Studies in Second Language Acquisition **0** (2019), 1–28
doi:10.1017/S0272263119000676

Research Article

THE RELATIVE SIGNIFICANCE OF SYNTACTIC KNOWLEDGE AND VOCABULARY KNOWLEDGE IN SECOND LANGUAGE LISTENING ABILITY

Payman Vafaee*

QA

Teachers College, Columbia University

Yuichi Suzuki

Kanagawa University

Abstract

The main purpose of the current study was to examine the relative significance of vocabulary knowledge (VK) and syntactic knowledge (SK) in second language (L2) listening ability, while accounting for the effect of several cognitive and affective factors. A total of 263 English-as-a-foreign-language learners took a standardized listening test (IELTS), as well as a battery of nine linguistic (two aural SK tests and two aural VK tests—covering both breadth and depth of VK), cognitive (two working memory tests and a metacognitive knowledge questionnaire), and affective measures (two L2 listening anxiety questionnaires). Structural equation modeling analysis revealed that both VK and SK were significant predictors of L2 listening ability; however, VK was a stronger predictor with an effect size being almost twice as much as the one for SK (.55 vs. .28). The results also showed that metacognitive knowledge, working memory, and L2 listening anxiety are significant predictors of L2 listening ability.

We would like to thank professors Steve Ross, Mike Long, Robert DeKeyser, and Catherine Doughty and the SSLA editors and anonymous reviewers for their thoughtful suggestions and constructive feedback. We are also grateful to the participants of our study, and Behrooz Yaghmayean, Mehdi Yaghoubi, Farokh Ziyaee, and Masaki Eguchi, who helped us with our data collection and the revision of the article. All errors and omissions are, of course, our own.

*Correspondence concerning this article should be addressed to Dr. Payman Vafaee, Zankel 313, Box 66, 525 W. 120th Street, New York, NY 10027. Teachers College of Columbia University. E-mail: pv2203@tc.columbia.edu

INTRODUCTION

As a result of a growing number of empirical studies in recent years, perhaps it is not true anymore to claim that second language (L2) listening is "the least understood and least researched skill" (Vandergrift, 2007, p. 191). However, still "little is known about the variables that contribute to the development of L2 listening ability" (Vandergrift & Baker, 2015, p. 390). These contributing variables are a set of general linguistic and cognitive factors that explain individual differences in L2 listening ability. The purpose of the current study was to investigate the role and relative significance of two major linguistic variables in L2 listening ability, that is, vocabulary knowledge (VK) and syntactic knowledge (SK). These two types of linguistic knowledge, along with phonological knowledge, are recognized as the most important linguistic components of L2 listening ability (Bloomfield et al., 2010).

VK has been presented in models of L2 listening comprehension as a way of assigning meaning, at least in part, to aural input. SK has likewise been recognized as essential for the perception and interpretation of aural language (Buck, 2001; Rost, 2013). However, empirical studies on the contribution of these two linguistic variables have been rare (Mecartty, 2000). Furthermore, existing studies had several methodological limitations that invite the further investigation of the topic. For example, in the existing studies, the two constructs of VK and SK were measured in a limited way. For one, VK is a multidimensional construct comprising of, at least, the two dimensions of size or *breadth* and quality or *depth*. However, in the majority of previous studies, only the measures of the breadth of VK were included. Second, according to models of L2 listening (e.g., Buck, 2001), *aural* VK and SK play a key role in L2 listening. However, in almost all previous studies, VK and SK were tested only in the *written* modality, which is a primary measure of *written* knowledge and does not provide direct evidence of *aural* VK and SK. The results of studies that included only written measures of the two constructs are incomplete and potentially misleading.

In addition, previous studies ignored the role of cognitive and affective variables when investigating the role of linguistic variables. For example, metacognitive knowledge (MK) and working memory (WM) are the two important cognitive variables known to affect L2 listening ability. L2 listening anxiety, an affective variable, which is a situation-specific type of anxiety, has also been shown to influence the performance of L2 listeners (Bloomfield et al., 2010). To investigate the unique variance explained by linguistic variables, the influence of important cognitive and affective variables should be accounted for. Finally, zero-order correlational and regression data analysis methods conducted at measured-variable level (in contrast to latent-variable level) used in previous studies have led to results contaminated with the impact of measurement error. Measurement error is the difference between a measured quantity and its true value. It includes both random error (error to be expected with any measurement) and systematic error (caused by an unreliable instrument). Thus, more complex statistical modeling methods, such as structural equation modeling (SEM), are useful to obtain results that control for the effect of measurement error.

More fine-grained analysis of the variables contributing to L2 listening ability conducted in the current study is essential for enhancing our theoretical understanding of the role and empirical weight of the underlying components of the construct. Additionally, the current study's findings may have practical implications for the pedagogy of

L2 listening. L2 teachers need an informed grasp of the variables and development of L2 listening ability to adequately serve their students. Traditionally, classroom listening pedagogy has followed a testing model in which, after performing listening tasks, answers are given, with little feedback on how to improve listening skills (Field, 2008). This limitation in pedagogy may be related to the fact that many classroom listening practices are influenced by large-scale proficiency testing. With teachers' primary focus on proficiency outcome tests, there is little opportunity for the development and improvement of underlying factors contributing to listening ability. It is also important to note that VK and SK are two linguistic components that can be explicitly taught and practiced in classroom settings, and that teachers can focus on the development of their students in these two areas to help them improve their L2 listening ability.

To bridge the gap in the literature, the current study examined the role of *aural* VK (both its *breadth* and *depth*) and SK in L2 listening, while the influence of MK, WM, and L2 listening anxiety was accounted for, using SEM as the main method of data analysis. To the best of our knowledge, this is the first study that investigated the *concurrent* roles of all of these linguistic, cognitive and affective variables.

BACKGROUND

THE ROLE OF VOCABULARY KNOWLEDGE IN L2 LISTENING

Empirical studies have shown that VK plays a central role in successful listening comprehension among listeners at different levels of ability (e.g., Mecartty, 2000; Milton, 2010; Milton, Wade, & Hopkins, 2010; Stæhr, 2008, 2009; Vandergrift & Baker, 2015). More specifically, aural word *recognition* skills are crucial in effective use of bottom-up processing, which allows listeners enough available WM capacity for further processing of the input (Vandergrift, 2004). Knowledge of word *meanings* has also been recognized as essential for deriving meaning, at least in part, from aural input (Rost, 1990).

Among the relatively recent studies, for example, a study by Stæhr (2008) reported a substantial correlation (around .70) between breadth of VK and listening comprehension among advanced EFL learners. As a follow-up study, Stæhr (2009) simultaneously investigated the role of *breadth* and *depth* of VK in L2 listening ability. Advanced EFL learners took the Cambridge Certificate of Proficiency in English (CPE) as a listening comprehension test, along with the *written* measures of breadth and depth of VK. The results revealed that both breadth and depth of VK were significantly correlated with listening test scores (.70 and .65, respectively). Milton et al. (2010) investigated the relationship between breadth of VK and L2 reading, listening, speaking, and writing abilities as measured by the IELTS test among intermediate and advanced level EFL learners. They also examined whether measuring the size of VK in two separate modalities, written and aural, can better explain performance in these four skills. X-Lex (Meara & Milton, 2003), a written, and A-Lex (Milton & Hopkins, 2006), an aural, VK size tests were used, and correlation analysis revealed that scores from the written VK test significantly correlated with all skills except for speaking. Among reading, writing, and listening, the latter had the smallest correlation (.48) with the written VK test scores. However, the scores from the aural VK test did not correlate significantly with the reading skill and had a strong significant correlation (.67) with listening. This suggests that *aural* VK is more directly related to listening ability test scores than written VK. In a similar vein, a more recent study by Vandergrift and Baker (2015) demonstrated a significant correlation between the size of L2 VK and the listening comprehension ability (.51).

While Milton et al. (2010) and Vandergrift and Baker (2015) included an *aural* measure of vocabulary breadth, and Stæhr (2009) included written measures of both breadth and depth of VK, none included *aural* measures of both *breadth* and *depth* of VK. In addition, their analysis was limited to zero-order correlational analysis, by which magnitude of associations are often attenuated due to measurement error. Thus, studies including *aural* measures of both *breadth* and *depth* of VK applying statistical analysis that accounts for measurement error (i.e., SEM) are needed to further illuminate the role of VK in L2 listening ability.

THE ROLE OF SYNTACTIC KNOWLEDGE IN L2 LISTENING

Compared to other variables, the role of SK in L2 listening is the least examined (Goh, 2005). One reason for this underrepresentation in the literature could be theoretical hypotheses about its minimal role. For example, Field (2008) claimed that L2 listeners appear to be more successful in processing and identifying *content* words than *function* words, and their relative success at identifying content words is not surprising because content words carry meaning. Vandergrift (2011) also argued that due to WM limitations, L2 listeners need to attend to aural input selectively by focusing on content words and ignore function words to optimize L2 listening comprehension.

