tesol QUARTERLY
tesol international association

The Role of Task and Listener Characteristics in Second Language Listening

TINEKE BRUNFAUT
Lancaster University
Lancaster, England

ANDREA RÉVÉSZ
Institute of Education
University of London
London, England

This study investigated the relationship between second language (L2) listening and a range of task and listener characteristics. More specifically, for a group of 93 nonnative English speakers, the researchers examined the extent to which linguistic complexity of the listening task input and response, and speed and explicitness of the input, were associated with task difficulty. In addition, the study explored the relationship between L2 listening and listeners' working memory and listening anxiety. The participants responded to 30 multiple-choice listening items and took an English proficiency test. They also completed two working memory tasks and a listening anxiety questionnaire. The researchers analysed listening input and responses in terms of a variety of measures, using Cohmetrix, WebVocabProfiler, Praat, and the PHRASE list, in combination with expert analysis. Task difficulty and participant ability were determined by means of Rasch analysis, and correlational analyses were run to investigate the task and listener variables' association with L2 listening. The study found that L2 listening task difficulty correlated significantly with indicators of phonological, discourse, and lexical complexity and with referential cohesion. Better L2 listening performances were delivered by less anxious listeners and, depending on L2 listening measure, by those with a higher working memory capacity.
doi: 10.1002/tesq.168

The construct of second language (L2) listening difficulty, defined here as the objective rather than perceived challenge posed by listening (Robinson, 2001), has received increasing interest from researchers in recent years. A number of factors associated with the

characteristics of the listening task and the listener have been proposed, and some demonstrated, to determine listening difficulty. Among the listening task factors that have been investigated are task input variables such as passage length, speech rate, linguistic complexity, and text content; variables related to task procedures, such as the nature of the task instructions and the number of times listening; and task response characteristics such as the item type, or the length and complexity of the required response (for comprehensive reviews, see Bloomfield et al., 2010; Vandergrift, 2007). Only a handful of empirical studies have dealt with listener-related characteristics. Variables that have been investigated include cognitive factors—for example, working memory (WM) capacity (Andringa, Olsthoorn, van Beuningen, Schoonen, & Hulstijn, 2012; Kormos & Sáfár, 2008); awareness of metacognitive strategies (Vandergrift, Goh, Mareschal, & Tafaghodtari, 2006); affective factors, including anxiety (Elkhafaifi, 2005) and motivation (Vandergrift, 2005); and linguistic factors such as language knowledge (Andringa et al., 2012; Staehr, 2009) and linguistic processing speed (Andringa et al., 2012).

The aim of our study was twofold. The first aim was to replicate and extend existing research by examining the role of a large array of task input and response features in determining L2 listening difficulty. We focused on a task type yet unexplored and used some linguistic measures new to listening comprehension research. The second aim was to contribute to the underresearched area of listener-related factors by exploring the influence of WM and listening anxiety on listening difficulty using a novel population of L2 users and novel assessments.

LISTENING COMPREHENSION

L2 listening is an interactive, cognitive process, which involves neurological, linguistic, semantic, and pragmatic processing (Rost, 2011) while drawing on resources such as linguistic knowledge, world knowledge, and knowledge about the communicative context (Buck, 2001; Rost, 2011; Vandergrift, 2007). Hence, the ability to integrate in real time information from the various knowledge sources is considered crucial for successful listening comprehension (Rost, 2005). This process is highly automatized, requiring little or no conscious attention from proficient listeners, but demands more controlled, conscious processing from those with more limited L2 knowledge and/or less efficient processing skills (Segalowitz, 2003). Given that WM is limited in capacity (Baddeley, 2000), less automated processing is more taxing. Buck (2001), Vandergrift (2007), and others have suggested that this leads to partial comprehension or miscomprehension by L2 listeners.

However, as already mentioned, the extent to which listening comprehension poses difficulty for L2 users is determined by a variety of factors, among them task- and listener-related factors.

LISTENING TASK CHARACTERISTICS

Linguistic Complexity

A large number of linguistic factors have been identified as affecting L2 listening comprehension, including phonological, lexical, syntactic, and discourse features of the passages. In our study, we looked into aspects of each of these four types. Although these aspects cover a wide range of characteristics, space limitations compel us to emphasize variables that are directly relevant to our investigation. For a more extensive review, we refer the reader to Bloomfield et al. (2010) and Révész and Brunfaut (2013).

In the literature on the phonological complexity of listening input, contracted forms (e.g., *can't*) have been proposed to have a potentially negative impact on decoding auditory information, because the recognition of lexical items and syntactic constructions might pose a greater challenge due to decreased phonological information (e.g., Rubin, 1994). Reflecting this prediction, Henrichsen (1984) found that the presence of reduced forms created greater difficulty for L2 learners than for first language (L1) listeners. Révész and Brunfaut (2013) and Kostin (2004), in contrast, observed no effects for missing phonemic information. In both of these studies, however, the listening texts were scripted and delivered by actors, which might have decreased item variation.

Several aspects of the lexical complexity of listening input have been found to relate to listening difficulty in prior research (see Révész & Brunfaut, 2013). One of these is lexical sophistication, which can be defined as the percentage of low-frequency words or formulaic expressions in a text. Previous research on lexical sophistication was motivated by the assumption that texts with a greater proportion of infrequent lexis will be more demanding, because low-frequency words are less likely to be known or recognised by L2 listeners (e.g., Muljani, Koda, & Moates, 1998). Some investigations have demonstrated that low-frequency words do contribute to listening difficulty (Kostin, 2004; Nissan, DeVincenzi, & Tang, 1996; Révész & Brunfaut, 2013), but others have found no effects for infrequent lexis (Yanagawa & Green, 2008; Ying-hui, 2006). It appears likely that L2 listening difficulty is also related to the amount and nature of formulaic expressions in the input, in light of L1 findings that the presence and corpus-derived frequency of formulae may impact oral processing (e.g., Conklin &

Schmitt, 2008). Previous studies have yielded results contrary to this prediction (Kostin, 2004; Révész & Brunfaut, 2013; Ying-hui, 2006).

In addition to lexical sophistication features of texts, it has been suggested that texts with more varied lexis cause more listening difficulty, because they require the listener to decode and process a greater quantity of distinct words. Indeed, Rupp, Garcia, and Jamieson (2001) and Révész and Brunfaut (2013) found a significant association between lexical diversity and L2 listening success.

Furthermore, texts with greater lexical density—that is, a higher ratio of content words to the total word count—are expected to put more pressure on processing mechanisms (Bloomfield et al., 2010), because they carry more information due to the higher proportion of content words for the total amount of words. This hypothesis was confirmed by Buck and Tatsuoka (1998), who found that listening difficulty was increased when a greater percentage of content words surrounded the information relevant to task completion. Révész and Brunfaut (2013) also identified lexical density as a strong predictor of listening difficulty.

