An investigation of integrative and independent listening test tasks in a computerized academic English test

This research provided a comprehensive evaluation and validation of the listening section of a newly introduced computerized test, Pearson Test of English Academic (PTE Academic). PTE Academic contains 11 item types assessing academic listening skills either alone or in combination with other skills. First, task analysis helped identify skills important for listening comprehension in academic settings. Aspects analyzed included the purpose of assessment tasks, skills/constructs assessed, and task stimuli employed in PTE Academic. The findings indicated that modern technologies enabled PTE Academic, a computer-based test, to assess students' academic listening abilities in real time using the integration of multi-modal sources. The statistical validation consisted of two stages. Exploratory factor analysis was performed first with a sample of over 5,000 students who took PTE Academic, to examine the underlying listening constructs as measured by the scores on the 11 item types item on different listening skills; these scores were subjected to Rasch analysis using CONQUEST. Second, the difficulties of the item types were estimated and the effectiveness of these item types was evaluated by calculating the information function by item type. The study has implications for test developers and test users regarding the interpretation of student performance on listening assessments.

Keywords: Computerized test, academic listening, integrative and independent tasks, construct validity

Introduction

The 21st century has witnessed a growth in the use of computer facilitated techniques (Buckingham & Alpaslan, 2017; Stickler & Shi, 2016; Wagner, 2007) and integrative skills items in international English language examinations, such as TOEFL iBT (Deluca, Cheng, Fox, Doe, & Li, 2013) and the Pearson Test of English Academic (PTE Academic). Both features have been widely believed to make listening assessment more valid from various perspectives. On the one hand,

Chappelle and other researchers (2006, 2016) stated that multimedia inputs, featured in computerized tests, can help portray some elements of authenticity of target language use situations in a test. They argued that computers can be used to display and process large amounts of data rapidly allowing for the input test takers receive on a language test to include rich contextual information consisting of images, sounds, and full-motion videos, potentially enhancing authenticity in both inputs and responses. On the other hand, integrated skills tasks or prototype tasks, which invite test takers to write a response to print or audio source texts, require a fuller understanding of the input from students, and therefore would adequately reflect their abilities in the areas that are required for academic success (Cumming, Grant, Mulcahy-Ernt & Powers, 2005; Yu, 2013; Zhu et. al. 2016). Although computerised tests have started to receive attention from researchers and test developers, the number of validation studies are still limited (Kim & Craig, 2012).

Most empirical studies on validating item types have been conducted in traditional paper and pen tests (Buck, 2001; Yu, 2013). The value of validating various item types in computerized tests is relevant to the current field of language testing and teaching for at least two reasons: first, embedding multimedia or internet-based resources in classroom language teaching and assessment has increased in popularity due to its potential to offer a higher level of authenticity, which can lead to higher levels of motivation and learning gains (Dixon & Hondo, 2014; Roman-Odio & Hartlaub, 2003). Some recent studies, however, argued that they may only appeal to language learners and students if they are perceived as easy to use and playful enough (Cigdem, Ozturk, & Topcu, 2016); Secondly, although some researchers expressed their concerns about language learners' over reliance on visual supports in listening

tasks, other researchers proposed updated solutions to evaluate the degree of reliance (Leveridge & Yang, 2013) and to reduce listener's dependence on captions (Li, 2014).

This study investigated construct validity, item difficulty and item effectiveness in relation to the use of integrative and independent listening items in a computerised high-stakes listening test. Academic listening skills are crucial in determining non-native English speaking students' academic success in English speaking environments (Berman & Cheng, 2001; Chang & Read, 2012; Jung, 2003). Listening skills, as one of the receptive skills, have been widely researched (Berne, 2008; Graham, 2006; Vandergrift, 2007). Wagner (2013) summarised seven areas that merit further investigation, among which at least three are directly relevant to the present study: 1) using integrated test tasks, 2) including unplanned spoken discourse in the spoken texts used in L2 listening assessment, and 3) the appropriateness of item types, response formats and the construct being tested.

Research Context

For the purposes of this investigation, data was gathered from PTE Academic, a computer-delivered and computer-scored English language test. This test is designed to measure the English language competency of test-takers in the four language skills – Speaking, Listening, Reading and Writing – for making admission decisions at tertiary level institutions and for organisations where English is the language of instruction. PTE Academic features 20 item types, reflecting different modes of language use, different response tasks and different response formats (Wang, Choi, Schmidgall & Bachman, 2012). Each item type assesses one language skill or a combination of language skills, representing the range of functions and situations that students will encounter in academic studies in an English-speaking environment.

PTE Academic reports scores on the Pearson's Global Scale of English (GSE), ranging from 10 to 90. The PTE Academic score report includes 11 scores on the GSE, they are, an Overall Score, four Communicative Skills scores and six Enabling Skills scores. The Overall Score reflects test takers' overall English language ability. The score is based on performance in all items in the test. Scores for Communicative Skills (Listening, Reading, Speaking, and Writing) are based on all test items (tasks) that assess these skills, either as a single skill or together with other skills. Scores for Enabling Skills (Grammar, Oral Fluency, Pronunciation, Spelling, Vocabulary and Written Discourse) are based on test items assessing one or more of these skills. The GSE scores have been empirically designed and developed to align with the Common European Framework of Reference (CEFR) for languages.

The listening skill in PTE Academic is tested using 11 integrated and independent items, including highlighting correct summary, multiple choice, choosing single and multiple answers, filling in the blanks, highlighting incorrect words, selecting the missing word, writing from dictation, summarising spoken text, repeating a sentence, retelling lecture and answering a short question. Among the 11 item types, five use audio input: fill in the blanks, write from dictation, summarise spoken text, repeat sentence and answer short question. Another five use a combination of audio and video input: highlight correct summary, MCQ with single/multiple answer (s), select missing word and retell lecture. The last item type, highlight incorrect words, presents test takers with both transcripts (visual aids) and audio inputs. Table 1 below summarises the tasks and sub-skills being tested in relation to the 11 item formats. Column 1 consists of item type labels and codes, column 2 describes what test takers need to do to complete each task. Column 3 lists the types of listening skills these item types were designed to measure. The item type

codes in column 1 consist of three parts: number in part one is the item type sequence number; the second part contains the initials of the skills assessed, for example, LR indicates it is an integrated item type designed to measure listening skill and reading skill integratively, while LL indicates it is an independent item type designed to measure listening skill only. The third part of the code briefly summarises what test takers are asked to do. For example, HILI, indicates that in this item type, test takers need to highlight their choice to complete the task, while MAMC indicates it is a multiple answer multiple choice question type.