Despite these arguments, it seems intuitively and theoretically plausible to assume that SK plays a major role in L2 listening. Contrary to the claims of Field (2008), the role of SK in L2 listening has been included both in componential (e.g., Buck, 2001) and process-oriented (e.g., Field, 2013) models of L2 listening. More specifically, SK plays a crucial role in segmenting and interpreting streams of speech (Call, 1985) and in predicting what they will hear next while listening and processing spoken utterances (Rost, 1990).

However, to date, very few empirical studies have directly investigated the role of SK in L2 listening. In the most frequently cited study on this topic, Mecartty (2000) compared the roles of VK and SK in L2 reading and listening. The participants were 154 upper-beginner learners of Spanish enrolled in a four-semester Spanish program in the United States. Results showed that while scores on the *written* measures of VK and SK correlated significantly with scores on the listening comprehension test, a hierarchical regression analysis demonstrated that VK remained the only significant predictor.

Mecartty's (2000) study is valuable insofar as it is among the few studies reporting the role of SK in L2 listening. However, its results are not free from critical limitations. For measuring SK, the researcher used a sentence-completion multiple-choice task and a grammaticality judgment task (GJT), both in *written* modality. However, similarly to VK, theoretical models of L2 listening (e.g., Buck, 2001; Goh, 2005) consider *aural* SK as one of the building blocks of listening ability. Thus, due to the written nature of the SK measures in Mecartty's study, it is unsurprising to find no significant predictive power of SK in L2 listening. Another limitation of this study is that Mecartty used a SK test that included *morphosyntactic* target structures such as articles, verb endings, nouns, adjectives, and adverbs. Inaccurate or incomplete knowledge of these morphosyntactic, not syntactic, target structures may or may not lead to misinterpretation of aural input, depending on whether an extended context is available. Although testing these target structures may reveal participants' overall level of morphosyntactic knowledge, there may not necessarily be a direct relationship between knowledge of these structures and L2 listening comprehension. Thus, studies including *aural* measures of SK developed using *syntactic* target structures that their inaccurate and/or incomplete processing can lead to miscomprehension of the aural input are required.

Finally, Andringa et al. (2012) investigated individual differences in listening ability, and they tested their participants (both native and nonnative speakers) on several linguistic and cognitive skills. They included measures of grammatical processing in their study; however, they did not examine the independent role of SK in relation to listening ability. In their SEM analysis, they grouped the linguistic and cognitive test scores under separate factors and, in their retained model, they combined the scores from grammar and vocabulary tests *together* as indicators of the language-related factors. This means, in their study, Andringa et al. did not examine the relative significance of different linguistic constructs. Rather, they compared the role of linguistic factors against cognitive factors (see Hulstijn, 2015 for the underlying theoretical framework in Andringa et al.'s study).

THE ROLE OF METACOGNITIVE KNOWLEDGE IN L2 LISTENING

Research on L2 listening ability initially examined the use of strategies by learners during aural comprehension tasks (Rubin, 1994). This led to extensive research on the role of MK in L2 listening and performance. MK is defined as "listener awareness of the cognitive processes involved in comprehension, and the capacity to oversee, regulate, and direct these processes" (Vandergrift & Baker, 2015, p. 395). MK involves both self-reflection and self-direction, and enables listeners to avoid unsuccessful comprehension strategies (Bloomfield et al., 2010; Vandergrift & Goh, 2012). Vandergrift (2006) showed that self-reported MK explained about 13% of the variance in L2 listening performance in a study of university-level language learners. This finding was confirmed by results from the comprehensive study by Vandergrift and Baker (2015). Therefore, when examining the role of linguistic factors in L2 listening, MK is one of the main nonlinguistic factors for which its role should be accounted.

THE ROLE OF WORKING MEMORY IN L2 LISTENING ABILITY

Theories of comprehension postulate that the ability to temporarily store and process information in WM underlies language comprehension (e.g., Andringa et al., 2012). According to Daneman and Merikle (1996), listening comprehension involves the ability to compute semantic and syntactic relations among successive words, phrases, and sentences, and the results of these computations enable listeners to construct a meaningful and coherent interpretation of the aural input. For integrating newly encountered information with the previously processed information, listeners must have access to the results of earlier processes; otherwise, they will fail to comprehend. Because WM is the cognitive system with oversight of processing, storage, and retrieval of information

(Baddeley, 2012), it should be an important source of individual differences in listening comprehension ability (Daneman & Merikle, 1996).

According to the most widely used model in L2 research (Baddeley, 2003), WM composed of four components: two modality-specific input/storage subsystems, the central executive, and the episodic buffer.[1] The two input/storage subsystems are the phonological loop and the visuo-spatial sketchpad, in which verbal information and visual/spatial information are processed, respectively. The central executive is an attentional control system coordinating the two input /storage subsystems by focusing attention, dividing it between two important targets or stimulus streams, and switching between tasks. It plays the key role in efficiently managing available attentional resources, rendering it the primary determiner of individual differences in WM capacity (Engle, 2002).

However, other WM models (e.g., Conway & Engle, 1994; Cowan, 2008; Engle, 2002) do not presuppose *modality-specific* (i.e., verbal vs. visual) input/storage subsystems and emphasize executive functions and processes of WM over its structure and content (Kane, Conway, Hambrick, & Engle, 2007). In other words, it is argued that WM capacity is not modality specific but a *domain-general* cognitive system. Evidence from large-scale empirical studies (e.g., Kane et al., 2004) supports this claim. In this view of WM, there is a strong emphasis on the role of executive functions in managing available attentional resources to resolve conflict or ambiguity in, for example, interpreting the so-called garden-path sentences (Bloomfield et al., 2010).

In terms of measurement, the executive function is typically assessed by so-called *complex* WM measures (Linck, Osthus, Koeth, & Bunting, 2014). These measures require active processing of input while remembering a string of letters, words, digits, or objects (Linck at al., 2014). Meta-analyses of the role of WM in both L1 (Daneman & Merikle, 1996) and L2 comprehension (Linck et al., 2014) have suggested that complex WM tasks (e.g., Operation Span Task) are better predictors of individual differences in WM capacity than *simple* tasks (e.g., forward/backward digit/word-span tasks), which only measure the ability to store and rehearse information.

Little research has examined the relationship between WM and L2 listening comprehension (Vandergrift & Baker, 2015). Among the few studies conducted to date, Brunfaut and Révész (2011) examined the role of WM and listening anxiety in listening task difficulty. They found that WM did not significantly predict listening ability. This null effect of WM could be due to the simple WM measures (visual forward digit-span and backward digit-span tasks) and/or a short passage used in the listening test. When cognitive demands of a listening comprehension task are low, it could easily lead to null findings for the role of WM. Similarly, Andringa et al. (2012) found very low associations between WM and listening ability. In their study, even though five WM measures were used, they were all simple tasks, which might have attenuated the role of WM. Vandergrift and Baker (2015) also used a backward digit recall (BDR) task and a nonword list recall (NLR) task to measure WM capacity. They argued that BDR[2] taps both central executive functions and phonological loop capacity, while NLR taps the phonological loop only. Curiously, despite this argument, the researchers combined the scores of these two tasks to create a composite WM score. Similar to Andringa et al.'s (2012) findings, Vandergrift and Baker did not find a significant correlation between WM capacity and L2 listening scores.

Although theoretical accounts of WM predict a significant role for this cognitive factor in L2 listening, the existent empirical studies have failed to capture this phenomenon. These null findings may have resulted from the use of (a) only simple WM tasks, (b) simple listening comprehension tasks, and /or (c) data analysis methods at the measured-variable level (Andringa et al., 2012, notwithstanding).

INFLUENCE OF L2 LISTENING ANXIETY ON L2 LISTENING PERFORMANCE

In addition to the preceding measurement and data-analysis issues, the lack of accounting for the influence of anxiety on WM may have led to null findings about the role of WM in previous research. It has been argued that L2 anxiety has different facets; while some learners experience L2 anxiety in general (trait), others only report feeling anxious while performing in a specific skill such as listening. In the current study, L2 anxiety is defined as the latter, *situation-specific* anxiety, meaning "the feeling of tension and apprehension specifically associated with second language context, including speaking, listening and learning" (MacIntyre & Gardner, 1994, p. 284). According to Scarcella and Oxford (1992), *listening* L2 anxiety usually occurs when learners feel they are faced with a difficult and unfamiliar L2 listening task. L2 listening anxiety increases when listeners are under the false impression that to complete a listening task, usually in a testing situation, they must understand every single word they hear (Vogely, 1998). Although there is a consensus among scholars that L2 listening anxiety, a situation-specific anxiety, impedes L2 listening comprehension (e.g., Elkhafaifi, 2005), it remains unclear how anxiety, which presumably interacts with other factors such as WM, exerts influence on L2 listening comprehension process.

