 listening ability based on previous L2 taxonomies and models, including Buck’s (2001) model. He posited a two-factor model of L2 listening ability in an academic listening domain. The two factors were “the ability to perform bottom-up processing” and the ability to “perform top-down processing”. The bottom-up processing factor was operationalized through two sub-skills “the ability to identify details and facts” in the text, and the ability to “recognize supporting ideas” in the spoken texts. The top-down processing factor was operationalized through four sub-skills: the ability to identify the controlling idea/gist; ability to make text-based inferences, the ability to make inferences about speakers’ attitudes and pragmatic meaning, and the ability to deduce vocabulary through context (Wagner, 2002, p. 12). Wagner operationalized this model of L2 listening ability and created a 20-item (multiple choice and limited production items) video listening test that he administered to 85 high school ESL students in the U.S. He then used the results of this test to perform a series of exploratory factor analyses (EFAs) to examine the validity of the theorized model of L2 listening. The results of these EFAs provided little evidence in support of the original model, but based on these results, Wagner re-interpreted the model as a two-factor model of L2 listening ability: the ability to listen for explicitly-stated information, and the ability to listen for implicit information. This revised model corresponds very closely to Buck’s (2001) model of L2 listening ability.

Wagner’s (2004) study built on his (revised) 2002 model. He posited a two-factor model of L2 listening ability: the ability to comprehend “explicitly stated spoken information”, and the ability to comprehend “implicit spoken information” (p. 9). He then tried to validate this model based on the results of test-takers’ performance on the listening sections of the MELAB ($n = 823$) and the ECPE ($n = 11,212$). The results of the EFAs showed limited support based on the MELAB data, but little evidence based on the ECPE data.

Rost’s (2011) book *Teaching and research listening* (2nd edition), is a very important book in the L2 listening literature, even though Rost did not explicitly define a model of listening ability. Instead, the first section of the book attempts to define listening by describing the various processing that listening entails, including neurological processing (i.e., consciousness, hearing, and attention); linguistic processing (i.e., perceiving speech, segmenting, grouping, and parsing); semantic processing (i.e., constructing meaning by integrating memory and prior experience); and pragmatic processing (i.e., constructing meaning by inferring speaker

intention). This exploration of the various processes involved in L2 listening is a useful resource for attempts at operationalizing models of L2 listening ability.

Vandergrift and Goh (2012) developed their working cognitive model for L2 listening comprehension based on Levelt's (1983, 1989, 1995) model of speech production. Although technically a model of speaking, Levelt's model of producing and monitoring speech is relevant because by integrating the speaking and comprehension components together in one model, it allows a conceptualization of both one-way and two-way (interactive) listening. In Vandergrift and Goh's (2012) model, listening begins with the perception of sound signals by an acoustic-phonetic processor. This information is stored very briefly in working memory while processed for meaning, and then the sounds are displaced by the subsequent incoming sounds. Analysis of the speech sounds then begins, involving both bottom-up and top-down processing. The next stage involves parsing, in which the phonetic representations of the perceived sounds are parsed for meaning, by segmenting the utterance. This segmenting is done using both syntactic and semantic cues. Again, the processing activity is done not linearly but in a parallel fashion, in which the bottom-up processing informs the top-down processing (and vice versa), until a reasonable mental representation of the meaning of the spoken utterance is created. This is done by the listener by linking the mental representation of the spoken sounds with the listener's existing knowledge from long-term memory. The listener creates a mental representation of their interpretation of what they have heard, and stores this in long-term memory.

Vandergrift and Goh stress that the aspect of their model that has not been adequately addressed in previous cognitive models of listening ability is the idea of metacognition, the idea that the listener can have conscious control of the listening process, including planning, monitoring, problem-solving, and evaluating. They argue that listeners with more metacognitive awareness are better able to control and regulate the cognitive processes involved in listening. Vandergrift and Goh stressed that it was a working model, because there is no widely accepted comprehensive theory that adequately explains the comprehension process. They also stressed that theirs was not a comprehensive model, acknowledging the fact that listening occurs in various social contexts, and thus a comprehensive model would have to include the affective components of listening.

1.3 L2 listening, L1 listening, and reading

Models of L1 reading and L1 listening have been influential in modeling and operationalizing L2 listening ability. Obviously, listening in an L2 is very similar to listening in one's L1. Indeed, virtually all of the models of L2 listening ability described here are based almost entirely on conceptualizations of L1 listening ability. Yet it also seems obvious that there are differences between L1 and L2 listening. Buck (2001) argues that they are fundamentally similar, yet differ in "emphasis" (p. 48), suggesting that L2 listeners are affected by an incomplete knowledge of the linguistic system, content or textual schemata, background knowledge and cultural knowledge.