Table 1: 11 listening tasks in Pearson Test of English Academic

Item type	**Task description**	**Listening skills tested***
Highlight correct summary (06-LR-HILI)	After listening to the audio/watching the video and reading the alternatives, test takers select the paragraph that is most specific to the audio/video.	Understand vocabulary, comprehend pronunciation, comprehend information, classify information, Identify structures
MCQ with single/multiple answer (s) (09-LL-SAMC) (10-LL-MAMC)	After listening to the audio/watching the video and reading the alternatives carefully, test takers select the option(s) that best answers each question.	Identify and summarise structures, identifying speaker's purpose, making connections between pieces of information, making inferences, generalisations or conclusions
Fill in the blanks (11-LL-GAPS)	Test-takers listen to the audio and complete the gapped written text by typing the missing word in each gap.	Understand vocabulary, identify words and phrases appropriate to the context; classify information
Highlight incorrect words (12-LR-HOTS)	Test takers see a reading text on screen. While listening to the audio, the test takers click on all the 'hotspot' words which differ from what they have heard.	Understand vocabulary, comprehend pronunciation, identifying errors in a transcription, classify information
Select missing word (13-LW-GAPS)	Test takers need to listen to a recording/watch a video where one or more words are replaced by an electronic beep.	Understand vocabulary, comprehend pronunciation, comprehend information, classify information, identify structures

Write from dictation (14-LW-DICT)	Whilst listening to the audio, test takers transcribe what is spoken and type the exact sentence in the space provided.	Understand vocabulary, comprehend pronunciation, classify information
Summarise spoken text (15-LW-SUMM)	Test takers listen to the audio recording, and then write a summary of what the speaker has said.	Understand vocabulary, classify information, identify and summarise structures
Repeat sentence (16-LS-REPT)	After hearing the sentence, test takers repeat the sentence exactly as they hear it.	Understand vocabulary, comprehend pronunciation
Retell lecture (20-LS-PRES)	Test-takers hear an audio recording/watch a video, and retell what they have just heard/watched in their own words	Understand vocabulary, comprehend pronunciation, comprehend information, classify information, identify structure
Answer short question (21-LS-SAQS)	Test takers answer the question with a single word or a short phrase.	Understand vocabulary, identify structure

*The information on skills being tested by different item types is quoted from *The Official Guide to PTE Academic*.

Relevant Literature

This section reviews the previous studies and discussion on the relationships between the construct of listening comprehension in an academic context, the use of integrative and independent listening items, item difficulty level, item effectiveness and recent understanding of embedding visual and multimedia input in language listening assessment tasks.

To start with, the construct of academic listening has been defined by previous researchers from three main perspectives: 1) a comprehensive list of skills, knowledge or abilities (Buck, 2001), 2) a list of tasks that language learners are expected to complete (Buck, 2001) and 3) the listening comprehension procedure (Vandergrift, 2007; Weir, 2005), which depicts the listening process as goal-setting or planning, acoustic/visual input, audition, pattern synthesis and monitoring. Another way to

interpret listening comprehension procedure is what is called the "hierarchical view" (Anderson & Lynch, 1998; Buck, 2001; Cai, 2012), in which listening comprehension is believed to be facilitated by information sources at three levels: systemic knowledge (language), context (co-text and situation) and schematic knowledge (background and procedural knowledge). This three-level sources of information appear to reflect Vandergrift's (2007) understanding of listening processes: bottom-up, top-down and bi-directional listening. Addressing the relationship between the three proposed listening strategies, a more recent study argued that bottom–up strategies do not exert direct effects on listening comprehension, but must be mediated by top–down strategies (Nix, 2016). Table 2 combines the information of 11 item types in PTE Academic listening tasks and targeted sub-listening skills being tested based on the hierarchical view. Second language learners with higher listening ability can be defined as those who are capable of using different listening skills to engage with various forms of input and demonstrate their comprehensions at appropriate levels. The last two columns of the table highlight what level of listening skills each item type aims to assess from a test developer's point of view (see The Official Guide to PTE Academic). For example, in the "Language" column, "strong" suggests that the corresponding item type has been specifically designed to assess listening skills at the language level. Meanwhile, in the column of "co-text & situation" (Anderson & Lynch, 1998; Buck, 2001; Cai, 2012;), "strong" suggests that the corresponding item type has been specifically designed to assess listening skills at the co-text and situation level and "weak" indicates that the corresponding item type has not been designed to assess listening comprehension at co-text and situation level.

Table 2: Item types and level of listening skills being tested

Item type	**Integrative fashion**	**Level of skills being tested**

		Language	Co-text & situation
Highlight correct summary (06-LR-HILI)	Listening/ Reading	strong	strong
MCQ with single/multiple answer (s) (09-LL-SAMC) (10-LL-MAMC)	Listening	strong	strong
Summarise spoken text (15-LW-SUMM)	Listening/ Writing	strong	strong
Retell lecture (20-LS-PRES)	Listening/ Speaking	strong	strong
Answer short question (21-LS-SAQS)	Listening/ Speaking	strong	strong
Fill in the blanks (11-LL-GAPS)	Listening	strong	weak
Highlight incorrect words (12-LR-HOTS)	Listening/ Reading	strong	weak
Select missing word (13-LW-GAPS)	Listening/ Writing	strong	weak
Write from dictation (14-LW-DICT)	Listening/ Writing	strong	weak
Repeat sentence (16-LS-REPT)	Listening/ Speaking	strong	weak

Interestingly, Buck (2001) observed that listening assessment has developed from the view of listening as recognising different linguistic elements, through perceiving listening as language processing, to the more current idea of listening as interpreting meaning in a communicative context. Some more recent studies recommend the inclusion of test takers' capabilities in interpreting and understanding nonverbal and visual information in the listening construct (Wagner, 2008; Cubilo & Winke, 2013). Most of the validation evidence in listening tests seems to be relevant to the skills, abilities and knowledge involved in listening comprehension, or the tasks that the test takers are expected to complete. Recognising the limitations of assessing the individual elements of listening comprehension, this approach has been extended as a result of recent debates over the validity of high-stakes language tests which include both integrative and independent listening tasks. Item types with more

integrative characters, such as dictation, statement evaluation, summarising spoken input – have been developed and employed in some high-stakes tests. These item types are designed to extend the assessment of listening from the discrete-point approach to one of assessing listening through language processing and interpretation (Buck, 2001) thereby better reflecting academic literacy activities, representing a higher level of authenticity, and reducing certain unwanted effects of test methods that are often associated with conventional item types (Yu, 2013).