Substantial evidence from cognitive psychology has shown that anxiety interferes with both executive-function and storage aspects of WM among English-speaking native speakers. According to Eysenck, Derakshan, Santos, and Calvo (2007), anxiety exerts its negative effect on performance of complex cognitive tasks by preempting the processing and temporary storage capacity of WM. Eysenck and Calvo (1992) argued that the main effects of anxiety are on central executive or attentional control mechanisms in WM. Eysenck et al. explained that anxiety redirects the allocation of inhibition and shifting mechanisms[3] of WM to threat-related stimuli, whether internal (e.g., worrisome thoughts) or external (e.g., threatening task-irrelevant distractors). As the result, performance on complex cognitive tasks demanding WM executive functions (e.g., inhibition and shifting functions in particular) become challenging when an individual experiences anxiety. This means that anxiety can *mask* individual differences in WM capacity, and it is not surprising that empirical studies with no control over L2 listening anxiety effects do not report a statistically significant role for WM in L2 listening. Thus, it seems plausible to control for the effect of anxiety on WM when the role of WM in L2 listening is investigated.

THE CURRENT STUDY

The purpose of the current study was to investigate the role and relative significance of VK and SK in L2 listening ability. The current study adopted an individual-differences approach (Cronbach, 1957), by which variation between individuals in their L2 listening ability and other relevant factors was investigated.

Unlike previous studies, measures of *aural* SK and VK, both its *breadth* (VB, hereafter) and *depth* (VD, hereafter) were included. Also, to clarify the role and relative significance of these two linguistic factors, the effects of two cognitive factors of MK and WM were accounted for. To obtain a clearer picture of the role of WM, the effect of L2 listening anxiety (AX, hereafter) on WM was also accounted for. SEM was used as the main method of data analysis to answer the following research questions:

1. Accounting for the role of SK, MK, WM, and AX, does VK make a significant contribution to L2 listening?
2. Accounting for the role of VK, MK, WM, and AX, does SK make a significant contribution to L2 listening?
3. Accounting for the role of MK, WM, and AX, what is the relative significance of VK and SK in explaining success in L2 listening ability?

To answer the preceding questions, the main SEM model tested in the current study is depicted in Figure 1.

PARTICIPANTS

The current study recruited 263 (Female = 162 and Male = 101) English learners from lower-intermediate to advanced levels of proficiency enrolled in communicative EFL classes in Iran. English classes in the formal education system of Iran (both secondary and higher education) are based on the grammar-translation method and, for this reason, for gaining communicative competence in English; learners attend communicative EFL

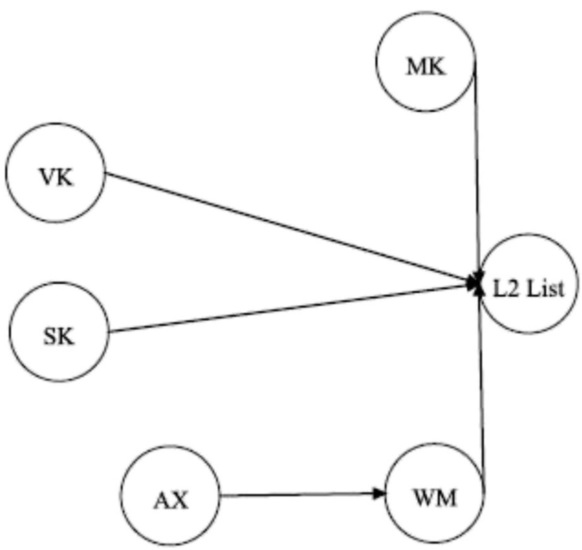

FIGURE 1. The main SEM model represented only at the latent variable level.

classes outside the formal education system. These typically two-hour classes, which usually meet three times a week, follow the communicative approach. The average experience of our participants attending the EFL classes (the ones outside the formal education system) was 3.8 years. The average age of the participants was 27.8 (SD = 5.3), and their minimum level of education was matriculating as an undergraduate student at a university. None of the participants reported to have had the experience of living, working, or studying in an English-speaking country. The listening subsection of the IELTS test was used as the measure of listening ability, and the participants' score indicated their listening skills varied sufficiently for analysis (median = 5, range = 2.5–9).

INSTRUMENTS

Listening Comprehension Test

The listening subsection of an institutional version of the IELTS test was used as a measure of EFL listening comprehension ability. The test was comprised of four parts administered in sequential order of difficulty. Each part included a recorded passage that was a conversation, a monologue, or a lecture by a range of English native speakers, and there were 10 questions for each passage (10 questions per part), with a total of 40 items in the test. Response format for the questions could be divided into two main types of selected response ($K = 16$) and limited production ($K = 24$). Questions tapped into a variety of listening abilities such as understanding main ideas and factual information; understanding the opinions and attitudes of speakers; understanding the purpose of an utterance; and following the development of ideas. Participants listened to each passage once and marked or wrote their answers directly on the question booklet, and the total time of the test was 45 minutes.

Vocabulary Knowledge Measures

In the current study, both VB and VD were assessed. VB was operationalized as the ability to recognize words and match them with their definitions, and the VD was operationalized as the ability to recognize words and match them with their synonyms and collocations (Read, 1993).

Vocabulary Breadth Test

An aural vocabulary breadth test was developed using the written Vocabulary Size Test (VST) developed and validated by Nation and Beglar (2007) and Beglar (2010). The written test is in multiple-choice format, and the target word in each item is placed in a short, nondefining context. Here is a sample item:

Miniature: It is a miniature.

A. A very small thing of its kind
B. An instrument for looking at very small objects
C. A very small living creature
D. A small line to join letters in handwriting

Sudden							
Beautiful □□	Quick □□	Surprising □□	Thirsty □□	Change □□	Doctor □□	Noise □□	School □□

The written VST consisted of 10 items taken from the 1st to the 14th 1,000-word families of English, but we adopted its first 80 items to develop the new aural version because Nation (2006) asserted that for successful comprehension of English spoken texts, a vocabulary size of 6,000–7,000 word families is required. A female native speaker of American English was recruited to create the audio portions of the new aural test (see McLean et al., 2015 for the same approach to create an auditory version). The speaker read the word in each item and its relevant nondefining sentence.

To complete the VBT, participants listened to the words and their nondefining sentences once and read and marked the multiple-choice answers on an answer sheet. There was a 10-second pause[4] between each item, a length of time determined based on a pilot study to be sufficient to read the four options and mark the answer. The total time of the test was 20 minutes. See Appendix 1 in Online Supplementary Material for more VBT sample items.

Vocabulary Depth Test

An aural vocabulary depth test (VDT) was developed to assess the two major aspects of depth of VK: synonym/polysemy and collocation. The VDT was created by adopting a written Word Association Test (WAT), developed and validated by Qian and Schedl (2004). Here is a sample item from this test:

In a WAT, each item consisted of one adjective target word at the top ("sudden" in the preceding sample item) and two boxes below each consisting of four words. Among the four words in the left box, one to three could be synonymous either to one aspect or the full meaning of the target word. Among the four words in the right box, one to three words collocate with the target word. Each item always has four correct choices. To decrease the probability of getting correct answers at random, the correct choices were systematically distributed between the two boxes: (a) the left and right boxes both contain two correct answers; (b) the left box contains one correct choice, while the right box contains three; and (c) the left box contains three correct answers, while the right box contains one correct choice.

To develop the aural VDT, a female native speaker of American English was recorded reading the target words ($k = 40$). At the time of test administration, participants heard each target word once, then saw answer choices on an answer sheet and marked their responses. The time limit for each item was 20 seconds, which had been determined to be a sufficient length of time for processing aural information, reading answer choices, and marking answers. The total time of the test was 15 minutes. See Appendix 2 in Online Supplementary Material for more VDT sample items and instructions on how to the test.

TABLE 1. Syntactic target structures used in the present study

Target	Canonical	Noncanonical
Active/passive[a]	The man is pointing at the boy.	The boy is tapped by the girl.
Subject/object relative clauses	This is the boy that points at the man.	This is the boy that was pointed at by the man.
Passive object relative clauses		I read the book which was written by a Russian writer.
Hypothetical sentences (types two and three)	If I had money, I would travel.	If they were not tall, they could not play well.
Causative/noncausative	His friend helped him to move.	He got his friend to help him move.

[a]The examples in the first two rows of the table were directly borrowed from Robertson and Joanisse (2010, p. 161).

Syntactic Knowledge Measures

Two tasks were used to assess the learners' SK: an aural GJT and an aural sentence comprehension task (SCT). In both tasks, decontextualized sentences were presented aurally, and time pressure was imposed for the responses. Time pressure and aural modality were assumed to have participants recourse to *aural* SK that can be accessed quickly. Five target structures, with their canonical and noncanonical versions, were chosen for the current study (see Table 1). The current study focused on syntactic structures that potentially play a role in listening comprehension. Inaccurate processing of these target structures would potentially lead to misinterpretation of the message. The previous studies (e.g., Mecartty, 2000) that examined the role of SK in L2 listening focused on morphosyntactic target structures, such as articles, verb endings, nouns, adjectives, and adverbs. Testing these target structures may reveal participants' overall level of morphosyntactic knowledge; however, there may not be a direct relationship between knowledge of these structures and L2 listening comprehension. Inaccurate or incomplete knowledge of the target structures may or may not lead to misinterpretation of aural input, depending on whether an extended context is available.

