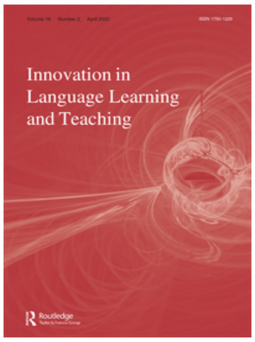

Innovation in Language Learning and Teaching

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rill20

Trialing alternative listening assessment tasks: interactions between text authenticity, item focus and item presentation condition

Stefan O'Grady

INNOVATION IN LANGUAGE LEARNING AND TEACHING
https://doi.org/10.1080/17501229.2022.2109643

Routledge
Taylor & Francis Group

Trialing alternative listening assessment tasks: interactions between text authenticity, item focus and item presentation condition

Stefan O'Grady

International Education Institute, The University of St Andrews, St Andrews, UK

ABSTRACT

Purpose: The current study applies an innovative approach to the assessment of second language listening comprehension skills. This is an important focus in need of innovation because scores generated through language assessment tasks should reflect variation in the target skill and the literature broadly suggests that conventional methods of assessing listening may fall short of achieving this principle (Field [2019]. *Rethinking the Second Language Listening Test from Theory to Practice*. Equinox).

Design: The study investigated interactions between different methods of presenting listening comprehension questions, the focus of the comprehension questions and the relative authenticity of the sound file in an English-medium university entrance listening test. In a balanced design, 61 participants completed a listening test featuring both scripted and unscripted sound files by answering explicit and implicit information comprehension questions under five counterbalanced question preview and presentation conditions. Test scores were analysed using ANOVA and examined for interactions.

Findings: The results revealed interesting relationships between text authenticity and item focus whereby item responses were most frequently correct overall on the explicit items on scripted tasks. However, the reverse was observed on the implicit items, which were more frequently correct on the unscripted tasks.

Value: The research findings have important implications for construct definition and highlight possible directions for the development of pedagogical listening tasks and assessments for English-medium academic study.

ARTICLE HISTORY
Received 15 July 2021
Accepted 31 July 2022

KEYWORDS
Proficiency testing; Language testing; Listening assessments; English for Academic Purposes; Listening proficiency

Introduction

This study investigates listening assessment task characteristics to inform the development of an English-medium university entrance test in Turkey. English-medium instruction (EMI) is widespread in Turkish higher education and a prerequisite for admissions acceptance to EMI courses of study is for prospective students to demonstrate sufficient levels of English ability by meeting the minimum score requirements on an internally produced language test. An important stage of assessment development is to establish the extent to which the test method efficiently measures the intended underlying test construct (O'Sullivan 2016; O'Sullivan and Weir 2011; Weir 2005). Investigating this requires test developers to trial various test methods to identify sources of construct irrelevant

CONTACT Stefan O'Grady ✉ so59@st-andrews.ac.uk International Education Institute, Kinessburn, Kennedy Gardens, St Andrews KY16 9DJ

variance and construct underrepresentation and to resolve issues relating to applicability and practicality (Field 2019; Taylor and Geranpayeh 2011). The literature broadly indicates that important decisions in listening test development are (a) the focus of comprehension questions (b) the availability of question preview (c) the decision to present comprehension questions in writing or as part of the sound file and (d) whether to use scripted listening material (Chang and Read 2013; Field 2013; Green 2017; O'Grady 2021; Yanagawa and Green 2008; Wagner, Liao, and Wagner 2020). Research into the interaction between these test methods in listening assessments is sparse and findings lack consensus. As a result, the empirical backing test developers crucially require for important decisions involving these aspects of listening task design is not currently available and this effectively jeopardizes claims about the validity of decisions based on the results of their tests (Knoch and Chapelle 2018). To address this limitation and to identify potential directions for the development of university entrance listening tests, the current study examines interactions between these test methods. The research was conducted in a Turkish EMI educational context; however, the findings are likely to be relevant to other standardized examinations that assess listening comprehension.

Literature review

The following section reviews the literature relating to four principal features of listening test design. The first is the focus of test items and the importance of assessing test taker understanding of both explicit information and implicit features of the source text. The subsequent discussion concentrates on variation in the presentation and preview of item contents and the hypothesized effect on the listening processes test takers engage. Finally, the section reviews the concept of authenticity in listening assessment and the use of scripted and unscripted texts.