Regarding the less integrative item types, Powers (1985, p. 9) states that the task of 'matching or distinguishing,' i.e., choosing a picture to correspond with what was heard, was seen as the least appropriate assessment task. The same task using a written response, however, was rated as somewhat more appropriate. The 'matching/distinguishing' task was viewed as "too low level", "lacking content", "never encountered" and "too elementary". Multiple choice item types are regarded as "information transfer" tasks. Brown (2004) comments that in answering multiple choice questions, the skill needed is a technique in which "aurally processed information must be transferred to a visual representation, such as labelling a diagram, identifying an element in a picture, completing a form, or showing routes on a map"(2004. p. 127). Brown (2004) suggests that this task "may reflect greater authenticity by using charts, maps, grids, timetables and other artefacts of daily life"(p. 128).

In reviewing 'fill in the blanks' and 'highlight incorrect words' item types, aural cloze questions were viewed as inappropriate because a listening cloze test would be too confusing and difficult for non-English native students. Potential problems with scoring on a large scale and the likely difficulty of correctly keying answers were also mentioned. In addition, it is believed that cloze tests entailed

greater production (writing) skills than receptive (listening) skills (Powers, 1985). Buck (2001) speculates that a listening cloze test would make a good listening test, as "test takers would clearly have to understand a short piece of spoken language, at least on the linguistic level, even if they did not have to apply this to a communicative situation"(2001. p. 69). The item type 'select missing word' is regarded as more likely to test word-recognition skills than general listening ability (Buck, 2001), as the completion of gap-filling tasks can often be processed on a perceptual or a phonological level without bringing to mind the actual meaning of the words, which would not constitute comprehension.

As to those items that require more integrative language skills, previous literature paints a rather mixed picture. From EFL teachers' and learners' perspectives, previous studies (e.g., Carrell, Dunkel & Mollaun, 2004; Cumming, Grant, Mulcahy-Ernt & Powers, 2005) suggest that they viewed "positively the new prototype tasks that required students to write or to speak in reference to reading or listening source texts"(Cumming et.al., 2005, p.2). In particular, Brown (2004) reviews the 'dictation' item type and summarises that when the passage is not long, dictation seems to provide a reasonably valid method for integrating listening and writing skills and for tapping into the cohesive elements of language implied in short passages. However, he also warns that only a moderate degree of cognitive processing is required and therefore claiming that the dictation fully assesses the ability to comprehend pragmatic or illocutionary elements of language, context, inference or semantics may be going too far.

Powers (1985), however, suggests dictation to be the least appropriate question type. The most frequently cited reason for the inappropriateness of dictation exercises is that students do not typically need to copy lecturers' verbatim, but rather

they need to comprehend and organise materials. Dictation was seen as something that simply is not ordinarily required by students (Powers, 1985). 'Transcribing' was generally seen as a low-level skill that does not adequately reflect the abilities required for academic success, or it was felt that success in this task does not reflect a student's level of understanding. Buck (2001) echoes that dictation assesses listening comprehension on a local, literal, linguistic level, and that when the text is longer, what dictation assesses is "the ability to recognise simple elements" (p. 78).

In reviewing the 'repeat sentence' item type, Buck (2001) questions the skills/construct actually being assessed. He states that if the sentence is short, this task is likely to assess the ability to recognise and repeat sounds, while if the sentence is longer, length of working memory (rather than cognitive listening skills) seems to be what is evaluated. When the sentences become even longer, it seems likely that chunking ability and the ability to deal with reduced redundancy will begin to become important. In reviewing the 'summarize spoken text' and 'retell lecture' item types, Powers (1985) suggests that answering questions involving recall of details and those involving inferences and deductions were viewed as somewhat more appropriate than were the other tasks mentioned, as was condensing, i.e. being able to reduce what is heard to an outline of main points. In terms of 'answer short question', the discussion found in the literature is on sentence evaluation tasks. Buck (2001) perceives the underling cognitive process as "basic sentence-processing skills" rather than a communicative academic listening skill.

In terms of evaluating the difficulty levels of various listening tasks, Revesz and Brunfaut (2013) examined whether the difficulty of an L2 listening task is affected by the speed of delivery, linguistic complexity, and explicitness of the input text, and by the characteristics of the textual information necessary for task

completion. One of their findings indicate that speed of delivery was not found a significant predictor as suggested by previous research, and they suggested looking into compounding factors, such as task types and task conditions. In addition, Brunfaut and Résvez (2015) further looked at 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. They also explored listener characteristics in relation to task difficulty and performance.

With respect to the relative level of difficulty of integrated skills tasks, no agreement has been reached. On the one hand, some researchers reported that different types of responses had a significant effect on test takers' performance. In comparing multiple choice questions and open-ended questions, in general, questions with constructed response formats are deemed difficult. Cheng (2004) maintained that the guessing factor and the availability of clues for prediction were two reasons that may lead to multiple choice questions being easy, and memory constraints imposed by open-ended questions made them difficult. On the other hand, Brindley and Slatyer (2002) maintained that the complexity of the interaction between text, item and response made it difficult to isolate the effects of specific variables, and the particular combinations of item characteristics appear either to accentuate or attenuate their effect on difficulty. They stated that there are three factors that may impact the difficulty level of a listening task: the nature of the input, the nature of the assessment task and the individual listener factors. To be specific, according to them, the variables that define the nature of the listening input include speech rate, length of the listening passage, syntactic complexity, vocabulary range, discourse structure, noise level, accent, register, propositional density and amount of redundancy. The following elements define the nature of the assessment task: amount of context provided, clarity

of instructions, response format, availability of question preview, amount of lexical overlap between the listening passage and the response format, length of text preceding the information requiring a response, length of required response, repetition of tested information, whether responses and repetitions of information are verbatim or paraphrases.

In addition to validity arguments, i.e. to what extent one test item type represents the language domain that test developers aim to measure, some test developers, especially those involved in psychometric analysis, are also interested in knowing how efficiently the test items perform when measuring the test takers' language ability with precision. Test item information function, also known as Fisher function, states that information is reciprocal to the precision with which a parameter could be measured (Baker, 2001). When we speak of having information, we imply that we know something about a particular object or topic. In testing, the term 'information' is used to mean that the test takers possess the required knowledge about the subject under focus. If a parameter is accurately estimated, more information will be acquired as to the numerical value of that parameter than when the same parameter is less accurately estimated (Baker, 2001).