The idea for choosing the target structures in the current study emerged from two lines of research: (a) cognitive psychology studies regarding factors contributing to aural sentence comprehension (e.g., Ferreira, 2003; Reali & Christiansen, 2007; Robertson & Joanisse, 2010); and (b) corpus studies comparing and contrasting the frequency of occurrence of important syntactic structures in written and spoken corpora (Biber et al., 1998; Roland et al., 2007).

Sentence comprehension studies in cognitive psychology have primarily focused on comparing and contrasting comprehension of canonical and noncanonical sentences. These studies have examined the difficulty of comprehending active and passive sentences, subject and object relative clauses (Robertson & Joanisse, 2010), conditional clauses (Haigh & Stewart, 2011), and causative sentences (Dittmar et al., 2008). The reason for targeting these structures is their presumed importance in successful aural comprehension.

Corpus studies (both written and spoken) comparing and contrasting variation of key syntactic structures have typically focused on three elements: (a) frequency of occurrence of different types of relative and conditional clauses; (b) causative sentences; and (c) active and passive sentences. Following these two lines of research, five target structures, with their canonical and noncanonical versions, were chosen for the current study. Table 1 contains the target structures and relevant examples.

To avoid confounding SK with VK, the frequency of words used in both GJT and SCT sentences were checked with corpus data, with only words from the first 1,000-word families used.

Aural Grammaticality Judgment Task

Each auditory sentence was presented once with participants instructed to mark answers as correct or incorrect on a paper answer sheet. Here are two sample sentences for the grammatical and ungrammatical sentences for the active/passive target structures:

The boy is tapped by the girl.

* The boy is tapping by the girl.

The time limit for each item, including the presentation of the stimulus, was 8 seconds, a length of time determined by a pilot study. There were four correct and four incorrect sentences for each of the nine target structures (see Table 1). Therefore, the test had a total of 72 items. Each sentence was recorded at a normal speed by a native speaker of American English.

See Appendix 3 in Online Supplementary Material for more GJT sample items.

Aural Sentence Comprehension Task

Participants listened to a sentence once and answered a short yes/no comprehension question. The questions were designed to tap accuracy of comprehension based on the target structures. Here is a sample item from the test:

Sentence (aural): If she were not rich, she could not travel.

Question (written): Can she travel now?

The time limit for each item, including the presentation of the stimulus, was 11 seconds, a length of time determined by a pilot study. There were six items for each of the nine structures, for a total of 54 items. Each sentence was recorded at a normal speed by a native speaker of American English. See Appendix 4 in Online Supplementary Material for more SCT sample items.

Metacognitive Knowledge Measure

MK of L2 listening was measured using the Metacognitive Awareness Listening Questionnaire (MALQ), developed and validated by Vandergrift (2006). The MALQ is a self-report instrument consisting of 21 randomly ordered items tapping MK related to L2 listening comprehension. Items measure perceived use of strategies and processes underlying five factors related to the regulation of L2 listening comprehension: Problem-Solving (PS), Planning and Evaluation (PE), Mental Translation (MT), Person Knowledge (PK), and Directed Attention (DA). Participants respond to the MALQ using a Likert scale from 1 to 6, with 6 signifying full agreement with the statement in the item. To avoid confounding MK with EFL proficiency, the items in the MALQ were translated into Persian.[5]

There was no time limit for completing the questionnaire, but, on average, it took about 10 minutes for participants to complete this task. The MK questionnaire had several negatively worded items, so responses were inverted to have positive correlations with the listening ability measures. See Appendix 5 in Online Supplementary Material for the complete set of English and Persian MALQ[6] items.

WORKING MEMORY MEASURES

As explained earlier, if the role of WM in complex cognitive tasks like L2 listening is to be examined effectively, complex measures tapping both processing and storage aspects of WM need to be employed. For this reason, two complex WM tasks—Blockspan (WM 1) and Shapebuilder (WM 2) (Atkins et al., 2012)—were used in the current study. Both are spatial tasks, requiring participants to simultaneously store and process spatial information. These visual complex WM tasks were chosen because nonverbal measures allow for avoiding confounding WM with a verbal construct such as L2 proficiency. Also, as discussed before, more recent models (e.g., Cowan, 2008) consider WM capacity domain general (modality general) and not domain specific. See Appendix 6 in Online Supplementary Material for descriptions of these two tasks.

Anxiety Questionnaires

For measuring AX, the current study employed two questionnaires that were used in ~~two~~ previous studies on the role of AX on L2 listening performance. The first one (AX 1) ($K = 9$) was adopted from Mills, Pajares, and Herron (2006), and the second one (AX 2) ($K = 18$) was adopted from Elkhafaifi (2005). There was no time limit for completing the questionnaires, but on average it took about 10 minutes for participants to complete them. Again, to avoid confounding AX with EFL proficiency, the items in the two questionnaires were translated into Persian. See Appendix 7 in Online Supplementary Material for the English and Persian versions of the AX 1 and AX 2.

PROCEDURE

Table 2 summarizes the timing and order of administration of the measures. Participants joined two sessions (group and individual ones) on the same day. A room equipped with high-quality audio devices was used for group-level administration of the measures. Administrations of group measures were limited to 10 participants at a time. Group-level administration took about 130 minutes, including all measures except for the WM tasks. There were two 10-minute breaks during the group administration of the measures. The WM tasks were administered in the individual session in the lab. This session lasted for about 30 minutes for each participant.

TABLE 2. Order and timing of task administration

	Group Session	
Order	Task	Mins.
1	Consent Form & Background Questionnaire	10
2	IELTS listening test	45
3	MK Questionnaire	10
4	AX 1 and AX 2 Questionnaires	10
	Break 1	**10**
5	VBT	20
6	VDT	15
	Break 2	**10**
7	SCT	10
8	GJT	10
	Break 3	**20**
	Individual Session	
9	WM 1	15
10	WM 2	15
	Total Time	200

Individual sessions were arranged after a 20-minute break following the group-level administrations. To promote complete understanding, the instructions were given in Persian for all the tasks.

DATA ANALYSIS

All linguistic measures were scored dichotomously, with data recorded as zero and one. For the IELTS listening measure (K = 40), scores were recorded in two ways: the whole test (including all 40 items), and the four separate parts (coded IELTS 1–4) with 10 items in each part. For VK measures, scores from the VBT (K = 80) and VDT (K = 160) were recorded separately and combined (K = 240). For SK measures also, scores from the GJT (K = 72) and SCT (K = 54) were recorded separately and combined (K= 126). Responses to the MK questionnaire were recorded as numbers ranging from one to six for all items (K = 21), as well as subsections: MK-PE (K = 5), MK-DA (K = 4), MK-PK (K = 3), MK-MT (K = 3), and MK-PS (K = 6). ~~The MK questionnaire had several negatively worded items, so responses were inverted to have positive correlations with the listening ability measures.~~ Responses to the two anxiety questionnaires, AX 1 (K = 9) and AX 2 (K= 18), were recorded as numbers in a range of zero to seven and one to five, respectively. The scores from the two questionnaires were also combined to create a composite score for the construct of AX (K = 27). Total scores from WM 1 and WM 2 were automatically calculated by the computer. Scores from these two measures were also combined to create a composite score for the construct of WM.

Before conducting the main analyses, a set of preliminary analyses were conducted. First, Rasch item and reliability analyses were conducted. Rasch item analysis showed that the infit mean-square of several items in the battery was outside the acceptable range of .75 to 1.3 (Bond & Fox, 2001). Therefore, before reliability analyses, misfitting items were deleted. The number of deleted items from each measure was as follows: one from IELTS 4, GJT, SCT, MK_DA, MK_PE, and MK_PK; two from VDT; three from AX1 and AX2; and eight from VBT. Reliability of the measures were then estimated after deleting the misfitting items. Rasch person ability logits (a logit is a unit of additive measurement on an interval scale) for all measures, except WM tasks, were generated to be used as data for the subsequent analyses.

Additionally, the infit mean-square of Rasch person ability logits was evaluated to detect misfitting persons (cases). By considering the acceptable range of .75 to 1.3, several misfitting cases were found in the dataset. However, for all measures, the number of misfitting cases was less than 5% of all participants. The maximum number of misfitting cases was found in VBT with 12 misfitting cases. As the number of misfitting cases for none of the measures exceeded 5% of the total number of cases, and because a few misfitting cases in a dataset should not be a point of concern (i.e., they have negligible impact on anything else) (Wright, Linacre, Gustafson, & Martin-Lof, 1994), it was decided to keep data from all participants for the subsequent analyses.