Vandergrift (2006) examined whether L2 listening proficiency was more aligned with overall L2 language proficiency, or L1 listening ability. He wanted to examine empirically the linguistic interdependence hypothesis, which predicts that L2 listening performance is largely predicted by L1 listening ability. The idea is that "learners do not relearn a language skill; rather the skill is available to them as needed within another language context" (Vandergrift, 2006, p. 6), and Vandergrift argues that this hypothesis has been researched extensively in the L2 reading literature (e.g., Bernhardt & Kamil, 1995; Lee & Schallert, 1997). Because this literature suggests that both L2 proficiency and L1 reading proficiency are predictors of an individual's L2 reading proficiency, Vandergrift wanted to see if the same applies for L2 listening. He divided 75 L1 English speaking 8th graders in Canada who were learning French into a high ability and a low ability group. The 75 participants took both a French (L2) listening test and an English (L1) listening comprehension test. He found that while both L2 proficiency and L1 listening ability predicted group membership (high ability or low ability), and that L2 proficiency accounted for about 25% of the variance, L1 listening ability accounted for about 14% of the variance. The results from Vandergrift (2006) seem to be in line with Hulstijn's (2015) BLC-HLC model of language ability, which posits that higher language cognition (HLC) is a complement or extension of basic language cognition (BLC – implicit and unconscious phonetic, prosodic, phonological, morphological, and syntactic knowledge, in conjunction with explicit and conscious lexical knowledge). According to this theory, linguistic knowledge explains most variance in language performance, but as proficiency increases, other peripheral factors such as general cognitive ability, L1 ability, metacognition, etc., are able to explain at least some of the variance in language performance.

Just as there are many similarities between L1 and L2 listening, there are also a number of similarities between listening and reading (including L2 listening and L2 reading), and conceptualizations and models of reading have been influential on conceptualizing and modeling L2 listening ability. Indeed, the two processing models that began this chapter are prime examples of this phenomenon. Anderson's

cognitive processing model is applicable for both reading and listening, although Paivio's dual coding theory goes much farther in acknowledging the different modalities involved.

Aotani (2011) reviewed the research related to the similarities and differences between L2 reading and listening. He described two opposing views in the literature. The unitary process view is based on the idea that L2 listening and L2 reading are more or less the same, except that the mode of the input is different (spoken versus written). In contrast, the dual process view holds that there are real and measurable differences between L2 reading and listening. Aotani argued that the dual process view is now more accepted in the field. As evidence for this, numerous researchers (e.g., Buck, 2001; McCarthy & Carter, 1995; Vandergrift, 2007; Wagner 2014a, 2014b, 2018) have described how L2 reading and listening differ, especially due to the online nature of listening, which prevents listeners from having the chance to go back and re-process the input. In addition, the non-verbal and paralinguistic (i.e., stress, rhythm, and intonation) components of spoken language are important parts of listening, and do not have equivalents in reading. Finally, the nature of written texts differs fundamentally from the nature of spontaneous spoken texts, as will be discussed below.

2. Measurement practices: A critical discussion of how the constructs and their components have been operationalized and measured by empirical researchers

L2 listening ability is more difficult to operationalize and measure than the other language skills (i.e., reading, writing, speaking) for a number of reasons. Like reading, it is an internal process. Listening is also difficult to measure because it requires the selection or creation of spoken texts to present to the listeners, and choosing or creating the appropriate texts is surprisingly more difficult than many researchers anticipate. Presenting the spoken texts is also challenging, as the researcher must decide whether to do it "live" (with a human speaking the text), or utilize technology such as audio or audio-visual recordings. Numerous researchers have argued that when testing listening ability, it is necessary to include and even focus on those aspects that are unique to listening ability (e.g., Buck, 2001; Rost, 2011; Wagner, 2014b), yet many of the characteristics that are unique to listening have not been included in the operationalization and measurement of listening ability. What follows is a critical discussion of the way L2 listening ability has been operationalized and measured in the field, focusing on: listening task response, differentiating between reading and listening, the length of the spoken text used as the input, assessing interactive speaking/listening ability, using audio-only versus audio-visual input, the

accent/dialect variety of the spoken input, and the use of real-world spoken input. This critical discussion is grounded in the idea of the need to include those aspects of listening that are unique to listening when operationalizing and measuring L2 listening ability.

2.1 Interactive speaking/listening ability

Listening has traditionally been operationalized and measured as a one-way activity. Obviously, one-way listening is common, and part of many domains of interest for L2 researchers and testers. Academic listening (including listening to lectures) often involves one-way listening. But listening ability is often intertwined with interactive speaking and listening, in which the participant is both a listener and speaker. There seems to be growing recognition that the operationalization of listening ability should also include listening ability as part of an interactive speaking/listening ability (Ockey & Wagner, 2018). There also seems to be growing recognition that while isolating a single skill might be appropriate in some cases, the overarching idea of interaction and communicative competence in second language acquisition research necessitates broadening the construct.