Finally, in the context of computerized listening assessment and instruction, the extent to which different types of input impact on listening comprehension level has been investigated extensively. The possible features which computerized listening tests can adopt include: various forms of subtitles and audio/video input. Three types of subtitles have been investigated with mixed results: full captioning, keyword captioning and no captioning. Perez, Peters and Desmet (2014) concluded that the full captioning group outperformed the other two groups on the global comprehension questions, with no significant difference being identified between the keyword

captioning and the no captioning group. Furthermore, the results of the detailed comprehension questions (with audio) revealed no differences between the three conditions. In contrast, Chen and other researchers (2012) argued that the presence of transcripts may help listeners to understand spoken English input better (as evidence in the immediate recall tasks) in a short period of time. But they are unlikely to construct schema knowledge and help listeners to complete similar subsequent listening tasks. In addition, the language of subtitles matter, as studies (Ghoneam, 2015; Hayati & Mohmedi, 2009) confirmed that L2 listeners achieve higher comprehension level when they are presented with English subtitles than subtitles in their first language.

With regards to the relationship between types of input and listening comprehension level, Jones and Plass (2002) found that a combination of pictorial and written annotations are the more effective in terms of assisting students to recall the short language audio clips than offering written and pictorial assistance separately. Wagner (2013) concluded that the test takers who received audiovisual input scored higher than those who only received audio input. Exploring the reasons behind the increased comprehension level under the condition of presenting audiovisual input, Taguchi (2016) argued that multimedia input can present both linguistic and nonlinguistic clues which can increase L2 listeners' comprehension of indirect meaning. The form of audio input also matters, as some studies (Papageorgiou et.al., 2012) pointed out that items associated with dialogic input are easier than those linked with monologic input. Moreover, in investigating the effects of providing subtitles and taking notes on listeners' cognitive load and performances, Lin and her colleagues (2016) concluded that the availability of animation with subtitles can reduce cognitive load and increase performance.

To summarise, advances in computer technology have equipped test developers with better tools for developing a wider range of listening tasks. However, the potential benefits of simultaneous visual and auditory input and the integrative items still require further investigation (Vandergrift, 2007). Moreover, not many studies have provided validation evidence for a listening test with both discrete and integrative items. The present study, therefore, examined the nature of the listening inputs and assessment tasks in the listening part of PTE Academic, and explored how these factors influenced the academic listening constructs examined, from the perspectives of difficulty levels and task effectiveness. The study addressed the following two research questions:

1) Are integrative and independent items measuring the same listening construct(s) in a computerized test? And how do integrative and independent items contribute to the overall listening score?
2) How do these item types perform in measuring academic listening skills in terms of item type difficulties and item type effectiveness?

Methodology

The data for analysis in this study contain 5697 test takers' records provided by the test developers to aid this project. These data were retrieved randomly from the test developers' item bank in 2013. Since these were secondary data, no ethical approval were sought. Only the data that were pertinent to the purpose of this project were provided by the testing organisation. Among the test taker data provided, there are 3433 male and 2264 female students coming from 111 countries. The top ten first languages of the test takers are presented in Table 3.

Table 3: Test takers' first languages

Test taker's first language	N
Chinese-Mandarin	405
Tagalog	387
Urdu	342
Hindi	270
Bengali	212
Arabic	165
Nepalese	149
Tamil	145
French	138
Korean	130

To examine the underlying listening construct(s) as measured by the 11 item types, Exploratory Factor Analysis (EFA) was performed by IBM SPSS Statistics 22. Maximum likelihood with direct oblimin rotation was performed and the number of sub-constructs was evaluated by the eigenvalues and scree plot. Moreover, the predictive power of each item type for the final overall listening score is calculated by multiple regressions using the stepwise method, with test takers' scores in the 11 item types being treated as the independent variables and test takers' final listening scores as the dependent variable. Stepwise regression begins with an empty model and adds variables in order of importance for prediction. It first selected the best predictor. In the second step of selection, by partialling out all variables already included in the previous regression equation, it calculates the unique relevant variance that the next variable made.

To examine item type difficulty and item type effectiveness, item scores were subjected to a Rasch analysis using a piece of software called CONQUEST developed by Australian Council for Educational Research (ACER). The difficulties of the item types were estimated from the item response theory perspective. Delta values, which range from -3 to +3, were used to gauge the item difficulty levels.

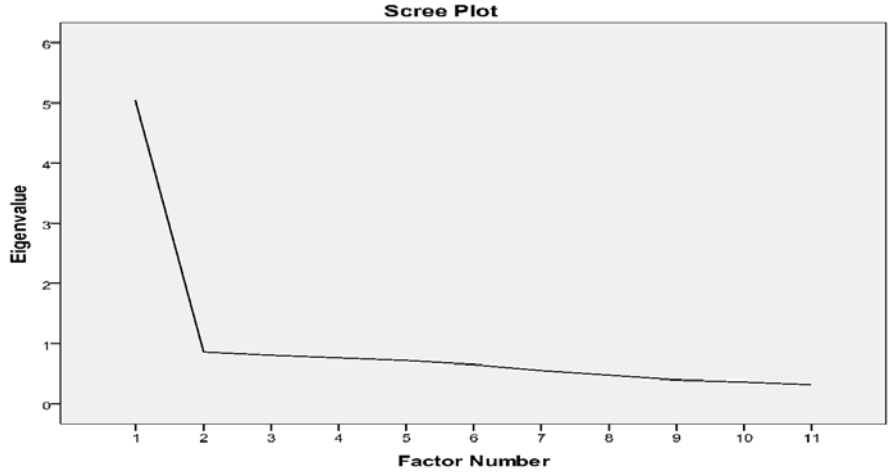
Test item effectiveness is usually approached from item discrimination statistics at each item level, or by looking for multiple choice distracters' effectiveness, i.e. how effective are the incorrect options in distracting higher and lower scorers statistically. In this study, the effectiveness of these item types was evaluated by calculating the information function of each item type by the time allocated to each item type, where all items examined in each particular item type were calculated and averaged to get the average item type information function.

Results

Examining Academic Listening Contruct(s)

In examining the underlying listening construct(s), statistical analyses were performed. Firstly, EFA of all listening item types was conducted to explore the number and nature of the underlying construct(s). Multiple regression was then conducted to examine the degree to which item type scores predicted the overall listening factor score. The EFA results indicated that there is a unidimensional factor underlying these item types, which accounts for 46% of the variance (see Figure 1 for the Scree plot).