With reference to another lexical aspect, that is, the concreteness of content words in listening input, psycholinguistics research has indicated that concrete words are generally processed and retrieved faster than abstract words (e.g., Paivio, Walsh, & Bons, 1994). Indeed, Freedle and Kostin (1996) reveal that TOEFL minitalks with higher abstractness ratings resulted in greater listening difficulty. Révész and Brunfaut (2013), however, found no effects for this factor, but, as the authors acknowledge, this might have been due to lack of sufficient variance in concreteness across the passages under scrutiny.

Syntactic processing is regarded as an important component of the process of decoding aural information (Rost, 2011), and syntactic complexity features of listening input have been suggested to have an effect on processing. For example, Carpenter and Just (1975) propose that the presence of negative markers has a negative influence on sentence processing. Therefore, it would appear probable that listening texts that are more syntactically complex and/or contain more negative expressions will exert higher processing demands. However, existing research, overall, suggests that listening difficulty is not significantly related to any subconstruct of structural complexity (Norris & Ortega, 2009)—complexity by subordination (Blau, 1990; Kostin, 2004; Révész & Brunfaut, 2013; Ying-hui, 2006; but see the significant relationship between syntactic simplification and comprehension in Cervantes & Gainer, 1992), phrasal complexity (Révész & Brunfaut, 2013), and overall complexity (Kostin, 2004; Révész & Brunfaut, 2013; Ying-hui, 2006). For negatives (markers such as *not* and prefixes such as *dis-*), some empirical studies confirm adverse effects (Freedle & Kostin, 1999; Kostin, 2004; Nissan et al., 1996), whereas others reveal

no relationship between listening difficulty and their presence (Révész & Brunfaut, 2013; Yanagawa & Green, 2008; Ying-hui, 2006).

In terms of discourse complexity, research has particularly focused on how cohesion may determine listening difficulty, expecting that more cohesive listening texts will be easier to comprehend. Nissan et al. (1996), however, identified no effects for either explicit lexical or structural cohesive links in TOEFL minidialogues. In Révész and Brunfaut's (2013) study, on the other hand, causal content emerged as a significant predictor. Listening passages with fewer causal verbs and particles were associated with more successful performance. In Ying-hui (2006), experts also assigned higher cohesion ratings to easier listening test items.

Explicitness

More explicit texts are anticipated to pose less challenge for listeners, because they demand less engagement in pragmatic processes such as inferencing and evaluation of contextual factors (Rost, 2011). Previous empirical studies indeed suggest that texts that require more inference result in increased comprehension difficulty (Garcia, 2004; Kostin, 2004; Nissan et al., 1996; Taguchi, 2005; Ying-hui, 2006). An exception to this trend was observed by Révész and Brunfaut (2013), who uncovered no relationship between explicitness and success in listening. As the researchers noted, this was probably due to a ceiling effect on the native speaker ratings.

Speed of Delivery

One input characteristic that has received a significant amount of attention in the research literature is speed of delivery. Faster delivery of speech is assumed to cause more listening difficulty, because it affords a shorter period of time to process the incoming information. Both experimental (e.g., Griffiths, 1992; Zhao, 1997) and nonexperimental research (Brindley & Slayter, 2002; Buck & Tatsuoka, 1998) corroborate this assumption (but see the lack of impact of speed found in Révész & Brunfaut, 2013).

Response Characteristics

Researchers have also found effects on listening task difficulty of task response aspects such as the format (Brindley & Slayter, 2002;

In'nami & Koizumi, 2009) and length of response (Buck & Tatsuoka, 1998; Jensen, Hansen, Green, & Akey, 1997). For example, In'nami and Koizumi's (2009) meta-analysis of the effect of multiple-choice versus open-ended questions on L2 listening performance indicates that multiple-choice formats are easier. Jensen et al. (1997) found an association between more difficult listening items and lengthy, multiple-word answers.

Furthermore, particular combinations of task input and response characteristics seem to contribute to listening task difficulty. For instance, Jensen et al. (1997) reveal a relationship between task difficulty and lexical overlap between words in the text and the response. In addition, research has specifically looked into the characteristics of those parts of the text that need to be understood for successful task completion (as determined by the required response; see, e.g., Buck & Tatsuoka, 1998). Révész and Brunfaut (2013), for example, found that the information necessary to respond in less difficult tasks contained more function words or formulaic expressions or had higher lexical density.

LISTENER CHARACTERISTICS

Working Memory

Working memory has been proposed to be a vital component of comprehension processes and thus a possible source of individual differences in comprehension ability (e.g., Daneman & Carpenter, 1980). For our research, we adopted Baddeley and Hitch's (1974) WM model. This model defines a multicomponent memory system comprising a central executive, two domain-specific subsystems (a phonological loop and a visual–spatial sketchpad), and an episodic buffer (Baddeley, 2000). Of particular interest in the context of listening comprehension are the phonological loop, which is responsible for the temporary retention and manipulation of verbal information, and the central executive, which coordinates complex cognitive operations such as attention focus and division, control of information flow, and regulation of processing routines (Gathercole, 1999). Both the phonological loop and the central executive are limited in capacity.

Measures typically employed to assess phonological short-term memory (PSTM) are forward digit span tasks and non-word repetition tasks, which require immediate recall of a series of numbers or non-words of increasing length, respectively. More complex verbal WM operations which also involve the central executive are typically captured by backward digit span or reading and listening span tasks.

Backward digit span tasks entail the recall of increasing numbers of digits in reverse order, whereas reading and listening span tasks require recalling the final words of increasingly longer series of sentences/utterances.

Individual differences in PSTM and complex WM capacity have been shown to predict L2 processing abilities. There is empirical evidence suggesting that PSTM (Kormos & Sáfár, 2008; O'Brien, Segalowitz, Freed, & Collentine, 2007) and complex WM capacity (Kormos & Sáfár, 2008) are positively associated with aspects of speech production. Complex WM capacity, additionally, has been linked to L2 comprehension abilities, including syntactic processing (e.g., Miyake & Friedman, 1998) and reading comprehension (e.g., Harrington & Sawyer, 1992; Kormos & Sáfár, 2008).

To our knowledge, only two studies have examined the relationship between WM and L2 listening ability. Kormos and Sáfár (2008) found a significant, moderate-sized correlation between performances on the listening section of the Cambridge First Certificate Exam (FCE) and complex WM capacity (measured by an L1 auditory backward digit span test), but not between PSTM (operationalized as an L1 non-word span test) and FCE listening. Andringa et al. (2012) analysed the role of WM in L2 listening as part of a multiple-component model, which examined the impact of L2 knowledge and processing skills, intelligence, and WM on listening ability. Working memory, expressed as a latent factor based on five WM measures (forward and backward auditory digit span, forward and backward visual digit span, and a non-word recognition task), did not explain any unique variance in scores on the Dutch State Exam of Listening Proficiency. However, L2 listening had weak correlations with the auditory and visual forward digit span and the auditory backward digit span measures. Clearly, further research is required to disentangle the nature of the relationship between WM and L2 listening.

Listening Anxiety

Listening anxiety is a type of situation-specific anxiety (MacIntyre & Gardner, 1991) that learners may uniquely experience when engaged in L2 listening. Although general foreign language (FL) anxiety is well researched (for a review, see Horwitz, 2010), relatively little research has been dedicated to investigating anxieties related to specific language abilities, and research on listening anxiety is particularly sparse. Additionally, in the few existing studies, the focus has often been the nature of FL listening anxiety itself (e.g., Kimura, 2008; Vogely, 1998) rather than the link between anxiety and comprehension.