Nevertheless, even with this increased awareness of listening being part of a broader communicative construct, many researchers have focused on the one-way, listening-only aspect of listening ability. There are probably many reasons for this, including the fact that listening researchers want to focus on listening only, and not on speaking. In addition, assessing listening ability as part of interactive speaking/listening ability certainly presents definite measurement challenges (Nakatsuhara, 2018; Ockey, 2018). Nevertheless, these are challenges that L2 listening researchers will have to address, because conceptualizing and operationalizing the incredibly broad and diverse construct of listening ability without including interactive speaking/listening ability results in an obviously impoverished construct.

2.2 Differentiating listening from reading

As described above, numerous researchers have described the many differences between reading and listening, and there is an increasing realization that it is necessary to think of them as different processes. Field (2008) describes how L2 listening ability has frequently been conceptualized as a set of sub-skills, and that these sub-skills combine and interact in the overall process of listening. This idea of identifiable and distinct sub-skills is common in reading, and thus has been applied to listening, but Buck (2001) criticizes the notion that the reading “subskills” can be

directly applied to listening. He argues that what these sub-skills are in listening is unclear, and also describes the difficulty in operationalizing and measuring them, and differentiating between them.

Yet many models and operationalizations of listening ability seem to continue to depict listening as essentially the same as reading, except for the input being verbal rather than written. A good example of this are studies that have examined the amount of shared variance between L2 reading and L2 listening ability (e.g., Aotani, 2011; Song, 2008). These studies have found that listening and reading do correlate quite highly. However, the tests of listening ability in these studies seem to be essentially L2 reading tests that have written input that is read aloud, rather than including real-world spoken input, and so it is possible that the two are overly conflated. Again, Buck (2001) and Rost (2011) argue that when operationalizing and assessing listening ability, it is important to focus on those aspects that are unique to listening ability. Indeed, many L2 listening tests continue to operationalize listening ability as if it were the same as reading ability.

In his (2001) default listening construct, Buck very specifically described how proficient L2 listeners had to be able to do automatic processing, and that listeners did not have the luxury of utilizing controlled processing in most listening contexts. Because of the nature of spoken texts, the rate of the input is not controlled by the listener, but by the speaker. This is in great contrast to reading, in which the reader controls the rate of input, and thus is more able to utilize controlled processing in processing the input. This leads to two natural critiques of how L2 listening has been operationalized. The first critique has to do with the number of times a spoken text is presented to test-takers. Buck (2001) argues that it is artificial and problematic to present the spoken text multiple times to test-takers. Doing so allows them to utilize controlled processing, and is inauthentic in that, in most real-life listening contexts, the listener does not have the opportunity to listen to the text multiple times. The second critique has to do with the speech rate of the spoken input. The research is clear that speech rate affects L2 listening comprehension (e.g., East & King, 2012; Griffiths, 1992; McBride, 2011). Basically, there is an inverse relationship between the rate of the input and comprehension – the higher the speech rate, the lower the comprehension (there are caveats, of course, but this is the general rule). Again, according to Buck (2001), using artificially low rates of spoken input allows L2 listeners to utilize controlled processing, when in real-life situations, this would not be possible. Similarly, Wagner (2016); Wagner and Wagner (2016), and Wagner and Ockey (2018) have criticized the use of carefully scripted and overly-enunciated spoken texts in L2 listening assessments, in part because they have lower speech rates than real-world speaking and listening contexts, and thus are not representative of the ability needed in most listening contexts.

2.3 Length of the spoken text

Buck's (2001) default listening construct also explicitly included the importance of listening assessments including "extended samples" of spoken language (p. 114). This seemed to be in response to earlier operationalizations and assessments that only included listening to very short (word or sentence level) spoken texts. Obviously, listening ability involves the ability to comprehend spoken texts of varying lengths, from word and sentence level utterances to very long texts (e.g., academic lectures). Using only very short oral texts is attractive from a practicality standpoint, as shorter texts are easier to create and easier to administer, and their use allows for better control of other variables (i.e., memory). But to accurately assess listening ability, spoken texts of varying lengths are needed, in order to adequately represent the types of listening that would be expected in the particular domain of interest. Somewhat surprisingly, there does not seem to be any recent empirical research investigating how the length of a spoken text can affect L2 listening performance, although obviously listeners need to be able to understand spoken texts of varying lengths.

2.4 Listening task response

Because it is not possible to "get inside the listener's head", listening tests usually require test-takers to listen to spoken input, and then do something with that input to demonstrate comprehension (i.e., answer comprehension questions, respond orally, fill out a chart, do a dictation, etc.). Based on their responses, the researchers make inferences about their actual listening ability. Thus, the type of task response that is utilized is very important in operationalizing and measuring the construct.