Item focus

The university entrance test needs to generate reliable information about large numbers of test takers' listening proficiency within a limited time frame to identify suitably able candidates to sit later, more challenging stages of the test involving rater-mediated assessment of integrated skills. For this reason, the listening section uses multiple-choice questions (MCQs) that can be scored by machine. MCQs may be characterized as targeting comprehension of (a) specific information that is explicitly stated in the text (e.g. locations, times and dates), or (b) information that is implicit (e.g. main ideas, speaker attitude and purposes) but also clearly retrievable from the text (Becker 2016; Buck 2001; Kang et al. 2019; Koyama, Sun, and Ockey 2016). This is a useful distinction for English language proficiency tests. Owing to limitations in language knowledge, the information that listeners at lower levels of second language ability extract from the contents of a spoken message without having to disproportionately rely on paralinguistic cues may often largely be restricted to short and familiar phrases (Field 2013; Shohamy and Inbar 1991; Wang and Treffers-Daller 2017). During the early stages of second language acquisition, efforts to match connected speech (often bearing little resemblance to citation forms taught in language classrooms) to forms stored in the emerging lexicon consumes substantial amounts of limited attentional resources (Field 2019; Levelt 1989). As listeners gain experience, lexical search becomes increasingly proceduralized and attentional resources become available for parsing of longer stretches of speech and the deeper, meaning-based processing that facilitates comprehension of implicit information (Aryadoust, Goh, and Kim 2012). This involves interpretation of pragmatic cues in the discourse such as politeness, formality and register and draws upon knowledge of the ways in which forms of address vary according to the speaker's perceived level of intimacy with their interlocutor or audience, power differentials in the social situation, and illocutionary intent (Bardovi-Harlig 2013; Becker 2016; Chikalanga 1992; Rost 2011).

Findings in the research literature indicate that comprehension of implicit information is indicative of developed second language proficiency. Taguchi (2008) found large, statistically significant correlations between scores on questions requiring comprehension of implicit information in a listening test and scores on the TOEFL. Becker (2016) investigated different levels of comprehension in a listening test of English for Academic Purposes by comparing responses to questions assessing comprehension of explicitly stated information, and questions requiring comprehension of implicit information provided by four groups of test takers with different levels of language proficiency. The results showed that the implicit information comprehension questions discriminated more successfully between the four groups than those requiring comprehension of explicit information.

Item presentation

Listening tests commonly feature a period of time in which test takers are able to preview written comprehension questions before hearing the sound files. This permits test takers to engage metacognitive strategies such as goal setting, provides a purpose for listening, and replicates the target context of undergraduate study where students listen for a reason and may be more attentive to information in the speech that they find relevant (Buck 2001; Wagner 2013; Yanagawa and Green 2008). However, preview of response options in MCQs may have an additional effect of encouraging construct irrelevant, ‘test-wise strategies’ involving lexical matching from the options to the spoken text (Field 2019, 74). This poses a problem for assessment developers because scores on such tasks may represent basic recognition skills rather than genuine listening comprehension.

Empirical accounts of the impact of preview on listening test scores are conflicting. In an early paper that investigated issues surrounding MCQ preview, Sherman (1997) found that higher scores were associated with question preview in her study. Li et al. (2017) investigated three multiple-choice preview conditions involving preview of stem and options, stem only and options only in a second language English listening test in a secondary school in Taiwan. The researchers reported no statistically significant difference between the three preview conditions. Similarly, in a test of listening featuring both audio-visual and visual only texts, Wagner (2013) found no difference between scores on MCQs under preview and non-preview conditions. Koyama, Sun, and Ockey (2016) compared scores under no preview, question stem only, and stem and option preview conditions on a video comprehension test and found that higher test scores were associated with the two preview conditions. The researchers’ analysis of the test questions indicated that the higher scores may have been attributable to lexical matching strategies facilitated by frequent mentions of keywords and phrases in the sound file. Yanagawa and Green (2008) reported higher scores under preview of stem and options, and stem only conditions in relation to a preview of options only condition. The researchers conclude that whereas the availability of stem preview provided test takers with a contextualizing focus, preview of options only may have encouraged test takers to engage lexical matching strategies that were detrimental in cases where lexical overlap occurred between the sound file and the contents of a distracter rather than the key. Overall, the research findings indicate that previewing question stems may support textual comprehension by guiding and focusing attention whereas (potentially detrimental) lexical matching strategies may be associated with option preview.