Figure 1. Exploratory Factor Analysis of the 11 listening item types



Subsequently, factor scores were generated for the 11 items. These factor scores were then used as dependent variables to explore which scores from which item type might act as the best predictors of test takers' performance in terms of the overall listening factor. The Stepwise method was used. As shown in Table 4, the 11 item types' scores all made their relative positive contributions to the overall listening factor score. Contrary to what might be expected, the three independent listening item types: 'fill in the blanks' and two types of 'multiple choice questions' are not the best predictors of the underlying listening construct. The top six predictors are all integrative item types.

Table 4: Multiple Regressions: 11 listening item types in predicting overall academic listening construct

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Integrative / Independent
	B	SE	Beta (*β*)			
Write from dictation (14-LW-DICT)	0.20	0.00	0.21	3.28	0.00	integrative
Repeat sentence (16-LS-REPT)	0.20	0.00	0.21	3.19	0.00	integrative
Select missing word (13-LW-GAPS)	0.19	0.00	0.20	3.18	0.00	integrative
Summarise spoken text (15-LW-SUMM)	0.15	0.00	0.16	2.69	0.00	integrative
Highlight incorrect words (12-LR-HOTS)	0.14	0.00	0.15	2.53	0.00	integrative
Answer short question (21-LS-SAQS)	0.11	0.00	0.12	2.16	0.00	integrative
MCQ - multiple answers (10-LL-MAMC)	0.06	0.00	0.07	1.35	0.00	independent
Retell lecture (20-LS-PRES)	0.09	0.00	0.09	1.71	0.00	integrative

MCQ - single answer (09-LL-SAMC)	0.06	0.00	0.06	1.25	0.00	independent
Highlight correct summary (06-LR-HILI)	0.05	0.00	0.05	1.16	0.00	integrative
Fill in the blanks (11-LL-GAPS)	0.05	0.00	0.05	1.15	0.00	independent

To be specific, ‘Dictation’ and ‘Repeat sentence’ were the two best predictors of the overall academic listening construct ($\beta = .21$, $p < .01$). To illustrate, the results suggested that one unit increase in the ‘dictation’ score or ‘repeat sentence’ score would result in .21*unit in the overall listening factor score. The next best predictor was ‘select missing words’ ($\beta = .20$, $p < .01$), which indicated that in the presence of ‘dictation’ and ‘repeat sentence’, ‘select missing words’ would contribute the greatest amount of unique relevant variance to the regression equation. A one unit increase in the ‘select missing words’ score would result in an increase of .20*unit in the overall listening factor score. Subsequently, the standardised coefficients decreased in the following order: ‘summarise spoken text’, ‘highlight incorrect words’, ‘answer short questions’, ‘multiple choice multiple answer’, ’retell lecture’, ‘multiple choice single answer’, ‘highlight correct summary’ and finally ‘fill in the blanks’. To be specific, each of the above beta weights denotes a unique contribution of that variable while partialling out all of the previously entered independent variable(s) in the equation. Interestingly, from a “hierarchical view” of listening comprehension, the top three predictors are targeting the language level rather than the situation or co-text level of the audio input. Meanwhile, four out of the five item types with the smallest predictive power assess listening comprehension at the or co-text or situational level.

Item Type Difficulties

Item facility indices were calculated to evaluate item type difficulties, where the difficulty estimates of all items in each item type were calculated and averaged. Table 5 below shows the results from the IRT analysis. Four of the 11 item types are dichotomously scored item types, they are, 06-LR-HILI, 09-LL-SAMC, 11-LL-GAPS, and 21-LS-SAQS, where only one delta estimate for each item type is generated. The other seven item types are polytomously scored, where the number of delta estimates equals the number of maximum score points minus one. To be able to compare the item difficulties, average delta estimates were calculated for each polytomously scored item type.

Table 5: Item difficulty - Delta estimates

Listening item type	**Independent/ Integrative**	**Delta**	**Language / Co-text level**
Summarise spoken text (15-LW-SUMM)	integrative	0.86	Language + Co-text
MCQ with multiple answers (10-LL-MAMC)	independent	0.83	Language + Co-text
Answer short question (21-LS-SAQS)	integrative	0.24	Language + Co-text
Write from dictation (14-LW-DICT)	integrative	0.22	Language
MCQ with single answer (09-LL-SAMC)	independent	0.21	Language + Co-text
Select missing word (13-LW-GAPS)	integrative	0.15	Language
Retell lecture (20-LS-PRES)	integrative	0.13	Language + Co-text
Fill in the blanks (11-LL-GAPS)	independent	0.01	Language
Highlight correct summary (06-LR-HILI)	integrative	0.00	Language + Co-text
Highlight incorrect words (12-LR-HOTS)	integrative	-0.34	Language
Repeat sentence (16-LS-REPT)	integrative	-0.49	Language

Note: delta estimates range from -3 to +3, with lower values indicating lower item difficulty levels

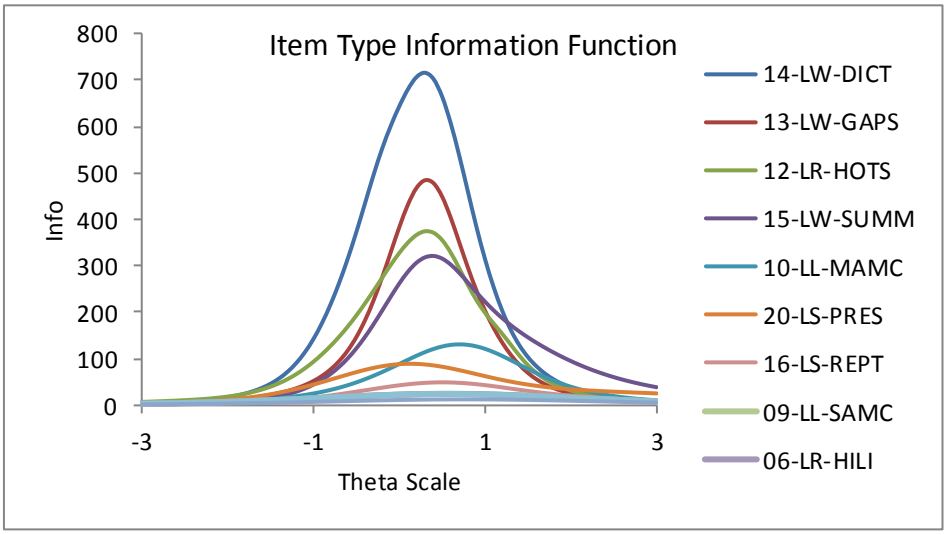
The results indicated that ‘summarise spoken text’ had the highest delta (0.86), followed by ‘multiple choice multiple answer’ (0.83) and ‘answer short question’ (0.24). It is not convincing to conclude that the difficulty level of item types relates to the inclusion of integrative characters, since both independent and integrative items stand as the 5 most difficult item types. Meanwhile, it seems to suggest that those item types that require test takers to interpret information at both the language and co-text levels of audio inputs show a higher difficulty level than the item types that target test takers’ abilities at the language level alone. For instance, among the five most difficult item types, four of them assess candidates’ abilities in interpreting information at both language and co-text levels. While three out of the five least difficult item types mainly target at the language level except for ‘highlight correct summary’ and ‘retell lecture’.