One seemingly obvious task type would be a written recall or dictation. But a dictation task presents problems when administering (when to insert pauses for writing, how many times to play the text), and with scoring because scoring a dictation reliably is surprisingly difficult (Buck, 2001). Another concern is that dictation includes much more than just listening ability (i.e., writing ability, memory, etc.). And of even more concern, when it comes to dictation, it is unclear to what extent a dictation task actually assesses listening comprehension. Requiring listeners to transcribe a spoken text word for word is very unlike real-world listening, where the listener has to segment the input and ascertain the semantic and syntactic meaning of this input, and compare this information with the listener's background knowledge (and contextual and co-textual knowledge) (Wagner, 2014a). In addition, dictation probably does not assess listeners' ability to make inferences (Buck, 2001), which would seem to be a vital component of listening ability. Even with

these limitations, dictations can be useful for L2 listening researchers as a listener's response can be analyzed to see what aspect of listening to the spoken text the listener had difficulty with (e.g., vocabulary, a particular grammatical or syntactic structure, etc.), although again listeners tend to focus on semantic processing, while writing the dictation would seemingly require syntactic processing. Listening cloze tasks are similar to dictations tasks, and have many of the same shortcomings as dictation tasks. Oral repetition tasks, in which the test-taker has to listen to a spoken text (usually at the word, phrase, or sentence level) and then orally repeat the input, also have much in common with dictation.

Listening summarization tasks are those in which the test-taker listens to a spoken text, and then has to write or say a summary of the text, or recall as much of the text as they can. They differ from dictations in that the listener is not expected to provide a word-for-word recall of the spoken text. Instead, the listener summarizes in their own words what they have heard. Such a task is more similar to real life listening contexts than dictation tasks, which require the word-for-word recall. Some disadvantages of these types of tasks include scoring challenges, due to the fact that writing (or speaking) ability is also being assessed. Also being assessed is summarization ability as well as memory, and it is unclear the extent to which these two abilities are part of the construct of L2 listening.

In contrast to these dictation-type tasks, comprehension tasks are probably more commonly used in many testing situations, as well as in research areas in which listening ability is operationalized and measured. Perhaps the most common type of comprehension task involves comprehension questions that the listener must answer based on what s/he heard. Most listening tests employ some variety of listening tasks involving discrete point comprehension questions, in which the listener listens to a spoken text and then chooses the correct answer (i.e., multiple-choice) or writes or states the correct answer (i.e., short answer constructed response). These types of tasks have many advantages, in that they are relatively easy to create and score reliably, and they seem to be able to tap into the listening comprehension process. Researchers have used these types of comprehension tasks extensively, in part because they have the advantage of allowing the researcher to assess the different components of their particular model of listening ability. Thus, a particular comprehension question (or questions) can be created that purportedly tap into a particular ability, skill, or sub-skill of listening. Within a single testing situation, multiple questions can be used to assess each of the particular abilities/skills of interest, thus resulting in a more reliable and valid assessment.

2.5 Audio-only versus audio-visual input

Whether listening ability involves the ability to comprehend nonverbal information is a question that has vexed L2 listening researchers for decades. On the one hand, listening ability has traditionally been focused on a person's ability to process and comprehend verbal information, and there is the sentiment that the non-verbal components of spoken language were somehow extraneous to actual listening ability (Buck, 2001; Coniam, 2001). In contrast, Wagner (2008, 2010a, 2010b, 2013) has argued that in most domains, the ability to process the visual, nonverbal information that accompanies a speaker's verbal output should be considered part of the construct of listening ability. In most real-world listening contexts, the listener can see the speaker, and the visual, nonverbal information provided by the speaker's body language, lip movements, and gestures, as well as the contextual information provided by the setting (i.e., the speaker's physical appearance, the physical background setting, etc.). The listener processes this visual information while simultaneously processing the verbal information, and uses the different input sources to try and make sense of the message (Paivio, 1971, 1986, 1991). Similarly, Ockey (2007) has argued that an "expansion of the construct" is necessary, in order to go beyond this traditional notion of listening ability including only aural input. Yet many models and operationalizations of listening ability include only aural input, or include nonverbal information only as an afterthought.

2.6 Accent and dialect variety

When operationalizing models of L2 listening ability, a researcher must identify the accent variety or varieties of the spoken texts used for the aural/audio-visual input. Traditionally, the perceived standard variety of the language in that research context has been used. Using English as an example, if a researcher was conducting research in the UK, then standard British English was used as the language variety of the input, and seemingly "accented" speech was avoided. One of the problems with this approach, of course, is identifying the "standard" variety, as well as who gets to determine what the standard variety is (or even if a "standard" variety actually exists). Widespread languages will have a variety of accents and dialects, based on national, regional, cultural, and other differences. In addition, many speakers of a language are non-native speakers of that language. Indeed, English is spoken by more nonnative speakers of English than by native speakers.