In a study involving test takers’ immediate retrospection on their listening performance, Yi’an (1998) described the focusing effect of question preview but also raised the concern that listening tests featuring a great deal of reading may to some extent confound listening ability with reading ability. To respond to MCQs in a listening test, test takers read question stems and options and listen for answers in a way that the textual processing involved in the test simultaneously requires both reading and listening. There are two clear effects on the test taker of this method. Firstly, the test takers’ scores are inevitably affected in some way by their reading ability with the consequence that test scores reflect more than the listening construct. Secondly, concurrently dividing attention between reading and listening in the way that MCQs encourage for sustained periods is uncommon in most academic settings (an obvious exception is lectures involving text that is projected on to a screen) and may lead to a form of cognitive overload whereby the test takers cannot focus on the sound file (Field 2019; Green 2014). At the very least, under these constraints, the test is unlikely to elicit the best possible performance from the test takers (Swain 1985).

Chang and Read (2013, 575) argued that MCQ tasks would constitute a 'pure' form of listening assessment (i.e. results would not be contaminated by reading proficiency) if test takers processed comprehension questions aurally. The researchers conducted a study in which sound recordings were made of test comprehension questions and were embedded into the relevant points of the sound file. One group of test takers completed the listening test by answering multiple-choice questions in the conventional, written mode and a second group in an experimental, spoken mode. The results showed no significant difference between the groups' scores. However, when interactions were investigated between test scores and participants' scores on a general English proficiency exam, the analysis indicated that lower proficiency participants recorded higher scores in the written mode, with a very large effect size. Chang and Read concluded that lower level listeners relied on the support provided in the written mode to complete the test. This conclusion was supported by Yeom (2016) who reported a statistically significant interaction between overall language proficiency and processing mode whereby lower proficiency participants' scores were higher in the written mode. Kim (2015) found higher scores when questions and options were presented in the written mode but also found statistically significant correlations between working memory capacity (measured with a conceptual span task) and scores on the spoken questions, suggesting that the ability to complete the task with spoken questions may to some extent be attributable to variation in test takers' ability to store information in working memory.

O'Grady (2021) investigated the interaction between item focus and item presentation in a listening assessment featuring MCQs that assessed comprehension of explicitly stated information and implicit information. The items were presented in one of four conditions: full written preview, written preview of item stems only, full aural preview, and no preview. The results of a many facet Rasch analysis of item responses showed that test takers recorded the highest scores after written preview of stems and aural presentation of options. Furthermore, there was evidence to suggest that responses to the implicit information questions were more frequently correct than explicit information questions when preview was available for item stems but not options. O'Grady (2021) suggested that this result was due the focusing effect of stem preview and the prevention of lexical matching strategies that would have proved unhelpful in items targeting implicit information.

Scripted and unscripted listening assessments

The extent to which listening texts may be regarded as representative of genuine spoken interaction is a key aspect of listening assessment validity (Field 2019). While the limitations of using scripted material to assess listening comprehension have been widely discussed, scripted texts remain prevalent in modern language assessment and pedagogical material (Wagner 2016, 2018). The use of scripted material is criticized for many reasons. Primarily, it is difficult to replicate the features of connected speech when following a script and test takers are assessed on their ability to follow speech that is unnatural and unrepresentative (Rossi and Brunfaut 2021; Wagner and Ockey 2018). The use of scripted material may also lead to negative washback whereby language learners fail to recognize linguistic forms in authentic connected speech because they have been encouraged to focus on citation forms in the language classroom (Wagner and Toth 2014).

Despite these limitations, research investigating the impact of using unscripted and authenticated texts on test results consistently indicates that test takers record higher scores on scripted tests (Read 2002; Wagner and Toth 2014; Wagner 2018). Most recently, Wagner, Liao, and Wagner (2020) compared test scores on a scripted and an authenticated version of the same test. The authenticated version was prepared from the script but involved 'speaking at a normal rate, enunciating normally, and including phonological features including connected speech, hesitation phenomena, oral turn openers, backchannels' (2020, 11). Results showed that test scores were lower on the authenticated version of the test. The researchers explain that this result was due to the increased speech rate in the authenticated sound files, which may have led to comprehension breakdown. However, the results are based on a test level analysis and did not examine the interactions between the sound file and the item types. This is an important focus because the researchers explain that the test featured explicit and implicit information item questions. Variation in characteristics of the sound file such as increased speech rate is likely to have an important impact on the kind of information test takers extract from a sound file.