Some empirical research, however, has explored whether listening anxiety and comprehension are related, and as expected by the researchers it yielded negative correlations. In a study of 253 Korean students of English as a foreign language, Kim (2000) found a moderate association between listening anxiety and comprehension. Listening anxiety was measured by a 33-statement Foreign Language Listening Anxiety Scale, and Kim determined the participants' listening proficiency based on TOEFL listening tasks.

Elkhafaifi (2005) further examined whether a relationship exists between listening anxiety and listening comprehension. The participants were 233 Arabic FL learners, whose FL listening anxiety was assessed by a 20-statement Foreign Language Listening Anxiety Scale. Listening ability was represented by learners' listening comprehension grade for their Arabic FL course. The analyses reveal a strong negative correlation between listening anxiety score and students' final listening achievement grade. Elkhafaifi interprets this as "lend[ing] support to the premise that increased anxiety adversely affects [students' listening] performance" (p. 214). Elkhafaifi rightly acknowledges, however, that using final grades as a measure of listening ability may have jeopardized the reliability of the findings.

RESEARCH QUESTIONS

To summarize, as compared to other language skills, our insights into the nature of L2 listening and what makes listening difficult for L2 learners are limited. Although the research base is growing steadily, little is known about how the success of L2 listening is influenced by the characteristics of the listening task and the second language listener. Only a limited number of listening task types have been looked into, few studies investigating the impact of task input characteristics on listening difficulty have been replicated, and researchers have rarely considered task input as well as response characteristics. In addition, the measures used to tap task input features in listening research have not always been optimal. Similar to listening task features, L2 listener characteristics, including WM and listening anxiety, have been the focus of relatively few studies. Investigations of the impact of WM on L2 listening have not only been scarce but also yielded contradictory conclusions. Although findings on the role of listening anxiety have been more consistent, research on this topic would benefit from using more precise L2 listening measures and exploring whether the findings apply to other populations (e.g., an English L2 population with more mixed backgrounds).

In view of these research gaps, we formulated four research questions (RQs), two concerning task characteristics (1 and 2) and two concerning listener characteristics (3 and 4):

1. Is there a relationship between listening task difficulty and characteristics of task input?
2. Is there a relationship between listening task difficulty and characteristics of task response?
3. Is there a relationship between listening performance and working memory?
4. Is there a relationship between listening performance and listening anxiety?

We addressed these research questions in relation to an unexplored listening item type and a novel population.

METHODOLOGY

Participants

Ninety-three students participated in the study. These were nonnative English speakers intending to enrol in, or studying towards, an undergraduate (53%) or postgraduate (47%) degree at a university in the United Kingdom. Fifty-one percent were Mandarin Chinese L1 speakers, 14% had other Asian L1 backgrounds, and 32% had European L1 backgrounds. Seventy percent were female and 30% male. Their ages ranged from 18 to 43 ($M = 22.83$, $SD = 4.42$). Fifty-four percent were at level B1 of the Common European Framework of Reference (CEFR) in terms of overall English proficiency, as measured by the Pearson Test of English (PTE) Academic Scored Practice Test. Approximately a fifth of the students were in the bands A2 (19%) and B2 (18%), respectively. A minority was at levels A1 (3%) and C1 (5%).

Instruments

Listening ability and item difficulty. The following instruments were administered to obtain measures of listening task difficulty (RQs 1 and 2) and listening performance (RQs 3 and 4).

Listening task. To control for task type effects (In'nami & Koizumi, 2009), our primary measure of L2 listening was restricted to one task type, a multiple-choice listening task. In practice, it concerned the Select

Missing Word (SMW) task of the PTE Academic, which constitutes a relatively independent measure of listening. This task requires the test taker to listen to a passage for which the end is missing and to select an appropriate ending from multiple-choice options. The following is a publicly available example (not used in this study):

> *Prompt*: You will hear a recording about an analysis of medical research findings. At the end of the recording the last word or group of words has been replaced by a beep. Select the correct option to complete the recording.
>
> *Transcript*: Speaker 1: My PhD student, Elaine Chong, did what was called a meta-analysis where you analyze the literature very carefully, to see what evidence there is to suggest that what we eat, particularly in terms of antioxidants, prevents you getting macular degeneration.
>
> Speaker 2: Because ophthalmologists have been using antioxidants quite a lot.
>
> Speaker 1: That's right. There have been studies looking at whether antioxidant supplements slow the progression once you have the disease, whereas this study was looking at trying to stop you getting it ...
>
> *Options*
>
> □ before you understood
> □ after diagnosis
> □ from anti-oxidants
> □ in the first place

Thirty items were selected for the study, representing a wide range of item difficulties (based on statistics provided by Pearson). The participants were presented with the items in a split-block design and completed the tasks at individual work stations in a computer lab. The results of this set of listening items were used to establish a measure of item difficulty (RQs 1 and 2) and a measure of listener ability (RQs 3 and 4; see preliminary data analysis in the Analyses and Results section below).

Proficiency test. Participants' L2 proficiency was measured by a scored practice version of the PTE Academic, testing independent and integrated skills. This test was delivered and scored online by Pearson software. Test performance was reported in terms of an overall score, with separate scores for communicative skills (listening, reading, speaking, and writing) and for enabling skills (grammar, oral fluency,

pronunciation, spelling, vocabulary, and written discourse; Pearson Education, 2012). The overall score served as a measure of L2 proficiency, and the listening score constituted a second, separate measure of L2 listening ability, in addition to the SMW measure (RQs 3 and 4).

Listening task characteristics. A wide range of tools was selected to establish the measures of the targeted input and response characteristics (RQs 1 and 2).

Listening input characteristics. Based on our review of the literature on listening text characteristics which relate to or impact on listening task difficulty, we operationalized the input characteristics as the prompts' phonological, lexical, semantic, morphosyntactic, and discourse complexity; speed of delivery; and explicitness of the texts. Table 1 gives an overview of the range and nature of the measures and the units of analysis we adopted for each of these characteristics.

Particular methodological improvements include the way in which we measured lexical sophistication. With the exception of Révész and Brunfaut (2013), listening researchers have regarded lexical sophistication as a dichotomous factor (occurrence or lack of low-frequency words), whereas we considered the quantity of low-frequency words. Also, we took into account the corpus-based frequency of multiword expressions in the texts, unlike previous studies which considered only whether formulae were present or absent. To measure lexical diversity, we used the D-formula, which has been argued to be more reliable and robust than type-token ratio (Malvern, Richards, Chipere, & Durán, 2004). In addition, we included both speech and articulation rate measures to assess speed of delivery. The majority of existing studies have exclusively used measures of speech rate which include silent pauses (e.g., Brindley & Slayter, 2002; Griffiths, 1992). As Bloomfield et al. (2010) imply, however, to fully capture delivery speed, studies would ideally combine a speech rate measure with an articulation rate index calculated by excluding silent pauses, because pausing may facilitate comprehension.