Descriptive statistics were computed, and assumptions of univariate and multivariate normality checked. Although the main method of data analysis in the current study was SEM, to explore the relationships between the *measured* variables of the current study, Pearson Product-Moment Correlation coefficients were also computed. Finally, a set of SEM analyses were conducted to answer the research questions. SEM analyses were implemented through the Mplus program (version 7.0).

To evaluate the SEM models, a profile of model fit tests and indices recommended by Hu and Bentler (1999) and Mueller and Hancock (2008) was used. Chi square, with its degrees of freedom and *p*-value, was checked. For a good model fit, the chi-square should not be statistically significant at a .05 level. However, in large samples and complex models, a chi-square is usually significant and not very informative. For this reason, the following descriptive fit indices were also used: the standardized root mean square residual (SRMR < .08), the root mean square error of approximation (RMSEA < .06), the comparative fit index (CFI > .95), the normal fit index (NFI > .90), the nonnormed fit index (NNFI > .95), and the goodness-of-fit statistic (GFI > .90).

RESULTS

PRELIMINARY ANALYSES

Table 3 summarizes the descriptive statistics and reliability estimates.

Skewness and kurtosis of all measures were within the acceptable range of +/–3 (Tabachnick & Fidell, 2007). In addition to examining skewness and kurtosis, Kolmogorov–Smirnov and Shapiro–Wilk statistical tests were conducted to examine whether the assumption of univariate normality of the data was met. These tests, along with a visual examination of histograms, revealed that this assumption was indeed met.

The results of Mardia's Multivariate Normality Test (Chi-square Skewness = 394.79, *p*-value = .000; z-score Kurtosis= 1.07, *p*-value = .28), however, revealed that the assumption of multivariate normality was not met. For this reason, for SEM, Robust Maximum Likelihood (RML) method of estimation was used. Because chi-squares from

TABLE 3. Descriptive statistics

	K	Mean	SD	Min	Max	Skewness	Kurtosis	Reliability
IELTS 1	10	6.31	2.54	1	10	–.4	–.67	.75
IELTS 2	10	4.76	2.63	1	10	.22	–.67	.76
IELTS 3	10	4.04	2.53	1	10	.67	–.27	.72
IELTS 4	9	4.74	2.87	1	10	.25	–.97	.77
IELTS Total	39	19.86	8.93	4	39	.43	–.66	.90
VBT	72	35.71	12.4	11	72	.76	.06	.92
VDT	158	77.35	24.78	14	149	.54	–.08	.97
VK Total	230	113.06	36.12	25	221	.67	.03	.98
SCT	53	40.25	9.08	13	53	–.69	.03	.91
GJT	71	40.4	11.75	7	71	.3	.39	.90
SK Total	124	80.66	19.41	25	124	.02	–.3	.95
MK_PE	4	17.12	4.9	5	28	–.06	–.48	.77
MK_DA	3	16.34	4.11	5	24	–.34	–.43	.77
MK_PK	2	10.11	4.37	3	18	.06	–1.1	.84
MK_MT	3	13.06	4.02	3	18	–.56	–.43	.61
MK_PS	6	24.42	5.52	10	36	–.29	–.55	.68
MK Total	18	81.05	13.22	51	114	.05	–.42	.75
AX 1	6	37.31	14.94	0	63	–.19	–.83	.91
AX 2	15	54.92	14.31	23	89	.08	–.44	.93
AX Total	21	92.23	28.16	32	151	–.09	–.75	.95
WM 1	N/A	12.05	4.03	1	27	–.19	.9	.73
WM 2	N/A	10.79	3.86	2	28	.82	2.09	.81
WM Total	N/A	22.72	6.59	6.65	43.6	.02	.33	.79

RML method cannot be used similarly to chi-squares from Maximum Likelihood method for nested-model comparison, chi-square difference test of nested models using the *scaled chi-square*[7] was computed (Satorra & Bentler, 2010).

Table 4 summarizes the Pearson Product-Moment Correlation coefficients, and all the measures had a significant correlation with the listening ability test scores.

STRUCTURAL EQUATION MODELING

In the SEM analysis, the main endogenous factor (dependent latent variable) was L2 listening ability (L2 List, for short) which was regressed on the exogenous factors (independent latent variables) of VK, SK, MK, and WM. Also, WM factor was regressed on the exogenous factor of AX. Figure 2 (Model 1) shows the measured variables loading on each of these endogenous and exogenous latent variables. In an earlier model, it was observed that MK-MT and MK-PS were not loading on the MK factor in a statistically significant way, so they were removed from Model 1.

To examine the role and the relative significance of VK and SK in explaining success in L2 listening comprehension, Model 1 was evaluated. The fit indices revealed that Model 1 fit the data well, as Table 5 indicates.

In Model 1, all regression paths between measured variables and their latent factors, as well as the ones between exogenous factors and the L2 listening latent factor, were

TABLE 4. Correlations among the total scores of the measured variables AU2

Task	IELTS	VBT	VDT	GJT	SCT	MK	WM	AX
IELTS	1							
VBT	.78**	1						
VDT	.8**	.87**	1					
GJT	.72**	.74**	.73**	1				
SCT	.73**	.7**	.66**	.75**	1			
MK	.37**	.44**	.36**	.4**	.39**	1		
WM	.31**	.14*	.2**	.22**	.27**	.14*	1	
AX	–.52**	–.47**	–.43**	–.47**	–.48**	–.56**	–.19**	1

statistically significant. The regression path between the AX factor to the WM had a negative value and was also statistically significant.

To examine whether each of the exogenous factors in Model 1 explained a significant amount of variance in the L2 listening latent variable, four alternative models to Model 1 were tested. In each of these alternative models, the structural regression path from an exogenous factor was constrained as carrying a zero regression weight. The scaled chi-square test was employed to evaluate any difference in fit between each of these more constrained models and Model 1. If constraining the regression weight for an independent variable at zero does not lead to deterioration of model fit, that independent variable adds little or nothing to the explanation of individual differences in L2 listening ability.

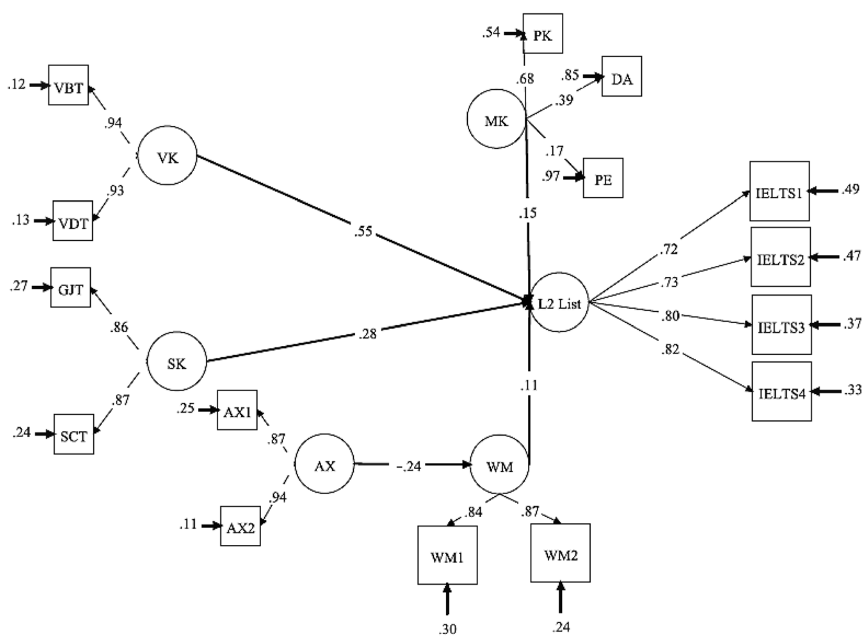

FIGURE 2. SEM Model 1.

TABLE 5. Summary of fit indices for SEM Model 1

Index	CFI	NFI	NNFI	GFI	RMSEA	SRMR	Chi-square
Criterion	≥ .95	≥ .90	≥ .95	≥ .90	≤ .06	≤ .08	None significant
Model 1	.98	.96	.97	.92	.07	.06	*$\chi 2$ = 189.65 , *df* = 79

The regression weights for VK, SK, MK, and WM were, respectively, set to zero in Models 2 to 5. The alternative models to Model 1 all had larger chi-squares, and the results of the scaled chi-square difference tests between these alternative models and Model 1 were statistically significant. This means that each of these exogenous factors significantly added to the explanation of individual differences in L2 listening ability. For this reason, all independent variables were kept, and Model 1 was used for examining the relative significance of VK and SK in L2 listening.

The regression weights (beta values or standardized loadings) of VK and SK were compared. These values in Model 1 for VK and SK were .55 and .28, respectively. Both beta values were statistically significant. This shows that VK accounted for individual differences in L2 listening better than SK.