Item Type Information function

Item information function for each item type was plotted against the test taker ability scale, i.e. the theta scale (see Figure 2). ‘Write from dictation’ (14-LW-DICT) had the highest information peak in a range of ability estimate of theta -1 to theta +1, followed by ‘select missing word’ (13-LW-GAPS), and ‘highlight incorrect words’ (12-LR-HOTS).

Theta -1 corresponds roughly to 36 on the Pearson's Global Scale of English (GSE), which is estimated to be in the CEF A2 range. Theta 0 corresponds roughly to 53 on the GSE, which is estimated to be in the CEF middle B1 range. Theta +1 corresponds roughly to 72 on the GSE, which is estimated to be in the CEF upper B2 range. As indicated, these 11listening items in PTE Academic provide more information in the range of CEF A2 to B2, which is the range that PTE Academic is designed to target.

Figure 2. Item information per item type



In any testing situation where testing time is fixed and always limited, it is of vital importance for test developers to configure a test where all test items can provide optimal information about test takers' abilities. This study approached this inquiry by looking into test information function per second.

The first column in Table 6 shows the average item information for100 items across the 11 item types. The second column indicates the allocated time per item in each item type as measured in seconds. The third column shows the information per second. The higher the amount of information, the more effective each item type is indicated to be. As can be seen, the dictation item has the highest information per second (3.55), followed by 'highlight incorrect words'(1.74), and 'select missing word'(0.95). 'Repeat sentence' (0.87) is the fourth most effective item type from this analysis.

Table 6. Item information function by item type

Item types	Average Information of 100 items	Allocated Time/item (seconds)	Info/second

14-LW-DICT	195.00	55	3.55
12-LR-HOTS	112.90	65	1.74
13-LW-GAPS	113.44	120	0.95
16-LS-REPT	21.63	25	0.87
10-LL-MAMC	47.38	120	0.39
21-LS-SAQS	7.54	23	0.33
20-LS-PRES	41.97	135	0.31
15-LW-SUMM	107.89	600	0.18
11-LL-GAPS	14.54	80	0.18
09-LL-SAMC	14.59	85	0.17
06-LR-HILI	14.56	150	0.10

Discussion

The findings of this study indicate that despite the variation in the measuring modes/formats, there is one overall underlying academic listening construct that these item types are designed to measure. The item type analysis indicated that among the 11 item types used in assessing listening, three item types are designed to assess listening independently, two item types are designed to assess listening together with reading, three item types are designed to assess listening together with writing, and the remaining three item types assess listening together with speaking. Although the 11 items seem to assess the same underlying construct, this study concludes that the combination of different items presents a more comprehensive measure of test takers' ability, measuring different levels of skills (local, literal, sentence level, linguistic level and communicative level) as they are all parts of the unified listening construct. Both literature and experience inform us that there is no single technique or method that can be claimed to be the best way to assess listening comprehension. As Buck (2001) pointed out, it is the test developers' responsibility "to consider the needs of the situation and determine which construct and which operationalisation of that construct will be the most suitable for their purpose"(p. 93). The use of 11 item types can reduce or cancel out any potential biases that may be caused by the interaction

between test taker background and test item formats, therefore achieving construct validity to a certain extent.

In the context of PTE Academic, the listening skills that are designed to be tested cover a wide range, which aims to reflect the listening skills test takers are required to demonstrate in real life. Therefore, a range of different item types that uses different modes and methods of testing is justified as covering a wide domain. The results of the study demonstrate that computer technologies have enabled PTE Academic to facilitate the assessment of academic listening using the integration of multi-modal sources. Chappelle and Douglas (2006) stated that multimedia, featured in computerised tests, can help portray some elements of authenticity in terms of target language use situations in a test. They argued that computers can be used to display and process large amounts of data rapidly allowing for the test takers to receive, as input within a language test, rich contextual information consisting of images, sounds, and full-motion videos, potentially enhancing authenticity in both inputs and responses.

Secondly, as demonstrated by the multiple regression results, those item types which assess lower levels of skills (language level) in an integrative fashion have higher predictive power than other item types that target both the language and co-text levels. In other words, those test takers who are more capable of building up their listening comprehension from the language level to complete writing or speaking tasks are more likely to obtain a higher score in this computerised listening test. This finding may be explained from two points of view: firstly, it reflects some of the features that can be embedded more easily in computerised tests, such as greater reflection of the authentic tasks (Chappelle & Douglas, 2006) and cognitive learning strategies (Carrell, Dunkel & Mollaun, 2004, Yu, 2013) in an academic setting. For

example, in a project of validating a new speaking and writing prototype task in a new version of computerized test, Cumming and his colleagues (2005) argued that the introduction of an integrative skills task can be seen as improving the test's level of authenticity and helps construct validity. Secondly, this finding offers further support for the idea that test takers might use more bottom-up or language-related strategies to deal with the integrated skills tasks (Plakans & Gebril, 2013), since high-performing test takers tend to demonstrate high-level cognitive processes, such as reformulating and reproducing information (Frost, Elder & Wigglesworth, 2012), and other regulation skills for managing reading, listening, and writing interactions (Yang & Plakans, 2012). In investigating the two types of computerized PTE Academic integrative skills tasks, Rukthong and Brunfaut (2015) also found that their participants had reported more 'lower-level processes' (input decoding and syntactic parsing) than 'higher-level processes' (such as structural building and integrating/linking pieces of information).