A highly proficient listener is able to comprehend multiple varieties of that language, not just the standard variety. Two areas of research are relevant here. One, multiple studies have found that familiarity with an accent leads to increased

comprehension of that accent (e.g., Gass & Varonis, 1984). Second, it is widely accepted in L2 pronunciation research that a speaker can be accented, but still be comprehensible. Accentedness and comprehensibility are correlated, but still separate constructs (Isaacs, 2008). For many languages (and certainly English), a proficient listener must have multidialectal listening skills (be able to understand a number of different speech varieties), as well as be able to accommodate or adapt to unfamiliar varieties of spoken language (Canagarajah, 2006).

2.7 Real-world spoken input

Perhaps the most problematic way that L2 listening ability has been modeled and operationalized involves the inauthentic nature of the spoken texts used in these operationalizations. Depending on the context of the speaking situation, a speaker might be able to plan exactly what s/he is going to say, or s/he might have virtually no chance to plan the speech. Wagner (2013, 2014a, 2014b, 2018) has described how there is a “continuum of scriptedness” with spoken language. At one end of the continuum is speech that is totally planned. This might involve writing a text, editing and polishing it, and then speaking it aloud, with careful enunciation. At the other end of the continuum is speech where the speaker composes the message and verbally utters it at virtually the same time, with no planning. This level of scriptedness affects the characteristics of the speech. As has been demonstrated repeatedly in the literature, unplanned and unscripted spoken language is “messy”, and is filled with connected speech, filled and unfilled pauses, repetitions, false starts, etc. In contrast, texts that are scripted and then read aloud might have very few of these phenomena.

That spoken texts vary in the extent to which they have these different phenomena is not in doubt or controversial. There is a long history of applied linguistics researchers investigating these phenomena in speech. It seems obvious that a proficient listener would be able to listen to and comprehend spoken texts with differing levels of scriptedness (from different points on the continuum of scriptedness). Yet when listening ability is operationalized through assessments that involve listening to spoken language, the spoken language almost invariably comes from the scripted end of the continuum (Field, 2013; Wagner 2016; Wagner & Wagner, 2016; Yanagawa, 2016). This is problematic for two main reasons. First, the unscripted spoken language is much more prevalent in most real world speaking and listening contexts. In most cases, the speaker does not have time to plan, and thus language that has connected speech, hesitation phenomena and non-linear organizational patterns is the norm, not the exception. Second, an increasing amount of research (e.g., Carney, 2018; Wagner, 2018; Wagner & Toth, 2014) demonstrates that many

L2 listeners have more difficulty comprehending unscripted spoken language (that has these characteristics of real-world, unplanned spoken language) compared to scripted spoken language. Researchers such as Gilmore (2007) and Wagner (2014a) have argued that many classroom L2 learners are exposed primarily to “textbook texts”, texts that are scripted and created especially for L2 learners, and that lack the characteristics of real-world, unplanned spoken language. This lack of exposure to and teaching about real-world spoken language can result in L2 learners that are proficient in understanding scripted spoken language, yet are unable to understand real-world spoken language (Brown & Trace, 2018; Wagner 2018; Wagner & Ockey, 2018). Again, a proficient listener needs to be able to comprehend different varieties of spoken texts, from a variety of genres, and from a variety of formality levels. As Wagner has repeatedly argued (Ockey & Wagner, 2018; Wagner, 2014b, 2016; Wagner & Wagner, 2016), to assess L2 listeners’ listening proficiency using only spoken texts that are scripted and read aloud results in an overly narrow operationalization of the construct of L2 listening ability. This critique applies not only to L2 listening assessment, but also to SLA research purporting to investigate L2 listening within a communicative competence framework.

3. Conclusion

It is easy to critique the way L2 listening ability has been modeled and operationalized in previous research, as none of these critiques are new, and are found widely in the literature. Numerous researchers have argued that if the goal is to measure listening ability, then the measurement must include and even focus on those aspects that are unique to listening ability (e.g., Buck, 2001; Rost, 2011; Wagner, 2014b). But as the critique above demonstrates, many of the characteristics that are unique to listening have been consciously or unconsciously avoided in the operationalization and measurement of listening ability. The question then becomes “Why have these aspects unique to listening been avoided in its modeling and operationalization?” Although it is probably impossible to fully answer this question, what follows is an (speculative, yet informed) examination of some of the probable reasons.