Table 1 also indicates the way in which the measures were obtained. The linguistic measures were established through expert analysis or by using the computer software Praat v5.0.25 (Boersma & Weenink, 2008), vocd of the CHILDES system (MacWhinney, 2000), WebVocabProfile v3 (Cobb, n.d.), and CohMetrix v2.0 (McNamara, Louwerse, Cai, & Graesser, 2005). Speed was determined using Praat. The explicitness of the texts was established through five native speakers' judgements on a 5-point Likert scale.

TABLE 1
Measures Used to Analyse the Listening Prompts

Measure	Unit of analysis[a]	Analytical procedure/ computational tool
Phonological complexity	Frequency of contractions	Praat, expert analysis
Lexical complexity		
Lexical sophistication	Proportion of • K1 words • K1 function words • K1 content words • K2 words • K1 + K2 words • Academic words • Off-list words	WebVocabProfile
	Proportion of • multiword expressions • K1, K2, K3, K4, K5 multiword expressions • multiword expressions as in spoken/written/written academic language	Calculated by researchers using PHRASE list (Martinez & Schmitt, 2012)
Lexical diversity	Tokens D	WebVocabProfile Vocd
Lexical density	Lexical density	WebVocabProfile
Concreteness	Mean concreteness value for all content words	CohMetrix
Syntactic complexity	Mean number of modifiers per noun phrase Mean number of higher level constituents per sentence, controlling for number of words Incidence score of: • negations • all connectives • logical operators	CohMetrix
	Words per clause Words per AS-unit Clause per AS-unit	Calculated by researchers
Discourse complexity	Causal and intentional content Causal, intentional, temporal, and spatial cohesion (Adjacent) anaphor reference (Adjacent) argument overlap (Adjacent) stem overlap	CohMetrix
Speed	Syllables per second (speech rate) Syllables per second without pauses (articulation rate)	Praat, expert analysis
Explicitness	Native speaker judgments	Analysed by researchers

[a]*K* in 1K, 2K, etc. stands for most frequent *x* thousand word families in English.

For further information on the nature of each of these measures and units of analysis, we refer the reader to Cobb (n.d.), MacWhinney (2000), McNamara et al. (2005), and Révész and Brunfaut (2013).

Listening response characteristics. Although in theory a wide range of response characteristics may be associated with difficulty, we restricted our selection to lexical frequency measures due to the specific nature of the multiple-choice options of the 30 SMW tasks (i.e., the length of each option in the task versions was limited to a few words or even one). WebVocabProfile v3 (Cobb, n.d.) was used to obtain measures for the proportion of the most frequent first and second thousand word families, academic words, off-list words, and the number of tokens for each item's response options. Informed by our literature review, we further measured lexical overlap between words in the text and the response options.

In addition, because in a (test) task context the relationship of the items to the input may determine the focus and difficulty of the task, five expert linguists associated with a British university determined what substituted those parts of the text necessary for task completion. Information was considered necessary if a minimum of three linguists identified it as such. The measures selected to operationalize the characteristics of the necessary information were lexical frequency (determined through WebVocabProfile) and formulaic expressions using Martinez and Schmitt's (2012) Phrasal Expression List.

Listener characteristics. To obtain measures for the listener characteristics working memory and listening anxiety (RQs 3 and 4), the following instruments were used.

Working memory tests. Two WM tests were administered: a forward digit span test as a measure for PSTM and a backward digit span test to assess complex WM capacity. Visual rather than auditory digit span tests were used to capture a more language-independent construct and thus reduce the overlap between the instruments used to assess listening and WM (Andringa et al., 2012). Auditory WM tests would have had to be administered in participants' L1, because L2 span results would have been confounded by L2 proficiency effects (Masoura & Gathercole, 1999; Olsthoorn, Andringa, & Hulstijn, 2014). However, participants' L1 could not be anticipated prior to task administration. Visual digit span tests have the advantage of avoiding language-specific input whilst allowing L1 verbalisation.

Both digit span tests consisted of a series of numbers which increased in length, with two equally long strings per level. These were presented by means of a PowerPoint presentation, displaying one digit

per second. After having seen a series of numbers, participants were instructed to write down the series in the order they had been shown for the forward digit span test and in reverse order for the backward digit span test. Participants' digit span was determined as the maximum list length at which they could produce at least one of the two sequences correctly.

Listening anxiety questionnaire. To assess listening anxiety, we administered the Foreign Language Listening Anxiety Scale, developed by Elkhafaifi (2005). The scale, originally designed for Arabic, was adapted to English. It included 20 statements which students needed to judge on a 5-point Likert scale (ranging from *strongly disagree* to *strongly agree*). An example of a statement is "I get upset when I'm not sure whether I understand what I'm hearing in English."

Procedure

Prior to the study, ethical approval was obtained from the Research Ethics Committee at the researchers' institution, and potential participants were provided with information sheets detailing the purpose, nature, and demands of the study. Written consent was obtained from willing participants.

The data were gathered during two sessions, within a period of 2 weeks. During one session, participants completed the background questionnaire and the proficiency test. During the other session, they completed the listening test, WM tests, and listening anxiety questionnaire.

ANALYSES AND RESULTS

Preliminary Data Analyses and Descriptive Statistics

Prior to focusing on the research questions, a number of preliminary analyses had to be conducted to obtain measures for the different variables in the study.

Listening

Listening task difficulty. In order to establish the difficulty of the 30 SMW items to obtain a measure of listening task difficulty for RQs 1 and 2, estimates were obtained by running a dichotomous Rasch model with the facets *participant ability* and *item difficulty* (Linacre, 1989). The

item difficulty estimates ranged from −1.86 to 1.76 logits (from easy to difficult; $M = 0$; $SD = 1.99$), and the items reliably differed from one another ($\chi^2(29) = 341.7$, $p < .01$, .93 separation reliability). The infit mean square values were all in the acceptable range of .83 to 1.15 (i.e., $M \pm 2SD$), except for three items, which had marginally misfitting values. In other words, the tasks overall performed in the expected manner, and a difficulty estimate could be established for each.

Listening performance. Participants' L2 listening ability (see RQs 3 and 4) was assessed by their performances on the 30 SMW items and the overall listening score of the PTE Academic Scored Practice Test.

On the basis of the SMW task performances, Rasch participant ability estimates were determined by running a dichotomous Rasch model (see above). The mean participant ability estimate was .16 logits, ranging from −1.59 to 2.95 (low to high ability; $SD = 0.99$; $SE = 0.43$). The participant abilities were spread out on the logit scale with acceptable consistency ($\chi^2(92) = 381.3$, $p < .01$, .84 separation reliability). The fit statistics identified one participant as slightly misfitting, representing 1% of the total number of participants and thus within the acceptable range of 2% (Pollitt & Hutchinson, 1987).

Listening scores for the PTE Academic Scored Practice Test resulted from performances on 11 task types, each consisting of 2–12 items. Scores ranged between 15 and 88 out of 90, with a mean of 52.34 ($SD = 14.22$). Based on PTE Academic–CEFR alignments, 52% of the students were at CEFR level B1, 20% at A2, and 19% at B2.