The extent to which scripted material may interact with item characteristics (the focus on implicit and explicit information, preview availability and aural or written presentation) is presently unclear. For example, items focusing on explicit information presented with full written preview may be answered using simple lexical matching strategies that are facilitated by the pace and rhythm of the scripted text. Further, withdrawing the focussing effect of item preview may hinder test takers' ability to identify pertinent information in a stream of connected speech. At present, research is yet to explore these possible interactions, and this represents an important gap in the literature.

Summary

To summarize this section, the research literature indicates that written question and option preview may enable test takers to achieve higher test scores than they would when preview is not available. However, by using written MCQs, test scores represent reading ability in addition to listening ability. The use of spoken multiple-choice questions removes the potential for listening test scores to reflect reading ability and may result in a pure listening assessment. Question preview is likely to interact with presentation format because written presentation may encourage lexical matching strategies whereas spoken presentation depends upon working memory capacity and test takers may find it difficult to store comprehension questions in working memory and listen to sound files simultaneously. For this reason, questions must be short to avoid overloading test taker working memory. Regarding the focus of the comprehension questions, questions targeting comprehension of implicit information require higher order processing and may be more challenging than questions targeting comprehension of explicit information. Responses to such questions may benefit from question preview as test takers are able to focus their attention specifically on words in the input text that overlap with or resemble the answer options. In contrast, responses to implicit information comprehension questions, often involving some form of summary are unlikely to be affected because there is very little potential for lexical overlap. Finally, the potential for interaction between item characteristics and text authenticity (defined as whether the sound file is scripted or unscripted) is underexplored in the literature and it is presently unclear whether variation in presentation facilitates or hinders test takers' ability to comprehend implicit and explicit aspects of genuine or scripted speech. This is an important research focus as language testers must strive for authenticity with their assessment tasks but avoid creating measurement instruments that are unfavorable to test takers (Weir 2005).

Research questions

The review of the literature indicates that variation in text authenticity, preview and presentation of comprehension questions may play a critical role in determining test results. To date, the relationship between test scores, text authenticity, question preview and presentation, and item focus has been relatively unexplored. To address the issues identified in the literature review, the current study investigates the following overarching research question and sub-questions.

(1) Do scores on items targeting comprehension of explicit and implicit information generated under different multiple-choice stem and option preview conditions in a test of second language listening vary according to whether the sound file is scripted or unscripted?
(2) Do explicit and implicit item scores vary according to whether the sound file is scripted?

(3) Do scores generated under different multiple-choice stem and option preview conditions vary according to whether the sound file is scripted?

On the basis of the literature review, it was hypothesized that scores on explicit information questions would be impacted more clearly than implicit questions by variation in preview conditions and sound file authenticity. This interpretation was predicated on the assumption of increases in lexical matching facilitated by the opportunity to preview test items (Field 2019), and decreases in the features of connected speech in scripted sound files increasing the correspondence between written forms in the test booklet and spoken forms in the sound file (Wagner, Liao, and Wagner 2020).

Method

Participants

Participants were 61 students studying in the English preparatory program of a Turkish university. The participants were aged between 18 and 21 years old and all had Turkish as their first language. In this context, English language learning begins during the first year of primary school and continues throughout secondary education. However, a high level of proficiency in the language is difficult to attain in the school environment and many prospective students must attend English language preparatory courses before beginning EMI study at the undergraduate level (O'Dwyer and Atlı 2018; British Council 2015). Exposure to spoken English is restricted to the educational environment and any language participants encounter through popular culture and over the internet. At the time of the study, participants were studying in English language classes at the B1 + level on the Common European Framework of Reference (CEFR, Council of Europe 2001). Participants' listening skills were assumed to be relatively parallel because placement into the English language classes is based on the students receiving similar scores on each section of the university entrance exam, which features tests of listening, reading, speaking, writing and integrated skills and has been benchmarked to the CEFR (Kantarcioglu et al. 2010). All participants were recruited to take part in the study on a voluntary basis.