The first reason is tradition. L2 listening ability has traditionally been operationalized (and tested, and taught) as one-way, non-interactive listening, using scripted, audio-only texts, that are revised and edited and then spoken aloud, often with a speaker especially trained to enunciate clearly. In relation to the spoken texts used in L2 textbooks and assessments, Gilmore (2007) and Wagner (2014) have argued that these industries are conservative, and reluctant to change the status quo. Related to this idea of conservatism, Wagner (Wagner, 2014b, 2018; Wagner & Wagner, 2018) has argued that the L2 testing industry might be reluctant to use

unscripted spoken texts because they might sound unprofessional. In other words, using recorded spoken texts that have filled pauses, overlaps, digressions, repetitions, and false starts, and with lots of reduced and connected speech, might cause test-takers (and other test users) to perceive the test as unprofessional and second rate. While it is certainly understandable that test developers might be reluctant to use spoken texts that sound unprofessional, Wagner (2018) argues that test developers can avoid this problem by stressing that they are using unscripted, authentic, spoken texts that have the characteristics of real-world spoken language, and thus their tests better assess the communicative competence of the test-takers.

Secondly, it is more difficult, less efficient, and more expensive to identify and find authentic spoken texts that are suitable for L2 listening tests than it is to create spoken texts that address the particular needs of a researcher or test developer. Wagner (2014b, 2018) describes how in most testing contexts, test developers have a number of pre-determined test specifications (including length/duration, number of questions, skills to be assessed, etc.), and it is much easier to create a spoken text that is specifically scripted to address these test specifications, than it is to find a real-world text that would be suitable. Similarly, using an audio-visual text presents cost and logistical issues for test developers – more expenses related to the creation and delivery of audio-visual texts. But if the goal is to operationalize a model of L2 listening ability that really tries to address communicative competence, and genuinely tries to assess a language user's ability to understand spoken language in real world contexts (rather than in a language laboratory), then the added expense involved in using authentic spoken texts seems necessary to incur.

Thirdly, it seems likely that many of the models of listening are based at least in part on models of reading. As described above, both are internal processes involving the processing of input, and they obviously have many similarities. Anderson's cognitive model and Paivio's dual coding theory are applicable to both reading and listening. It seems likely that research into reading has been more prevalent than listening research, because reading must be taught even in the L1, while listening is rarely taught in the L1. Within the cognitive paradigm, the idea is that the cognitive processing goes on inside the mind of reader/listener, and the focus is then on the input and the processing, and does not acknowledge some of the "messier" aspects of real-life communication. Again, this is more defensible with reading, where the written input is on the printed page, and there is a greater degree of uniformity than with listening, which has a multitude of characteristics that can affect the input (e.g., accentedness, background noise, mumbling, etc.). This reliance on reading as the dominant processing paradigm seems to have led to listening being operationalized as just a variation of reading, involving the processing of oral, rather than written text. The problem is that an unplanned, spontaneous oral text is very different from a planned, scripted, edited, and polished written text that is read aloud slowly with

clear enunciation. Again, listening and reading have many similarities, but rather than operationalizing them as if they are virtually identical (except for the channel of the input) it is necessary to include the characteristics that are unique to listening when operationalizing listening.

This also seems analogous to the competence/performance distinction made by Chomsky and his acolytes in transformational grammar. The way listening has been operationalized focuses on the idea of competence (the internal state of the mind), while real-world listening performance (and listening to spoken input with real-world linguistic characteristics) is neglected. There seems to be the recognition that these real-world characteristics of unplanned spoken language (e.g., hesitation phenomena, oral grammatical norms, connected speech, non-linear discursive organizational patterns) exist, but somehow they are superfluous to the underlying competence involved in listening. The problem with this view, of course, is that listening happens in the real world, where spoken texts are messy, with backtracking and false starts, repetitions, slang, connected speech, and mumbling. Field (2013) criticizes the way L2 listening has been operationalized on many listening tests, and argues for the need that these operationalizations have cognitive validity, which he defines in relation to assessing listening as “the extent to which the tasks employed succeed in eliciting from candidates a set of processes which resemble those employed by a proficient listener in a real-world listening event” (p. 77).

To conclude, when modeling and operationalizing L2 listening ability, it is important that researchers consider the nature of real-world listening. Real-world listening is “messy”, and requires a host of skills beyond the previous idealized, listening-as-reading operationalizations.