Listening task characteristics. As shown in Table 1, the characteristics of the listening input (RQ1) were operationalized as the prompts' linguistic complexity (i.e., phonological, lexical, semantic, morphosyntactic, and discourse complexity), speed of delivery, and explicitness of the texts; and a range of units of analysis and tools were used to establish measures for the various input characteristics. As a result of this process, two measures, incidence of K4- and K5-band multiword expressions, were removed from further analyses due to the very low incidence of formulae from these bands in the passages. Outliers were identified using boxplots (i.e., values below the 25th percentile or above the 75th percentile by at least 1.5 times the interquartile range). These outliers were removed for each measure. The resulting variabilities were checked for each input measure, and the mean and standard deviation values for all indices indicated sufficient variation to conduct meaningful correlational analyses. As explained in the Instruments section, due to the nature of the SMW item type, the characteristics of the listening response and the necessary information were limited to lexical complexity measures.

Listener characteristics

Working memory capacity. Given that WM scores may vary as a function of first language (Randall, 2007), only scores for the largest shared-L1 group, 47 Chinese participants, were included in further analyses. The average list length for the forward and backward digit span tests (RQ3) were close to 9 (M = 8.87; SD = 1.36) and 7 digits (M = 6.78; SD = 1.57), respectively.

Listening anxiety. The measure for listening anxiety (RQ4) was based on the 5-point Likert scale answers of the Foreign Language Anxiety Scale. The total scores ranged between 28 and 89 out of a maximum possible of 100 (higher scores indicate higher anxiety), with a mean of 53.94 (SD = 11.12). The internal consistency of the scale was .84 (Cronbach's alpha).

Having established all above measures, we were able to conduct the analyses needed to be able to look into the relationship between listening task difficulty and task-related factors (RQs 1 and 2) and the relationship between listening performance and listener-related factors (RQs 3 and 4).

Listening Task Difficulty and Task-Related Factors

Listening task difficulty and task input characteristics. To look into the relationship between listening task difficulty and characteristics of listening task input (RQ1), a series of Spearman rank-order correlations were performed between the 30 listening passages' text characteristics and task difficulty. The task difficulty estimates from the Rasch analyses were correlated with one of the listening text characteristics in each analysis. Table 2 displays the correlation coefficients for the analyses which yielded significant results.

Eight text characteristics proved to be significantly associated with task difficulty. *Frequency of contractions* (e.g., *we'll*) had a strong negative correlation with task difficulty; passages with a larger number of contractions were significantly less demanding. Four measures associated with *multiword expressions* were significantly correlated with task difficulty:

1. The proportion of *multiword expressions in the K3 band* had a moderate negative relationship with item difficulty. Passages with a higher proportion of K3 multiword expressions (e.g., *rely on*, *on the basis*, *in a way*) proved less difficult.
2. The proportion of multiword expressions which are *most common in written general discourse* was moderately related to task difficulty. Passages which contained more of the most common multiword

TABLE 2
Significant Correlations Between Task Input Text Characteristics and Task Difficulty Estimates

Task input text characteristic	Spearman's rho
Contractions	−.537**
Multiword expressions	
K3 multiword expressions	−.367*
Most common in written general discourse	−.425*
Less common in written academic discourse	−.611**
Rare in written academic discourse	.397*
Argument overlap	−524**
Adjacent argument overlap	−.377*
Stem overlap	−.465*

**$p < .01$, *$p < .05$.

expressions in written general discourse (e.g., *as well as*, *deal with*, *over the years*) posed significantly less difficulty.

3. The proportion of multiword expressions which are *less common in written academic discourse* had a strong negative relationship with listening difficulty. The higher proportion of less common academic multiword expressions used in written academic discourse occurred in the texts (e.g., *in fact*, *on the other hand*, *manage to*), the less challenging the task was likely to be.
4. The proportion of multiword expressions which are *rare in written academic discourse* was moderately related to task difficulty. Passages which included a larger number of multiword expressions rare in written academic discourse (e.g., *sort of*, *think about*, *go off*) were significantly more demanding.

The three remaining text characteristics associated with task difficulty have to do with overlap. *Argument overlap* (the proportion of all sentence pairs in a paragraph that share one or more arguments, such as a [pro]noun or noun phrase) had a strong relationship with task difficulty; passages with a higher proportion of argument overlap proved less difficult. *Adjacent argument overlap* (the proportion of adjacent sentences that share one or more arguments) was moderately related to task difficulty; the higher proportion of adjacent argument overlap, the less demanding the task was found to be. *Stem overlap* (the proportion of all sentence pairs in a paragraph that share one or more word stems; for example, *division* and *divide*) had a moderate correlation with item difficulty; passages with a higher proportion of stem overlap were less demanding.

Listening task difficulty and task response characteristics. We investigated the relationship between listening task difficulty and task

response characteristics (RQ2) by means of Spearman rank-order correlations between the listening task difficulty estimates from the Rasch analysis and one of the multiple-choice lexical measures at a time. The analysis was conducted for all options together and for the correct response and the distractors separately. It was found that none of the lexical complexity measures correlated significantly with listening difficulty.

To cast more light on the role of response characteristics, the interaction between response and task input was examined. In particular, we analysed the relationship between task difficulty and lexical overlap between words in the text and the response options. The passages did not share lexis with the correct options, but did with some distractors. However, no significant association was found between passage-response lexical overlap and task difficulty.

In addition, we assessed the relationship between task difficulty and the lexical complexity of those parts of the text necessary for task completion by running correlations between the Rasch task difficulty estimates and the lexical complexity measures. This yielded a moderate negative correlation between task difficulty and the proportion of multiword expressions identified as necessary to answer the item correctly (Spearman's rho = $-.375$, $p < .05$). Tasks for which the necessary passage information included higher proportions of multiword expressions were less demanding for participants.

Listening Performance and Listener-Related Factors

Listening performance and listeners' working memory. To investigate the relationship between L2 listening performance and WM (RQ3), we examined the association of the listening scores with the digit span tests. Specifically, the Rasch participant ability estimates for the SMW task performances and the PTE Academic overall listening scores were, in separate analyses, correlated with the forward digit span scores and with the backward digit span scores.

No statistically significant relationship could be identified between the performances on the SMW task and either of the WM measures. However, a significant association was found between the forward digit span scores and the PTE Academic listening scores (Spearman's rho = .297, $p < .05$), indicating a positive, close to moderate relationship between PSTM and L2 listening performance. Similarly, the backward digit span scores and the PTE Academic listening scores correlated significantly (Spearman's rho = .312, $p < .05$). Complex verbal WM capacity showed a moderate, positive link with L2 listening performance.

Listening performance and listening anxiety. To explore a potential relationship between L2 listening performance and listening anxiety

(RQ4), students' scores on the Foreign Language Anxiety Scale were correlated with the Rasch participant ability estimates for the SMW performances and their PTE Academic listening scores. We found that listening anxiety correlated significantly with the listening performances on the SMW task (Spearman's rho = −.440, $p < .001$) and on the PTE Academic (Pearson r = −.544, $p < .001$). Higher levels of listening anxiety were associated with lower L2 listening performance.