Then, we examined whether the larger effect of VK than SK (in terms of effect size) is statistically significant. An alternative SEM model to Model 1 was tested. In Model 6, the regression paths of VK and SK to L2 listening ability were set to be equal. This means that in Model 6, the significance of VK and SK in explaining success in L2 listening ability was considered equal. If the chi-square of Model 6 increased in comparison to Model 1, and if the scaled chi-square difference test between the two models were significant, it would mean that the effect size difference between VK and SK were also statistically significant. However, if the scaled chi-square difference test between Models 1 and 6 were nonsignificant, there were no inferential statistical evidence to conclude that VK plays a more significant role than SK in explaining success in L2 listening ability. After setting the regression paths of VK and SK to the L2 listening ability latent factor as equal, the regression weights for both independent factors were estimated to be .43 and statistically significant.

Table 6 summarizes fit indices for Model 6. In terms of fit indices, Models 1 and 6 are almost identical; however, the chi-square for Model 6 is slightly larger.

A scaled chi-square difference test between Models 1 and 6 was then conducted, and the results showed that these two models were not statistically significantly different from each other. These results suggest that although VK plays a more important role than

TABLE 6. Summary of fit indices for SEM Model 6

Index	CFI	NFI	NNFI	GFI	RMSEA	SRMR	Chi-square
Criterion	≥ .95	≥ .90	≥ .95	≥ .90	≤ .06	≤ .08	None significant
Model 1	.98	.96	.97	.91	.07	.06	*$\chi 2$ = 191.56 , df = 80

SK in explaining success in L2 listening, as shown by the effect sizes, this difference is not statistically significant.[8]

DISCUSSIONS

The purpose of the current study was to examine the role of VK and SK in L2 listening ability, as well as their relative significance. The role of these two linguistic factors has been recognized in theoretical models of L2 listening; however, limited empirical evidence is available to support these theoretical accounts. Unlike previous studies, the current project employed assessment tasks to measure *aural* VK and SK. In terms of VK, to avoid underrepresenting the construct, measures of both *breadth* (VB) and *depth* (VD) were included.

Additionally, the current study examined the role of VK and SK by accounting for individual differences in two important cognitive factors of L2 listening, that is, MK and WM. Also, to reveal the significant role of VK and SK more fully, the current study accounted for the negative impact of AX on WM, and in turn, on L2 listening.

The results of testing SEM Model 1 showed that the regression path between the MK independent factor and L2 listening was .15 and statistically significant. According to this finding, it can be concluded that MK is a significant variable of L2 listening success, and it was justifiable to include it in the SEM models as a control variable. This finding is in line with findings from previous studies (e.g., Vandergrift & Baker, 2015) that found a significant role for MK in L2 listening ability.

The results of SEM Model 1 also showed that the regression path from the WM independent factor to the L2 listening factor had a standardized loading of .11 and statistically significant. This finding stands in contrast to what Andringa et al. (2012) and Vandergrift and Baker (2015) found. In both of these two studies, although WM had a significant association with L2 listening, it was not a significant predictor. Aside from other methodological shortcomings, the nonsignificant results for the role WM in Andringa et al. (2012) and Vandergrift and Baker (2015) may have resulted from them not accounting for anxiety effects on WM. Studies in cognitive psychology (e.g., Eysenck et al., 2007) have shown the detrimental effect of anxiety on both executive functions and storage aspects of WM. Thus, anxiety can mask individual differences in WM capacity, resulting naturally in nonsignificant results. If individuals with relatively larger WM capacity experience anxiety during an L2 listening comprehension task, they cannot make an effective use of that capacity. However, individuals with relatively smaller WM capacity who do not experience anxiety can make the fullest use of their WM capacity.

This claim was supported by the findings from SEM Model 1 that showed the regression path from the anxiety independent factor to the WM factor had a standardized loading of −.24 that was also statistically significant. These results led us to conclude that WM is a significant variable of L2 listening, and it is justifiable to be included in the main analysis as a control variable. Also, because results showed that the role of WM in L2 listening can be negatively influenced by the effects of anxiety to an extent that its role may not be observable, in SEM models in the current study, the WM independent factor was regressed on the anxiety factor prior to the L2 listening factor being regressed on the WM independent factor.

With the preceding preliminary considerations, the answers to the research questions are in order. The first research question was:

Accounting for the role of SK, MK, WM, and AX, does VK make a significant contribution to L2 listening?

SEM Model 1 was examined to answer this question. In this model, the magnitude of the standardized regression path between the VK factor and the L2 listening factor was .55 and statistically significant. These results led us to conclude that VK is a strong and significant variable of L2 listening ability. These results are also supported by findings in previous studies (e.g., Mecartty, 2000; Staehr, 2009; Vandergrift & Baker, 2015) showing that VK plays a critical role in understating aural input. For example, Stæhr (2009) had found correlations of .7 and .65 between listening comprehension and breadth and depth of VK, respectively.

The important role of VK in L2 listening has similarly been acknowledged in componential theoretical models of L2 listening (Buck, 2001), and can be explained within process-oriented models (Field, 2013) of listening comprehension. According to these models, for successful listening comprehension to take place, L2 listeners rely on different knowledge sources through top-down and bottom-up processes. Successful listening comprehension is then the result of a complex interaction between these two types of processes (Vandergrift, 2007). Aural VK facilitates bottom-up processing of aural input, allowing L2 learners to use top-down processing for successful L2 listening comprehension (Field, 2013).

The second research question was:

Accounting for the role of VK, MK, WM, and AX, does SK make a significant contribution to L2 listening?

SEM Model 1 was also examined to answer this question. In this model, the magnitude of the standardized regression path between the SK independent factor and the L2 listening factor was .28 and statistically significant. The comparison of the standardized regression paths in SEM Model 1 revealed that SK was the second-strongest variable of L2 listening ability (after VK) among all other variables. These results stand in contrast with the findings of Mecartty (2000). In this study, although measures of both VK and SK correlated with the measure of L2 listening, only VK measure significantly explained success in L2 listening scores. As previously discussed, the lack of predictive power of SK in his study could be due to the methodological shortcomings (e.g., *written* SK tests were used). The *aural* SK tests utilized in the current study probably contributed more directly to successful L2 listening comprehension and were more compatible with L2 listening comprehension models (e.g., Buck, 2001; Field, 2013). Another novel aspect of the current study was that we developed SK tests that solely focused on English syntax for which difficulty in processing would lead to miscomprehension.

The significant role for SK in predicting success in L2 listening found in the current study can be explained through theoretical accounts, both in componential and processing-oriented models. For example, in Field's (2013) model, listening comprehension is broken down into the following five stages: input decoding, lexical search, parsing (low-level or bottom-up processes), meaning construction, and discourse representation (high-level or top-down processes). The *parsing* stage is most relevant for the role of SK.

Field (2013) defined parsing as processing of the establishment of relationships between the meaning of individual words and whole utterances and draws upon stored linguistic knowledge, similarly to other low-level or bottom-up L2 listening processes (i.e., input decoding and lexical search). While input decoding and lexical search primarily rely on phonological and lexical knowledge, parsing heavily draws on SK. According to Field (2013), the signature of L2 listening comprehension process is (real-time) incremental linguistic processing, which means listeners must constantly create hypotheses about what they will hear next. These hypotheses can be at the level of the phoneme, word, or whole utterance. To formulate hypotheses at the whole utterance level, listeners must impose a syntactic pattern on the auditory input, which can be done only in a piecemeal fashion because aural input unfolds in real time and fleets away quickly (Field, 2013).

In sum, for successful parsing to take place, listeners match provisional hypotheses about aural input at the utterance level against subsequent evidence. Aural SK can assist L2 listeners to more accurately anticipate upcoming syntactic structures in aural input. In case of a mismatch between the hypothesis and the aural evidence, aural SK can assist L2 listeners to quickly formulate new hypotheses (Field, 2009).

The third research question was:

> Accounting for the role of MK, WM, and AX, what is the relative significance of VK and SK in explaining success in L2 listening ability?

To answer this question, in SEM Model 1, the relative significance of these two linguistic factors was examined primarily using the effect sizes. In this model, the standardized regression paths from VK and SK to the L2 listening factor were .55 and .28, respectively. Both were also statistically significant. Standardized regression weighs can be interpreted as effect sizes and according to Cohen's (1992) convention, .55 and .28 represent strong and moderate associations between variables, respectively. Therefore, in terms of effect sizes, VK played a more important role than SK in explaining success in L2 listening. This result was confirmed by the amount of unique variance VK and SK explained in the L2 listening factor. VK uniquely accounted for 6% of the total variance in the L2 listening factor, while SK uniquely accounted for 2%.

To examine whether VK statistically significantly explained L2 listening success better than SK, SEM Model 6 was tested. In this model, the regression paths of VK and SK to L2 listening ability were set to be equal, and the comparison between Models 1 and 6 revealed no statistically significant difference in the scaled chi-squares. This suggests that the difference between VK and SK factors was not statistically significant. However, this nonsignificant difference could be due to lack of statistical power. For example, a larger sample size or less complex SEM model may have provided more statistical power to estimate the true difference between the contributions of these two linguistic factors. For this reason, we find the information from the effect sizes more meaningful and useful for interpreting the results of the current study.