References

- Aitken, K. (1978). Measuring listening comprehension. *English as a second language. TEAL Occasional Papers* (Vol. 2). British Columbia Association of Teachers of English as an Additional Language. (ERIC Document Reproduction Service No. ED155945)
- Anderson, J. (1985). *Cognitive psychology and its implications* (2nd ed.). Freeman.
- Anderson, J. (1990). *Cognitive psychology and its implications* (3rd ed.). Freeman.
- Anderson, J. (1995). *Cognitive psychology and its implications* (4th ed.). Freeman
- Aotani, M. (2011). Factors affecting the holistic listening of Japanese learners of English (Unpublished doctoral dissertation). Temple University Japan, Tokyo, Japan.
- Bernhardt, E., & Kamil, M. (1995). Interpreting relationships between L1 and L2 reading: Consolidating the linguistic interdependent hypotheses. *Applied Linguistics*, 16(1), 15–34. <https://doi.org/10.1093/applin/16.1.15>
- Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics*, 18, 171–191. <https://doi.org/10.1017/S0267190500003536>

- Brown, J. D., & Trace, J. (2018). In G. Ockey & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp. 45–63). John Benjamins.
- Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing*, 8(1), 67–91. <https://doi.org/10.1177/026553229100800105>
- Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing*, 11(2), 145–170. <https://doi.org/10.1177/026553229401100204>
- Buck, G. (2001). *Assessing listening*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511732959>
- Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, 15(2), 119–157. <https://doi.org/10.1177/026553229801500201>
- Buck, G., Tatsuoka, K., Kostin, I., & Phelps, M. (1997). The sub-skills of listening: Rule-space analysis of a multiple-choice test of second language listening comprehension. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment*. Universities of Tampere and Jyväskylä.
- Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly*, 3(3), 229–242. https://doi.org/10.1207/s15434311laq0303_1
- Carney, N. (2018). Diagnosing L2 bottom-up listening abilities of Japanese university EFL learners (Unpublished doctoral dissertation). Temple University Japan, Tokyo, Japan.
- Coniam, D. (2001). The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: A case study. *System*, 29, 1–14. [https://doi.org/10.1016/S0346-251X\(00\)00057-9](https://doi.org/10.1016/S0346-251X(00)00057-9)
- Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of an L2 listening comprehension construct: A tentative model for test specification and development. *The Modern Language Journal*, 77(2), 180–191. <https://doi.org/10.1111/j.1540-4781.1993.tb01962.x>
- East, M., & King, C. (2012). L2 learners' engagement with high stakes listening test: Does technology have a beneficial role to play? *CALICO Journal*, 29, 208–223. <https://doi.org/10.11139/cj.29.2.208-223>
- Field, J. (2008). *Listening in the language classroom*. 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 University Press.
- Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing*, 16(1), 2–32. <https://doi.org/10.1177/026553229901600102>
- Gass, S., & Varonis, M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, 34(1), 65–89. <https://doi.org/10.1111/j.1467-1770.1984.tb00996.x>
- Gilmore. (2007). Authentic materials and authenticity in foreign language learning. *Language Teaching*, 40(2), 97–118. <https://doi.org/10.1017/S0261444807004144>
- Griffiths, R. (1992). Speech rate and listening comprehension: Further evidence of the relationship. *TESOL Quarterly*, 26(2), 385–391. <https://doi.org/10.2307/3587015>
- Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research*. John Benjamins. <https://doi.org/10.1075/llt.41>

- Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *Canadian Modern Language Review*, 64(4), 555–580. <https://doi.org/10.3138/cmlr.64.4.555>
- Kelly, P. (1991). Lexical ignorance: The main obstacle to listening comprehension with advanced foreign language learners. *IRAL*, 29, 135–149.
- Lee, J., & Schallert, D. (1997). The relative contribution of L2 language proficiency and L1 reading ability to L2 reading performance: A test of the threshold hypothesis in an EFL context. *TESOL Quarterly*, 31(4), 713–739. <https://doi.org/10.2307/3587757>
- Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, 14(1), 41–104. [https://doi.org/10.1016/0010-0277\(83\)90026-4](https://doi.org/10.1016/0010-0277(83)90026-4)
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. The MIT Press.
- Levelt, W. J. M. (1995). The ability to speak: From intentions to spoken words. *European Review*, 3(1), 13–23. <https://doi.org/10.1017/S1062798700001290>
- Lund, R. (1991). A comparison of second language listening and reading comprehension. *The Modern Language Journal*, 75(2), 196–204. <https://doi.org/10.1111/j.1540-4781.1991.tb05350.x>
- McBride, K. (2011). The effect of rate of speech and distributed practice on the development of listening comprehension. *Computer Assisted Language Learning*, 24(2), 131–154. <https://doi.org/10.1080/09588221.2010.528777>
- McCarthy, M., & Carter, R. (1995). Spoken grammar: What is it and how can we teach it? *ELT Journal*, 49(3), 207–218. <https://doi.org/10.1093/elt/49.3.207>
- Nakatsuhara, F. (2018). Investigating examiner interventions in relation to the listening demands they make on candidates in oral interview tests. In G. Ockey & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp. 205–226). John Benjamins. <https://doi.org/10.1075/llt.50.14nak>
- Nissan, S., DeVenicenzi, F., & Tang, K. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension* (TOEFL Research Report No. 51). Educational Testing Service.
- Ockey, G. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, 24(4), 517–537. <https://doi.org/10.1177/0265532207080771>
- Ockey, G. (2018). The degree to which it matters if an oral test task requires listening. In G. Ockey & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp. 193–204). John Benjamins. <https://doi.org/10.1075/llt.50.13ock>
- Ockey, G., & Wagner, E. (2018). *Assessing L2 listening: Moving towards authenticity*. John Benjamins. <https://doi.org/10.1075/llt.50>
- Paivio, A. (1971). *Imagery and verbal processes*. Holt, Rinehart, and Winston.
- Paivio, A. (1986). *Mental representations: A dual-coding approach*. Oxford University Press.
- Paivio, A. (1991). Dual coding theory: retrospect and current status. *Canadian Journal of Psychology*, 45(3), 255–287. <https://doi.org/10.1037/h0084295>
- Peterson, P. (1991). A synthesis of methods for interactive listening. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (2nd ed., pp. 106–122). Newbury House.
- Richards, J. C. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly*, 17(2), 219–240. <https://doi.org/10.2307/3586651>
- Rost, M. (2002). *Teaching and researching listening*. Pearson Education.
- Rost, M. (2011). *Teaching and researching listening* (2nd ed.). Pearson.
- Song, M. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing*, 25(4), 435–464. <https://doi.org/10.1177/0265532208094272>