DISCUSSION AND CONCLUSIONS

Listening Task Difficulty and Task-Related Factors

Listening task difficulty and task input. In line with previous findings that lexical complexity has an effect on task difficulty (e.g., Nissan et al., 1996; Révész & Brunfaut, 2013), we found that four lexical complexity characteristics of the listening passages significantly correlated with task difficulty. The nature of the relationship between task difficulty and individual phrase-related characteristics seems to depend on the corpus-based frequency of the expressions (see Martinez & Schmitt, 2012), and may explain conflicting results of past research. On the one hand, higher occurrences of the more frequently used multiword expressions (K3 band, most common in written general discourse, and less common in written academic discourse[1]) had a moderate negative relationship with item difficulty, thus passages with a higher proportion of these phrases were easier. This finding corroborates results of L1 processing studies indicating that multiword expressions are associated with lower processing load (e.g., Conklin & Schmitt, 2008). The lack of significant findings for multiword expressions in the K1–K2 bands could potentially be because these were generally well understood by all participants, whereas comprehension levels of K3 band phrases differed according to proficiency. It should also be noted that the terminology may be misleading in the sense that *less common* is relative to *most common*; the PHRASE list breaks down frequency information into four categories: most common, less common, infrequent, and rare or nonexistent in (a certain genre) (Martinez & Schmitt, 2012).

On the other hand, a larger presence of low-frequency phrases (rare in written academic discourse) co-occurred with more difficult tasks. Similar frequency factors might explain Kostin's (2004) conclusion that the presence of idioms seems to increase task difficulty; the multiword expressions in her study may have been lower frequency ones.

[1] Note that the very low occurrence of the same category of multiword expressions for the other discourse types (spoken, written general, or academic) means that there is little potential for a relationship with task difficulty.

The significant associations of the discourse complexity measures *(adjacent) argument overlap* and *stem overlap* with task difficulty suggest that greater referential cohesion (Kintsch & Van Dijk, 1978) may decrease task demands. The measures give an indication of the extent to which (pro)nouns, noun phrases, and word stems are shared between different utterances of the listening passages. These referential overlaps may point to more uniform thematic foci of (parts of) passages, and greater cohesion of this nature may ease demands on the listening process.

Phonological complexity of the passages also correlated with task difficulty. More specifically, speakers' use of contractions was higher in easier tasks. Contractions are typical features of spoken grammar (Leech, 2000). With reference to the oral-literate continuum (Tannen, 1982), more spoken-like academic speech (i.e., containing typical spoken language features) is potentially easier to process for listeners than more written-like speech (i.e., lacking such spoken language features), but this needs further verification.

The lack of impact for input characteristics which have previously been found to affect L2 listening (e.g., speed, explicitness) cannot be ascribed to lack of variation, because there was sufficient variance in these characteristics between the 30 items of our study. Potentially, task differences between our study and other research may explain these conflicting findings.

Despite the need for more research, these findings may lead to some practical implications (in the first instance for the task investigated), for example, by informing text selection or text characteristic decisions at the task design or pretesting stage. However, due to the breadth of task characteristics and variety of passages we examined, a one-to-one relationship between each individual task variable and task difficulty is unlikely; manipulations of individual task characteristics would be expected to bring about changes in other characteristics (e.g., purposeful changes in argument overlap may result in simultaneous changes in lexical diversity). Therefore, further, more controlled task manipulation research, which isolates task variables, is recommended. Findings such as ours may guide initial variable selection for such experimental studies.

Listening task difficulty and task response. When assessing listening, we require learners to provide an answer to a task designed to make comprehension observable. Past research suggests that characteristics of the required response and the combination of input and response may contribute to task difficulty (see, e.g., Brindley & Slayter, 2002; Buck & Tatsuoka, 1998). Looking into PTE Academic SMW tasks, no relationship was found between task difficulty and the lin-

guistic complexity of the multiple-choice options (number of words and lexical sophistication). Our analyses also reveal that there was limited lexical overlap between passage and response options, and lexical overlap had no relationship with task difficulty. This contrasts with Jensen et al. (1997), who report increased difficulty when nonverbatim responses were needed. They suggest that item characteristics as opposed to text characteristics may be decisive for task difficulty and speculate that item writing guidelines may have evened out potential text characteristics effects. In the present study, the relatively limited lexical overlap may explain the results (and differences in item type with Jensen et al.), but measures taken at the item writing stage could similarly have cancelled potential response effects (which is reassuring when aiming to avoid construct-irrelevant variance). We acknowledge, however, that a restricted number of response characteristics were examined, and not all possible task input–response relationships were assessed. Similarly, limitations to the nature and size of our data set inhibited exploring potential interactions between response characteristics and their combined effect on task difficulty.

Given that our study is one of the first to consider the role of phrasal expressions in L2 listening and specifically in relation to listening task input and response characteristics and their interaction, one interesting finding is that less difficult tasks were those with more phrasal expressions in the necessary information. Furthermore, the majority of these phrases in the necessary textual information are classified as commonly occurring (Martinez & Schmitt, 2012). This finding corresponds to the role of multiword expressions we observed for the relationship between overall text characteristics and task difficulty. Three variables concerning common multiword expressions in the passages were associated with task difficulty; that is, texts with higher frequencies of these multiword expressions were less difficult. Given our results and estimates that 58% of spoken English discourse is made up of formulaic sequences (Erman & Warren, 2000), more research on the role of multiword units in L2 comprehension is desirable.

Listening Task Performance and Listener-Related Factors

Listening performance and listeners' working memory. The Chinese L1 participants who had higher listening scores on the PTE Academic were also those with higher PSTM and complex WM capacity. In this respect, our findings lend support to studies that have concluded that individual differences in WM predict L2 processing abilities. However, when focusing on one particular task type, the SMW

task, no correlation was found between listening performance and our WM measures. Potentially, task type may play a role in this, including differences in the nature of listening text and response characteristics and type of listening assessed by the task.

The SMW task constitutes a passage completion multiple-choice task, and, according to Pearson Education's (2012, p. 33) specifications, taps into the following listening subskills:

> identifying the topic, theme or main ideas; identifying words and phrases appropriate to the context; understanding academic vocabulary; inferring the meaning of unfamiliar words; comprehending explicit and implicit information; comprehending concrete and abstract information; following an oral sequencing of information; predicting how a speaker may continue; forming a conclusion from what a speaker says; comprehending variations in tone, speed and accent.

The PTE Academic listening scores, on the other hand, are based on performances on 11 different selected- and constructed-response task types, which also target additional listening subskills:

> identifying specific details, facts, opinions, definitions or sequences of events; identifying supporting points or examples; identifying a speaker's purpose, style, tone or attitude; classifying and categorizing information; summarizing the main idea; identifying the overall organization of information and connections between pieces of information; inferring the context, purpose or tone; identifying errors in a transcription. (Pearson Education, 2012, pp. 10; 14–16; 27–35)

As these lists illustrate, overall comprehension appears important for successful completion of the SMW task. Many other PTE Academic listening task types target this skill, too, but several also specifically assess local comprehension, for example, of specific details and supporting points. Understanding and retaining such details may put higher demands on WM.