CONCLUSIONS

The central finding of the present study was that both VK and SK play a significant role in explaining success in L2 listening. Unlike previous studies, current results showed that, along with VK, SK is one of the key linguistic sources L2 learners draw on for effective parsing of aural input to create idea units and understand propositions in L2 listening comprehension tasks. Note that a significant role for SK in L2 listening was found even after accounting for the effect of the other pertinent factors (VK, MK, WM, and AX). Critically, these findings are demonstrated through SEM analyses that took all pertinent variables into account simultaneously.

As to the relative significance of VK and SK, the results in the current study led to the conclusion that VK plays a more important role than SK in L2 listening, at least, when the difference in the effect sizes of the two variables are considered. However, future studies with lager sample sizes and maybe more reliable measures may show that the differences between the role of VK and SK is also statistically significant.

Sufficient aural VK and SK promote efficient bottom-up processing, which in turn frees up sufficient WM space for top-down processing of aural input. As a result, L2 listeners can complete these tasks: access the contextual information of the aural input by paying attention to top-level cues; activate prior knowledge about the topic; effectively use cognitive and metacognitive strategies; and transfer L1 listening skills to the L2 listening task for more successful L2 listening comprehension.

It should also be noted that there is limited set of syntactic structures in each language and mastering these structures may take less time and effort compared to learning a vast number of words and other lexical units. Although not tested empirically, it can be logically assumed that SK plays a more significant role in L2 listening among lower proficiency learners. As learners' proficiency grows, compared to VK, individual-differences in SK diminish faster. For this reason, individual differences in VK remain as a more important factor in explaining individual differences in L2 listening ability even at higher levels of proficiency.

The relative significance of these two linguistic components is somewhat of a less important issue because both were found to play an important role in L2 listening success. SK and VK are highly correlated constructs; as such, teasing them apart completely in empirical studies is nearly impossible. Therefore, the important implication of the current findings is that *both* VK and SK should be improved both quantitatively and qualitatively for more successful L2 listening comprehension to occur. Here, by quantitatively we mean expanding the size of VK and knowing (about) how a variety of syntactic structures encode meanings, whereas by qualitatively, we mean deepening the VK (e.g., by learning more about the associations between a word and other words), and increasing the speed and efficiency by which both VK and SK can be deployed during listening comprehension.

In interpreting these conclusions, one important issue merits close attention. The design of the current study was cross-sectional and, for this reason, no one-way causal inference with regard to the relationship between L2 linguistic knowledge and L2 listening ability can be drawn. These constructs typically covary and are bidirectional. On the one hand, at least in the context of learning a new language as a second language (versus a foreign language), listening is the main channel of input for learning new vocabulary and syntactic structures, and frequent exposure to words and syntactic structures in aural input can lead to proceduralization and automatization of linguistic knowledge. On the other hand, developing VK and SK (whether or not through listening) leads to improved listening ability. Although current conclusions do not imply a one-way causal relationship between L2 linguistic knowledge and listening ability, they do not discourage such conclusions particularly in the context of current study.

The evidence for a significant contribution of both VK and SK to success in L2 listening has important pedagogical implications. Awareness of the importance of VK and SK in L2 listening ability is fundamental to a theoretically grounded pedagogy of listening comprehension. There is a general consensus in L2 listening research literature that listening instruction has favored the development of top-down processes by focusing its attention on the development of cognitive and metacognitive strategies at the expense of developing bottom-up processes (e.g., Field, 2001; Rost, 2002). It is important to note that the current study revealed that *aural* knowledge of L2 vocabulary and syntax significantly contributes to explaining success in L2 listening. Therefore, language instructors should bear in mind the aim of promoting aural VK and SK, as well as written one.

LIMITATIONS AND FUTURE RESEARCH

Several issues not addressed (adequately) in the current study warrant more in-depth futures investigations. For one, in the current study we used a single listening test (i.e., IELTS). The listening passages in this test have their own specific characteristics in terms of linguistic features, genera, and content information. Future research needs to examine the extent to which the linguistic characteristics of listening passages may lead to different findings in comparison to the current study. See Appendix 9 in Online Supplementary Material for information about several lexical and syntactic features of the listening passages used in the current study.

Next, one way that the current study tried to improve on previous studies was to use *aural* measures of VK. This way, we tried to measure VK more directly in relation to listening ability. However, we acknowledge that our aural VK measures were limited in different ways. We strongly concur with the recent call for more rigorous development and validation of VK tests in general (Schmitt, Nation, & Kremmel, 2019). Future studies should devise more valid ways for measuring aural VK knowledge.

For the BVK test, the options in our BVK test were in English rather than Persian, which was the first language of the test takers. This makes making inferences about the knowledge of the test takers of the target words more difficult because it is not clear whether failure on an item was due to lack of knowledge of the target word or not knowing the words in the options. The context in which the target words were embedded was limited; testing the target words in a more extended context may be beneficial for capturing L2 vocabulary size more comprehensively.

Additionally, the way the construct of DVK was operationalized in the current study was limited. Through our test of DVK (i.e., WAT), we measured only two aspects of DVK, namely, knowledge of collocations and synonyms, and the latter aspect of DVK (i.e., synonyms) tended to overlap with the construct we measured through our test BVK more strongly. One of the reviewers astutely pointed out that the WAT format is far from ideal for the purpose of measuring depth of VK, and there is an urgent need to develop a more sophisticated measurement of DVK. We propose two key points for future considerations of measurement of DVK. First and foremost, because DVK can be conceptualized in a number of ways, the target construct of DVK needs to be explicitly specified (Schmitt, 2014). For instance, vocabulary researchers recently utilized the corpus information (e.g., collocation frequency and mutual information score) to improve the validity of a receptive collocation knowledge test (Nguyen & Webb, 2017). Second, when a study measures both BVK and DVK, these two constructs should be measured more distinctly. For instance, some aspect of DVK (e.g., derivative knowledge, see Sasao & Webb, 2017 for the Word Part Level Test) may show less overlap with BVK (Schmitt, 2014).

The other limitation of the current study was that the SCT, one of our aural SK measures, required the test takers to listen to isolated sentences, comprehend their meaning, and respond to a comprehension question about the meaning of the sentences. Although the comprehension questions for this test were designed in a way to tap SK, this still created an overlap with the task of listening comprehension ability as measured through the IELTS test. As the result, we acknowledge that our measurement may have created a bias in favor of role of SK in listening comprehension. In future studies, SK should be measured in both written and spoken modalities, and the results of the two methods of measurement should be compared.

In addition, in the models of L2 listening ability (e.g., Field, 2001; Rost, 2002), *proceduralized* aural VK and SK play key roles in successful listening comprehension. In this study, no strong claim can be made for the current aural VK and SK tasks measuring *proceduralized* VK and SK. Future research may apply some time pressure (i.e., learners hear the aural input only once and have limited time to record their responses) on VK and SK tasks and/or utilize reaction-time measures (see, e.g., Godfroid, 2019; Lim & Godfroid, 2015; Suzuki, 2017; Suzuki & Sunada, 2018; Vafaee, Suzuki, & Kachinske, 2017) to make a valid claim for tapping *proceduralized* knowledge.

In terms of data analysis, given the limited sample size in the current study, SEM multisample analysis was not conducted. In this kind of analysis, learners can be divided into several L2 listening ability levels, and the contribution of VK and SK to L2 listening can be examined across different levels. Such analysis can provide indirect evidence of the role of VK and SK in L2 listening from a developmental perspective and across different levels of listening comprehension ability. This issue needs empirical testing because of its important theoretical and practical implications.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0272263119000676

NOTES

[1]The episodic buffer holds integrated episodes and acts as a buffer store between working memory components, perception, and long-term memory.

[2]Linck et al. (2014) classify all word/digit-span tasks, including backward digit/word-span tasks as simple memory tasks.

[3]Inhibition and shifting are two main working memory attentional control mechanisms. Miyake et al. (2000) defined inhibition as “one’s ability to deliberately inhibit dominant, automatic, or prepotent responses when necessary” (p. 57) and shifting as "shifting back and forth between multiple tasks, operations, or mental sets” (p. 55).

[4]Ideally, both VK and SK should be tested under time pressure, and in addition to accuracy data, reaction-time data should be recorded. This approach in measurement enables us to tap *proceduralized* VK and SK (e.g., Godfroid, 2019; Lim & Godfroid, 2015; Suzuki & Sunada, 2018) that is more relevant to the demands of real-time listening comprehension. However, due to the logistical limitations of the current study, we used aural input for our VK and SK tests, but with paper-and-pencil answer sheets. This means we could not record reaction time or enforce meaningful time pressure for the test items. Accordingly, although we tried to measure VK and SK that can be accessed quickly by limiting the time on each item, we have no evidence to claim that we measured *proceduralized* VK and SK.