Thirdly, the results of the comparison of statistical features of the item types indicate that the 11 item types have varying average item difficulties and the association between difficulty level and item types is not supported. This result is not surprising considering that researchers consistently consider item format as one of the salient factors affecting item difficulty (e.g., Brindley & Slatyer, 2002). However, the findings appear to support another conclusion, that is, since different item formats have differing processing demands (e.g., linguistic and/or cognitive demands) on test takers, the level of skills being assessed can partly explain the difficulty level. For instance, the two easiest item types in this study are 'highlight incorrect words' and 'repeat sentence', which only require test takers to demonstrate word recognition at language level; while the two most difficult item types, 'summarise spoken text' and

'MCQ with multiple answers', are designed to assess test takers' ability to comprehend audio input at both the language and co-text levels. The results also revealed the varying item type information function. The dictation item type seems to possess the highest test information per second, indicating that given the same test time, compared to other item types, dictation can provide more information on test takers' academic listening ability in a test situation.

Conclusions and Implications

To conclude, this study examined the 11 item types employed in PTE Academic for assessing listening either independently or integratively. The underlying listening construct was analysed and its relevant item types were evaluated in terms of item type difficulty and item type effectiveness. To summarise the answers to the research questions: first of all, in the computerised listening tests, items with both integrative fashion and different item response format appear to assess one construct, academic listening, and they vary in their contribution to the overall listening score; secondly, there is no evidence suggesting that difficulty levels relate to item formats, however, this study speculates that the difficulty level of listening tasks may depend on the skills and abilities being assessed in tests. The examination of item type information function demonstrates that these item types vary greatly in the information function per second.

The finding of this study has implications for the teaching of academic listening skills to learners of English. Teaching can never be separated from testing and findings from testing research can always inform or help improve teaching practices. Integrative listening items in computerized listening tests can be highly recommended to EFL teachers, especially when their students' listening skills are relatively low. As

demonstrated in this study, from a test takers' or learners' perspective, the introduction of integrative listening items in a computerized environment can offer greater opportunities to practice their bottom-up listening skills. Moreover, since the combination of integrative and independent items seem to assess the same academic listening construct, teachers could design their classroom activities to improve students' listening comprehension by focusing on different aspects of listening or using more integrative tasks. Furthermore, depending on specific assessment requirements and availability of time, different listening item types can be used for different purposes. For example, since item types which assess lower levels of skills (language level) in an integrative fashion could have higher predictive power, EFL teachers, who are interested in designing effective replacement tests, are recommended to assign more computerized listening tasks in the form of dictations, repeat sentences and summaries of spoken text. In circumstances where a shortened listening computerized assessment is needed, item types that provide more test information should be opted for. The findings of this research also suggest that those item types which assess test takers' listening comprehension at both language and co-text levels are more effective and informative. Finally, the varying degrees of item type difficulty could be better utilised in constructing assessment tools for different diagnostic or teaching purposes. Although this study rejects the hypothesis that item difficulty level is associated with item response format (independent or integrative), it does link the difficulty level of listening items to the level of listening skills being assessed. Therefore, for language learners with higher listening abilities, this study recommends that teachers design more tasks to practice other higher level cognitive strategies, such as reformulating and reproducing audio inputs. The authors want to conclude this article by stating one of the limitations this study has, that is, the

findings rely on testing data only. The arguments could be strengthened with qualitative investigations to support or contradict some of the conclusions drawn. Nevertheless, this study provides some statistical evidence that could potentially lead to further investigation of the validation of computer based testing systems.

References

Anderson, K., & Lynch, T. (1998). *Listening*. Oxford: Oxford University Press.

Baker, F. (2001). The Basics of Item Response Theory. *ERIC Clearing house on Assessment and Evaluation*, University of Maryland, College Park, MD.

Berman, R., & Cheng, L. (2001). English academic language skills: perceived difficulties by undergraduate and graduate students, and their academic achievement. *Canadian Journal of Applied Linguistics, 40*(1), 25-40.

Berne, J. E. (2008). Listening comprehension strategies: a review of the literature. *Foreign Language Annals, 37*(4). 521-531. doi: 10.1111/j.1944-9720.2004.tb02419.x

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment, *Language Testing, 19*(4), 369-394. doi: 10.1191/0265532202lt236oa

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. London: Person.

Brunfaut, T., & Résvez, A. (2015). The role of task and listener characteristics in second language listening. TESOL Quarterly. 49(1), 141-168

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

Buckingham, L., & Alpaslan, R, S. (2017). Promoting speaking proficiency and willingness to communicate in Turkish young learners of English through asynchronous computer-mediated practice, *System, 65*, 25-37, doi: 10.1016/j.system.2016.12.016

Cai, H. W. (2012). Partial dictation as a measure of EFL listening proficiency: evidence from confirmatory factor analysis, *Language Testing*, 30(2), 177-199. doi: 10.1177/0265532212456833

Carrell, P., Dunkel, P., & Mollaun, P. (2004). The effects of note taking, lecture length, and topic on a computer-based test of EFL listening comprehension. *Applied Language Learning*. 14, 83–105.

Chang, A., & Read, J. (2012). The effects of listening support on the listening performance of EFL learners. *TESOL Quarterly*, *40*(2), 375-397. doi: 10.2307/40264527

Chapelle, C., & Douglas, D. (2006). *Assessing Language through Computer Technology.* Cambridge: Cambridge University Press.

Chapelle, C., & Voss, E. (2016). 20 years of technology and language assessment in Language Learning & Technology. *Language Learning & Technology, 20*(2), 116-128.

Chen, I., Cheng, C., & Yen, J. (2012). Effects of presentation mode on mobile language learning: a performance efficiency perspective. *Australasian Journal of Educational Technology*, 28*(1),* 122-137.

Cheng, H. F. (2004). A Comparison of multiple-choice and open ended formats for the assessment of listening proficiency in English. *Foreign Language Annals, 37*(4), 544–555. doi: 10.1111/j.1944-9720.2004.tb02421.x

Cigdem, H., Ozturk, M., & Topcu, A. (2016). Vocational college students' acceptance of web-based summative listening comprehension test in an EFL course, *Computers in Human Behavior, 61*, 522-531. doi: 10.1016/j.chb.2016.03.070

Cubilo, J., & Winke, P. (2013). Redefining the L2 listening construct within an integrated writing task: considering the impacts of visual-cue interpretation and

note-taking. *Language Assessment Quarterly*, 10, 371-397. doi: 10.1080/15434303.2013.824972

Cumming, A., Mulcahy-Ernt, P., & Powers, D. (2005). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL.*TOEFL Monograph Series* 26. Princeton, NJ: Educational Testing Service.