A second observation is that the PTE Academic listening performances are based on items of 11 different task types with input texts that vary considerably (between task types)[2] in, amongst other things, passage length, number of speakers, genre, and text type. For example, some task types have longer passages and texts, characterised by more turn-taking, than the SMW task type. These features may place higher demands on WM and potentially explain the positive association between WM and PTE Academic performances. Thus, the targeted listening subskills and specific input text characteristics of some items and individual task types contributing to the PTE Academic listening

[2] Note that there is variation of these input characteristics between the items of each individual task type, but that this is not as vast as between items of different task types.

score may make higher demands on WM than do a collection of items of the SMW task.

Some similarities and differences in listening measures between our study and Andringa et al.'s (2012) and Kormos and Sáfár's (2008) studies are also notable. Andringa et al. assessed L2 listening through "traditional" multiple-choice (MC) tasks with a stem formulated as a question, followed by three sentence-length options. They found that auditory and visual forward and auditory backward digit span results correlated with L2 listening (albeit weakly). Kormos and Sáfár, who found a moderate-size significant correlation between L2 listening and backward digit span scores, measured L2 listening using a combination of multiple-choice and short-answer questions. The task types in both these studies are similar to some of those in the PTE Academic, but differ from the SMW task, which requires passage completion by selecting from MC options one or a few words in length and thus involving barely any reading. Potentially, reading and keeping in mind lengthier MC items during listening (as with Andringa et al.'s and Kormos and Sáfár's items) or production tasks (like those in Kormos and Sáfár) place higher WM demands on L2 listeners than the SMW type MC items.

Another observation is that, similar to the combination of listening tasks in the PTE Academic, the L2 listening tasks in Andringa et al. (2012) and Kormos and Sáfár (2008) target a large mixture of global and local listening skills such as listening for main ideas and for specific details. As mentioned above, however, the SMW task has a strong emphasis on global listening, which is probably less taxing for WM than deciphering and maintaining detailed information.

It should be emphasised, however, that the above-mentioned similarities and differences may only to some degree play an explanatory role in the partly conflicting findings of our research and those of Andringa et al. (2012) and Kormos and Sáfár (2008). Besides task type and subskill, other task- and listener-related variables might have also contributed to the similarities and differences observed.

Listening performance and listening anxiety. Higher levels of listening anxiety were found to correlate with lower L2 listening performance. In other words, less anxious listeners were also those who performed better on the SMW task and on the PTE Academic listening tasks. This confirms Kim's (2000) and Elkhafaifi's (2005) conclusions on the relationship between listening anxiety and L2 listening performance; that is, lower listening anxiety is associated with better L2 listening performance and vice versa. Importantly, our results also extend Elkhafaifi's and Kim's findings to a different type of listening task (SMW task), a variety of listening tasks (11 PTE

Academic task types), and ESL test takers from a variety of L2 backgrounds.

Although all three studies concern correlational analyses, the common negative association between listening anxiety and L2 listening performance strongly suggests that "listening anxiety has the potential to hinder efficient cognitive processing of the incoming aural input" (Kim, 2000, p. 153). As a consequence, from a theoretical point of view, listening anxiety appears to deserve a role in L2 listening modelling. At the same time, language educators and assessors may want to give more thought to ways in which listener anxiety can be limited or reduced. Indeed, Kim (2000) proposes a number of pedagogical implications, including level-specific and guided listening exercises with diagnostic feedback, broader recognition of listening as a process, and increased exposure to authentic listening materials.

LIMITATIONS AND FUTURE DIRECTIONS

While our study was large in scope, a number of issues need to be followed up in future research. Although focusing on one particular task allowed us to eliminate task-type effects, our conclusions may not apply to other task types. Moreover, although we examined the role of a wide range of text characteristics, our listening task sample size restricted us to looking into the relationship with one characteristic at a time. Thus, we were not able to explore potential interactions between multiple characteristics and their combined relationship with task difficulty. We cannot rule out that our findings on the presence or absence of significant associations between task characteristics and listening task difficulty resulted from complex interactions between different task variables. Finally, it is important to keep in mind that our use of correlational analyses indicates relationships, but does not lead to cause–effect explanations. Nevertheless, we believe that our findings can serve as a starting point for experimental task-manipulation research which might inform and improve practices in designing pedagogic and test tasks for L2 listening.

ACKNOWLEDGMENTS

This research was supported by the Pearson External Research Projects grants. We wish to thank Pearson for its financial assistance and for providing access to the PTE Academic Scored Practice Test. We also give special thanks to Chihiro Inoue and Janina Iwaniec for helping with data entry and Eivind Torgersen for assistance with phonetic analyses.

THE AUTHORS

Tineke Brunfaut is a lecturer in the Department of Linguistics and English Language at Lancaster University, in England. Her main research interests are language testing and reading and listening in a second or foreign language. In recent studies, she has looked into factors affecting second language listening task difficulty and factors that have an impact on first and foreign language academic reading proficiency.

Andrea Révész is a lecturer in the Department of Culture, Communication and Media at the Institute of Education, University of London. Her main research interests lie in the interface of second language acquisition and second language instruction, with particular emphasis on task-based language teaching and the roles of input, interaction, and individual differences in second language learning.

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*(Suppl. 2), 49–78. doi:10.1111/j.1467-9922.2012.00706.x

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*, 417–423. doi:10.1016/S1364-6613(00)01538-2

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47–90). New York, NY: Academic Press.

Blau, E. K. (1990). The effect of syntax, speed and pauses on listening comprehension. *TESOL Quarterly*, *24*, 746–753. doi:10.2307/3587129

Bloomfield, A., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2010). *What makes listening difficult? Factors affecting second language listening comprehension.* University of Maryland Center for Advanced Study of Language. http://www.dliflc.edu/file.ashx?path=archive/documents/CASL_study_FINAL_Lit_Rev.pdf

Boersma, D., & Weenink, P. (2008). *Praat: Doing phonetics by computer version 5.0.25.* Retrieved from http://www.praat.org

Brindley, G., & Slayter, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, *19*, 369–394. doi:10.1191/0265532202lt236oa

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

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*, 119–157. doi:10.1177/026553229801500201

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review*, *82*, 45–73. doi:10.1037/h0076248

Cervantes, R., & Gainer, G. (1992). The effects of syntactic simplification and repetition on listening comprehension. *TESOL Quarterly*, *26*, 767–774. doi:10.2307/3586886

Cobb, T. (n.d.). *Web VP classic*. Retrieved from http://www.lextutor.ca/vp/eng

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, *29*, 72–89. doi:10.1093/applin/amm022

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466. doi:10.1037/0278-7393.9.4.561

Elkhafaifi, H. (2005). Listening comprehension and anxiety in the Arabic language classroom. *Modern Language Journal*, *89*, 206–219. doi:10.1111/j.1540-4781.2005.00275.x

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, *20*, 29–62. doi:10.1515/text.1.2000.20.1.29

Freedle, R., & Kostin, I. (1996). *The prediction of TOEFL listening comprehension item difficulty for minitalk passages: Implications for construct validity* (TOEFL Research Report RR-96-29). Princeton, NJ: Educational Testing Service.