[5]To ensure that the English and Persian versions were parallel forms and measured the same construct, the translation was done by a professional translator, and then his work was checked and revised by two other native speakers of Persian. They were both PhD students of applied linguistics in the US universities and had advanced levels of English proficiency and a solid familiarity with developing questionnaires for L2 research. The same procedure was followed for translating the anxiety questionnaires.

[6]In the English version of MALQ, French was used as the foreign language, but because our participants were English learners, in the Persian version of the questionnaire, we used English as the foreign language.

[7]For step-by-step instructions on how to compute *scaled* chi-square difference test, visit https://www.statmodel.com/chidiff.shtml.

[8]One of the reviewers astutely pointed out that the knowledge of active/passive and causative target structures included in our SK tests is not purely syntactic, but partially lexical. According to the reviewer, the inclusion of these target structures may have created an overlap between the measurement of VK and SK in our study, which in turn, may have influenced the results of SEM Model 1. To explore this issue empirically, we tested alternative SEM models to Model 1 by excluding the results from these two target structures. The results of these analyses were not different from the ones in SEM Model 1 and as reported in the current article. For these analyses and their results, see Appendix 8 in Online Supplementary Material.

REFERENCES

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: An individual differences approach. *Language Learning*, *62*, 49–78. doi: 10.1111/j.1467-9922.2012.00706.x

Atkins, S. M., Harbison, J. I., Bunting, M. F., & Dougherty, M. R. (2012). Using online games to assess cognitive ability: Validity of two web-deployable measures of cognitive ability. A technical report. AU3

Baddeley, A. D. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, *36*, 189–208.

Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1–29. doi: 10.1146/annurev-psych-120710-100422

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*, 101–118.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge, UK: Cambridge University Press.

Bloomfield, A., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2010). What makes listening difficult? Factors affecting second language listening comprehension. Retrieved from https://apps.dtic.mil/dtic/tr/fulltext/u2/a550176.pdf

Bond, T., & Fox, C. (2001). *Applying the Rasch model*. Mahwaw, NJ: Lawrence Erlbaum Associates.

Brunfaut, T., & Révész, A. (2011). EAP listening task difficulty: The impact of task variables, working memory and listening anxiety. Paper presented at the Language Testing Research Colloquium. Ann Arbor, MI, June 23–25.

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

Call, M. E. (1985). Auditory short-term memory, listening comprehension, and the input hypothesis. *TESOL Quarterly*, *19*, 765–781.

Cohen, J. (1992). Quantitative methods in psychology: A power primer. *Psychological Bulletin*, *112*, 155–159.

Conway, A. R., & Engle, R. W. (1994). Working memory and retrieval: a resource-dependent inhibition model. *Journal of Experimental Psychology: General*, *123*, 354–373.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research*, *169*, 323–338.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671.
Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*, 422–433.
Dittmar, M., Abbot-Smith, K., Lieven, E., & Tomasello, M. (2008). German children's comprehension of word order and case marking in causative sentences. *Child Development*, *79*, 1152–1167.
Elkhafaifi, H. (2005). Listening comprehension and anxiety in the Arabic language classroom. *The Modern Language Journal*, *89*, 206–220.
Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*, 19–23.
Eysenck, M. W., & Calvo, M. G. (1992). Anxiety and performance: The processing efficiency theory. *Cognition & Emotion*, *6*, 409–434.
Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion*, *7*, 336–353.
Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*, 164–203.
Field, J. (2001). Finding one's way in the fog: Listening strategies and second-language learners. *Modern English Teacher*, *9*, 29–34.
Field, J. (2008). Listening in the language classroom. *ELT Journal*, *64*, 331–333.
Field, J. (2009). A cognitive validation of the lecture-listening component of the IELTS listening paper. *IELTS Research Reports*, *9*, 17–65.
Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language* (pp. 77–151). Cambridge, UK: Cambridge University Press.
Godfroid, A. (2019). *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide*. New York, NY: Routledge.
Goh, C. (2005). Second language listening expertise. In K. Jonson (Ed.), *Expertise in second language learning and teaching* (pp. 64–84). New York, NY: Palgrave Macmillan.
Haigh, M., & Stewart, A. J. (2011). The influence of clause order, congruency, and probability on the processing of conditionals. *Thinking & Reasoning*, *17*, 402–423.
Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. doi: 10.1080/10705519909540118
Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research* (Vol. *41*). Amsterdam: John Benjamins.
Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*, 189–217.
Kane, M. J., Conway, A. R. A., Hambrick, D. Z., & Engle, R. W. (2007). Variation in working memory capacity as variation in executive attention and control. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 21–46). New York, NY: Oxford University Press.
Lim, H., & Godfroid, A. (2015). Automatization in second language sentence processing: A partial, conceptual replication of Hulstijn, Van Gelderen, and Schoonen's 2009 study. *Applied Psycholinguistics*, *36*, 1247–1282.
Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review*, *21*, 861–883. doi: 10.3758/s13423-013-0565-2
MacIntyre, P. D., & Gardner, R. C. (1994). The subtle effects of language anxiety on cognitive processing in the second language. *Language Learning*, *44*, 283–305. doi: 10.1111/j.1467-1770.1994.tb01103.x
Meara, P. & Milton, J. (2003). *X_Lex, The Swansea Levels Test*. Newbury: Express.
McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, *19*, 741–760.
Mecartty, F. H. (2000). Lexical and grammatical knowledge in reading and listening comprehension by foreign language learners of Spanish. *Applied Language Learning*, *11*, 323–348.
Mills, N., Pajares, F., & Herron, C. (2006). A reevaluation of the role of anxiety: Self-efficacy, anxiety, and their relation to reading and listening proficiency. *Foreign Language Annals*, *39*, 276–295.
Milton, J. (2010). The development of vocabulary breadth across the CEFR levels: A common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, and textbooks across Europe . In

I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 211–232). Eurosla Monograph Series I. Amsterdam: European Second Language Association.

Milton, J. & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: do vocabulary tests underestimate the knowledge of some learners. *The Canadian Modern Language Review*, *63*, 127–147.

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, M. Torreblanca-López, & M. López-Jiménez (Eds.), *Further insights into non-native vocabulary teaching and learning* (pp. 83–98). Clevedon, UK: Multilingual Matters.

Mueller, R. O., & Hancock, G. R. (2008). Best practices in structural equation modeling. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 488–508). Thousand Oaks, CA: Sage.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*, 59–82.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*, 9–13.

Nguyen, T. M. H., & Webb, S. (2017). Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research*, *21*, 298–320.

Qian, D. D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, *21*, 28–52.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*, 355–371.

Reali, F., & Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language*, *57*, 1–23.

Robertson, E. K., & Joanisse, M. F. (2010). Spoken sentence comprehension in children with dyslexia and language impairment: The roles of syntax and working memory. *Applied Psycholinguistics*, *31*, 141–165.

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, *57*, 348–379.

Rost, M. (1990). *Listening in language learning*. London, UK: Longman.

Rost, M. (2002). *Teaching and researching Listening*. London, UK: Longman.

Rost, M. (2013). *Teaching and researching: Listening*. London, UK: Routledge.

Rubin, J. (1994). A review of second language listening comprehension research. *The Modern Language Journal*, *78*, 199–221. doi: 10.1111/j.1540-4781.1994.tb02034.x

Sasao, Y., & Webb, S. (2017). The word part levels test. *Language Teaching Research*, *21*, 12–30.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*, 243–248.

Scarcella, R. C., & Oxford, R. L. (1992). *The tapestry of language learning: The individual in the communicative classroom.* Boston: Heinle & Heinle.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*, 913–951.

Schmitt, N., Nation, P., & Kremmel, B. (2019). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, 1–12. Advance online publication. doi: 10.1017/S0261444819000326

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*, 139–152.

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, *31*, 577–607.

Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics*, *38*, 1229–1261.

Suzuki, Y., & Sunada, M. (2018). Automatization in second language sentence processing: Relationship between elicited imitation and maze tasks. *Bilingualism: Language and Cognition*, *21*, 32–46. doi: 10.1017/S1366728916000857

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). New York, NY: Harper & Row.

Vafaee, P., Suzuki, Y., & Kachisnke, I. (2017). Validating grammaticality judgment tests: Evidence from two new psycholinguistic measures. *Studies in Second Language Acquisition*, *39*, 59–95.

Vandergrift, L. (2004). Learning to listen or listening to learn. *Annual Review of Applied Linguistics*, *24*, 3–25.

Vandergrift, L. (2006). Second language listening: Listening ability or language proficiency? *The Modern Language Journal*, *90*, 6–18. doi: 10.1111/j.1540-4781.2006.00381.x

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, *40*, 191–210.

Vandergrift, L. (2011). Second language listening: Presage, process, product, and pedagogy. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 455–471). New York, NY: Routledge.

Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning*, *65*, 390–416. doi: 10.1111/lang.12105

Vandergrift, L., & Goh, C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. New York, NY: Rutledge.

Vogely, A. J. (1998). Listening comprehension anxiety: Students' reported sources and solutions. *Foreign Language Annals*, *31*, 67–80.

Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch measurement transactions*, *8*, 370.