DeLuca, C., Cheng, L, Y., Fox, J., Doe, C., & Li, M. (2013). Putting testing researchers to the test: an explorative study on the TOEFL iBT. *System, 41*(3), 663-676. doi: 10.1016/j.system.2013.07.010

Dixon, E., & Hondo, J. (2014). Re-purposing an OER for the online language course: a case study of Deutsch Interaktiv by the Deutsche Welle, *Computer Assisted Language Learning*, 27*(2)*, 109-121.doi: 10.1080/09588221.2013.818559

Frost, K., Elder, C., & Wigglesworth, G. (2012). Investigating the validity of an integrated listening-speaking task: a discourse-based analysis of test takers' oral performances. *Language Testing, 29*(3), 345-369. doi: 10.1177/0265532211424479

Ghoneam, N. (2015). The effect of subtitling on the enhancement of EFL learners' listening comprehension. *Arab World English Journal,* 6*(4),* 275-290.

Graham, S. (2006). Listening comprehension: the learners' perspective. *System, 34*(2). 165-182. doi: 10.1016/j.system.2005.11.001

Hayati, A., & Mohmedi, F. (2009). The effect of films with and without subtitles on listening comprehension of EFL learners, *British Journal of Educational Technology*, 42*(1),* 181-192. doi: 10.1111/j.1467-8535.2009.01004.x

Jones, L., & Plass, J. (2002). Supporting listening comprehension and vocabulary acquisition in French with multimedia annotations, *Modern Language Journal*, 86*(4),* 546-561. doi: 10.1111/1540-4781.00160

Jung, E, H. (2003). The role of discourse signalling cues in second language listening comprehension. *The Modern Language Journal*, 87(4), 562-577. doi: 10.1111/1540-4781.00208

Kim, J., & Craig, D. (2012). Validation of a videoconference speaking test, *Computer Assisted Language Learning*, 25, 257-275. doi: 10.1080/09588221.2011.649482

Li, C. (2014). An alternative to language learner dependence on L2 caption-reading input for comprehension of sitcoms in a multimedia learning environment, *Journal of Computer Assisted Learning*, 30*(1)*, 17-29. doi: 10.1111/jcal.12019

Lin, J., Lee, Y., Wang, D., & Lin, S. (2016). Reading subtitles and taking Enotes while learning scientific materials in a multimedia environment: cognitive load perspectives on EFL. *Educational Technology & Society*, 19*(4)*. 47-58.

Leveridge, A., & Yang, J. (2013). Testing learner reliance on caption supports in second language listening comprehension multimedia environments, *RECALL*, 25*(2)*, 199-214.doi: 10.1017/S0958344013000074

Luoma, S. (2004), *Assessing speaking*. Cambridge: Cambridge University Press

Nix, J, M. (2016). Measuring latent listening strategies: development and validation of the EFL listening strategy inventory, *System, 57*, 79-97. doi: 10.1016/j.system.2016.02.001

Papageorgiou, S., Stevens, R., & Goodwin, S. (2012). The relative difficulty of dialogic and monologic input in a second language listening comprehension test, *Language Assessment Quarterly*, 9(4), 375-397. doi: 10.1080/15434303.2012.721425

Pearson (2012). *The Official Guide to PTE Academic*. Harlow, UK: Pearson Education ESL.

Perez, M., Peters, E., & Desmet, P. (2014). Is less more? Effectiveness and perceived

usefulness of keyword and full captioned video for L2 listening comprehension. *RECALL*, 26*(1)*, 21-43. doi: 10.1017/S0958344013000256

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: source text use as a predictor of score. *Journal of Second Language Writing*, *22*(3), 217-230. doi: 10.1016/j.jslw.2013.02.003

Powers, D. (1985). A survey of academic demands related to listening skills. (*TOEFL Research Report 20*). Princeton, NJ: Educational Testing Service.

Revesz, A., & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. Studies in Second Language Acquisition 35, 31-65

Roman-Odio, C., & Hartlaub, B. (2003). Classroom assessment of computer-assisted language learning: developing a strategy for college faculty, *Hispania a journal devoted to the teaching of Spanish and Portuguese*, 86*(3)*, 592-607.

Rukthong, A. & Brunfaut, T. (2015, November). *Anybody listening? The role of listening in integrated listening-to-write and listening-to speak tasks*. Paper presented at LTF 2015, Oxford, UK.

Stickler, U., & Shi, L. (2016). TELL us about CALL: an introduction to the virtual special issue (VSI) on the development of technology enhanced and computer assisted language learning published in the System Journal, *System*, *56*, 119-126. doi: 10.1016/j.system.2015.12.004

Taguchi, N., Gomez-Laich, M., & Arrufat-Marques, M. (2016). Comprehension of indirect meaning in Spanish as a foreign language. *Foreign Language Annals*, 49*(4)*, 677-698. doi: 10.1111/flan.12230

Vandergrift, L. (2007). Recent development in second and foreign language listening comprehension research. *Language Teaching*, 40, 191-210. doi:

10.1017/S0261444807004338

Wagner, E. (2007). Are they watching? Test-taker viewing behaviour during an L2 video listening test. *Language Learning & Technology*, *11*(1), 67-86.

Wagner, E. (2008). Video listening tests: what are they measuring? *Language Assessment Quarterly*, *5*(3), 218-243. doi: 10.1080/15434300802213015

Wagner, E. (2013). Assessing listening. In A. Kunan (Ed.), *Companion to language assessment* (pp. 47-63). Oxford, UK: Wiley-Blackwell.

Wagner, E. (2013). An investigation of how the channel of input and access to test questions affect L2 listening test performance, *Language Assessment Quarterly*, *10*(2), 178-195. doi: 10.1080/15434303.2013.769552

Wang, H., Choi, I., Schmidgall, J., & Bachman, L. (2012). Review of Pearson test of English Academic: building an assessment use argument. *Language Testing, 29* (4), 603-619. doi: 10.1177/0265532212448619

Weir, C. J. (2005), *Language testing and validation: an evidence-based approach*, New York: Palgave Macmillan.

Yang, H., & Plakans, L. (2012). Second language writers' strategy use and performance on an integrated reading-listening-writing task. *TESOL Quarterly*, *46* (1), 80-103. doi: 10.1002/tesq.6

Yu, G. (2013). From integrative to integrated language assessment: are we there yet. *Language Assessment Quarterly*, 10, 110-114. doi: 10.1080/15434303.2013.766744

Zhu, X., Li, X., Yu, G., Cheong, C., & Liao. (2016). Exploring the relationships between independent listening and listening-reading-writing tasks in Chinese language testing: toward a better understand of the construct underlying integrated writing tasks. *Language Assessment Quarterly, 13*(3), 167-185. doi:

10.1080/15434303.2016.1210609