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*, 2–32. doi:10.1177/026553229901600102

Garcia, P. (2004). Pragmatic comprehension of high and low level language learners. *TESL-EJ*, *8*(2). Retrieved from http://www.tesl-ej.org/wordpress/

Gathercole, S. E. (1999). Cognitive approaches to the development of short-term memory. *Trends in Cognitive Sciences*, *3*, 410–419. doi:10.1016/S1364-6613(99)01388-1

Griffiths, R. (1992). Speech rate and listening comprehension: Further evidence of the relationship. *TESOL Quarterly*, *26*, 385–390. doi:10.2307/3587015

Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skills. *Studies in Second Language Acquisition*, *14*, 25–38. doi:10.1017/S0272263100010457

Henrichsen, L. E. (1984). Sandhi-variation: A filter of input for learners of ESL. *Language Learning*, *34*, 103–126. doi:10.1111/j.1467-1770.1984.tb00343.x

Horwitz, E. K. (2010). Foreign and second language anxiety. *Language Teaching*, *43*, 154–167. doi:10.1017/S026144480999036X

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*, 219–244. doi:10.1177/0265532208101006

Jensen, C., Hansen, C., Green, S., & Akey, T. (1997). An investigation of item difficulty incorporating the structure of listening tests: A hierarchical linear modeling analysis. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 151–164). Jyväskylä, Finland: University of Jyväskylä.

Kim, J. H. (2000). *Foreign language listening anxiety: A study of Korean students learning English* (Unpublished doctoral dissertation). Austin: University of Texas.

Kimura, H. (2008). Foreign language listening anxiety: Its dimensionality and group differences. *JALT Journal*, *30*, 173–195.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*, 363–394. doi:10.1037/0033-295X.85.5.363

Kormos, J., & Sáfár, A. (2008). Phonological short term-memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition*, *11*, 261–271. doi:10.1017/S1366728908003416

Kostin, I. (2004). *Exploring item characteristics that are related to the difficulty of TOEFL dialogue items* (TOEFL Research Report No. RR-79). Princeton, NJ: Educational Testing Service.

Leech, G. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning, 50*, 675–724. doi:10.1111/0023-8333.00143

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

MacIntyre, P. D., & Gardner, R. C. (1991). Methods and results in the study of anxiety and language learning: A review of the literature. *Language Learning, 41*, 85–117. doi:10.1111/j.1467-1770.1991.tb00677.x

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum.

Malvern, D. D., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Houndmills, England: Palgrave Macmillan.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*, 299–320. doi:10.1093/applin/ams010

Masoura, V. M., & Gathercole, S. E. (1999). Phonological short-term memory and foreign language learning. *International Journal of Psychology, 34*, 383–388. doi:10.1080/002075999399738

McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. (2005). *Coh-Metrix version 2.0*. Retrieved from http://cohmetrix.memphis.edu

Miyake, A., & Friedman, N. P. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. Healy & L. Bourne (Eds.), *Foreign language learning* (pp. 339–364). Mahwah, NJ: Lawrence Erlbaum.

Muljani, D., Koda, K., & Moates, D. R. (1998). The development of word recognition in a second language. *Applied Psycholinguistics, 19*, 99–113. doi:10.1017/S0142716400010602

Nissan, S., DeVincenzi, F., & Tang, K. L. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension* (TOEFL Research Report No. RR-51). Princeton, NJ: Educational Testing Service.

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

O'Brien, I., Segalowitz, N., Freed, B., & Collentine, J. (2007). Phonological memory predicts second language fluency gains in adults. *Studies in Second Language Acquisition, 29*, 557–582. doi:10.1017/S027226310707043X

Olsthoorn, N. M., Andringa, S., & Hulstijn, J. H. (2014). Visual and auditory digit-span performance in native and non-native speakers. *International Journal of Bilingualism, 18*, 663–673. doi:10.1177/1367006912466314

Paivio, A., Walsh, M., & Bons, T. (1994). Concreteness effects on memory: When and why? *Journal of Experimental Psychology: Learning Memory and Cognition, 20*, 1196–1204. doi:10.1037/0278-7393.20.5.1196

Pearson Education. (2012). *PTE Academic score guide*. Retrieved from http://pearsonpte.com/PTEAcademic/scores/Documents/PTEA_Score_Guide.pdf

Pollitt, A., & Hutchinson, C. (1987). Calibrated graded assessment: Rasch partial credit analysis of performance in writing. *Language Testing, 4*, 72–92. doi:10.1177/026553228700400107

Randall, M. (2007). *Memory, psychology and second language learning*. Amsterdam, the Netherlands: John Benjamins.

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

Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 287–318). Cambridge, UK: Cambridge University Press.

Rost, M. (2005). L2 listening. In E. Hinkel (Ed.), *Handbook of research on second language teaching and learning* (pp. 503–528). Mahwah, NJ: Lawrence Erlbaum.

Rost, M. (2011). *Teaching and researching listening* (2nd ed.). Harlow, England: Pearson Education.

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

Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing, 1*, 185–216. doi:10.1080/15305058.2001.9669470

Segalowitz, N. (2003). Automaticity and second language learning. In C. Doughty & M. Long (Eds.), *The handbook of second language acquisition* (pp. 382–408). Oxford, England: Blackwell.

Staehr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition, 31*, 577–607. doi:10.1017/S0272263109990039

Taguchi, N. (2005). Comprehending implied meaning as a foreign language. *Modern Language Journal, 89*, 543–562. doi:10.1111/j.1540-4781.2005.00329.x

Tannen, D. (1982). *Spoken and written language*. Norwood, NJ: Ablex.

Vandergrift, L. (2005). Relationships among motivation orientations, metacognitive awareness and proficiency in L2 listening. *Applied Linguistics, 26*, 70–89. doi:10.1093/applin/amh039

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching, 40*, 191–210. doi:10.1017/S0261444807004338

Vandergrift, L., Goh, C., Mareschal, C., & Tafaghodtari, M. H. (2006). The Metacognitive Awareness Listening Questionnaire (MALQ): Development and validation. *Language Learning, 56*, 431–462. doi:10.1111/j.1467-9922.2006.00373.x

Vogely, A. J. (1998). Listening comprehension anxiety: Students' reported sources and solutions. *Foreign Language Annals, 31*, 67–80. doi:10.1111/j.1944-9720.1998.tb01333.x

Yanagawa, K., & Green, A. (2008). To show or not to show: The effects of item stems and answer options on performance on a multiple-choice listening comprehension test. *System, 36*, 107–122. doi:10.1016/j.system.2007.12.003

Ying-hui, H. (2006). An investigation into the task features affecting EFL listening comprehension test performance. *Asian EFL Journal Quarterly, 8*(2), 33–54.

Zhao, Y. (1997). The effects of listeners' control of speech rate on second language comprehension. *Applied Linguistics, 18*, 49–68. doi:10.1093/applin/18.1.49