- Vandergrift, L. (2006). Second language listening: Listening ability or language proficiency. *The Modern Language Journal*, 90(1), 6–18. <https://doi.org/10.1111/j.1540-4781.2006.00381.x>
- Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension. *Language Teaching*, 40(3), 191–210. <https://doi.org/10.1017/S0261444807004338>
- Vandergrift, L., & Goh, C. (2012). *Teaching and learning second language listening: Metacognition in action*. Routledge.
- Wagner, E. (2002). Video listening tests: A pilot study. *Working Papers in TESOL & Applied Linguistics, Teachers College, Columbia University*, 2(1). Retrieved on 1 May 2019 from <http://www.tc.edu/tesolalwebjournal>
- Wagner, E. (2004). A construct validation study of the extended listening sections of the ECPE and MELAB. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, 2, 1–26.
- Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly*, 5(3), 218–243. <https://doi.org/10.1080/15434300802213015>
- Wagner, E. (2010a). Test-takers' interaction with an L2 video listening test. *System*, 38(2), 280–291. <https://doi.org/10.1016/j.system.2010.01.003>
- Wagner, E. (2010b). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing*, 27(4), 493–513. <https://doi.org/10.1177/0265532209355668>
- Wagner, E. (2013). An investigation of how the channel of input and access to test questions affect L2 listening test performance. *Language Assessment Quarterly*, 10(2), 178–195. <https://doi.org/10.1080/15434303.2013.769552>
- Wagner, E. (2014a). Using unscripted spoken texts to prepare L2 learners for real world listening. *TESOL Journal*, 5(2), 288–311. <https://doi.org/10.1002/tesj.120>
- Wagner, E. (2014b). Assessing listening. In A. Kunnan (Ed.), *Companion to language assessment* (Vol. 1, pp. 47–63). Wiley-Blackwell.
- Wagner, E. (2016). Authentic texts in the assessment of L2 listening ability. In J. Banarjee & D. Tsagari (Eds.), *Contemporary second language assessment* (pp. 438–463). Continuum.
- Wagner, E. (2018). A comparison of L2 listening performance on tests with scripted or authenticated spoken texts. In G. Ockey & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp. 29–44). John Benjamins. <https://doi.org/10.1075/llt.50.03wag>
- Wagner, E., & Wagner, S. (2016). Scripted and unscripted spoken texts used in listening tasks on high stakes tests in China, Japan, and Taiwan. In V. Aryadoust & J. Fox (Eds.), *Current trends in language testing in the Pacific Rim and the Middle East: Policies, analyses, and diagnoses* (pp. 103–123). Cambridge Scholars.
- Wagner, E., & Ockey, G. (2018). An overview of the use of authentic, real-world spoken texts on L2 listening tests. In G. Ockey & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp. 13–28). John Benjamins. <https://doi.org/10.1075/llt.50.c2>
- Wagner, E. & Toth, P. (2014). Teaching and testing L2 Spanish listening using scripted versus unscripted texts. *Foreign Language Annals*, 47(3), 404–422. <https://doi.org/10.1111/flan.12091>
- Weir, C. (1993). *Understanding and developing language tests*. Prentice Hall.
- Yanagawa, K. (2016). Examining the authenticity of the Center Listening Test: Speech rate, reduced forms, hesitation and fillers, and processing levels. *JACET Journal*, 60, 97–115.