Research instrument and procedure

The listening test

The listening test consisted of five monologues and five dialogues. The texts were situated in the academic domain and featured interactions in campus situations, seminar discussions and academic presentations. Each sound file lasted between one and two minutes in duration. The sound files were recorded by a group of English teachers working in the institution who comprised a variety of nationalities including American, British, Irish, Canadian and Turkish. For the test to be representative of different forms of global English common in the target language situation it was important for the sound files to feature a range of accents (Harding 2012). Fifteen multiple-choice questions targeting explicit information and fifteen questions targeting implicit information were produced by the researcher. Following Becker (2016) and Koyama, Sun, and Ockey (2016), questions required test takers to comprehend explicitly stated specific details and identify suitable paraphrases (e.g. 1a) or to identify the speakers' attitudes and purposes, recognize main ideas and draw conclusions based on information that was implicit in the text (e.g. 1b).

1a Which language did the speaker study in high school?

A. Spanish
B. Chinese
C. Japanese

1b What can we understand about Joe and Steve?

A. They design software.

B. They recently argued.
C. They live together.

To avoid overloading test takers' working memory when answering spoken comprehension questions, the multiple-choice questions contained three options (Haladyna, Rodriguez, and Stevens 2019), and each consisted of a maximum of three words. A period of ten seconds of silence was inserted between each sound file and a period of five seconds was inserted between each spoken question. A speaker that was unfamiliar with the sound files recorded the comprehension questions. Upon construction of the test, two teachers in the institution took the test and submitted their answers. Agreement with the proposed answer key was 100%. Having completed the test, the teachers were further asked to classify the 30 questions as requiring comprehension of explicit or implicit information. Agreement with the intended classification was 100%. Results from a pilot study conducted with 30 participants from the same institution that were studying in classes targeting the same proficiency level indicated that the test met acceptable levels of reliability using a criterion of Cronbach's (α) ≥ .85.

The test conditions

Five experimental presentation conditions were developed for a within-participants design. This necessitated the preparation of different test booklets (see Table 1). Condition one is the current format used in the university entrance test. Condition two provides test takers with a purpose for listening and contextualizes the text without encouraging the lexical matching associated with option preview. In Condition three, processing of the text and questions involved listening only. The spoken preview contextualizes the text and provides a purpose to listen. Condition three was designed to compare the effects of option preview between the written and spoken presentation modes. Condition four was developed to compare the effects of stem preview between the spoken and written presentation modes. Condition five did not involve any form of reading or preview.

The test procedure

The listening test was administered in classrooms with the researcher acting as the invigilator. The sound file was delivered over a speaker system in each classroom and test takers responded to the

Table 1. Overview of conditions and booklet samples.

Condition	Description	Booklet sample
1	Full written preview of question stems and options	*1. Which language did the speaker study in high school?* *A. Spanish* *B. Chinese* *C. Japanese*
2	Written preview of question stems Options are presented in speech as part of the sound file after the input text	*1. Which language did the speaker study in high school?* *A.* *B.* *C.*
3	Full spoken preview of question stems and options	*1.* *A.* *B.* *C.*
4	Spoken preview of stems only. Options are presented after the sound file.	*1.* *A.* *B.* *C.*
5	No preview of question stems or options. Questions are spoken as part of the sound file after the input text	*1.* *A.* *B.* *C*

comprehension questions by circling their answer directly onto the question sheet. The tests were completed within 25 min. In a balanced design, each participant experienced each condition in the listening test in a different order (see Table 2: c = condition), however the order and structure of the items was the same between the participants.

Statistical analysis

To establish the psychometric properties of the test, item responses were analysed using Winsteps, software for conducting Rasch measurement (Linacre 2021a). To answer the research questions, the test results were analysed using a three-way ANOVA, with item responses as the dependent variable and item presentation, item focus and text authenticity as the independent variables. Coding of the dependent and independent variables was as follows: item responses were coded as correct = 1 and incorrect = 0, item presentation involved five levels coded as conditions 1-5, item focus involved two levels, coded as implicit and explicit, and text authenticity involved two levels, coded as scripted and unscripted. A critical alpha value of .05 was set for all analyses and effect sizes were calculated using eta squared (η^2).

Results

The Winsteps analysis showed that the mean test score was 16.5 and scores ranged from 5 to 28. The test reliability statistic, corresponding to Cronbach's alpha (Linacre 2021b), was .78 and the standard error of measurement was 2.48. These values indicate that overall, the test takers found the test difficult but that the test scores were reliable. Item facility values, point biserial correlations, and fit statistics are reported in Table 3. Item facility values ranged from .33 to .84, indicating that the test featured a range of difficult items and simple items. The point biserial correlations demonstrate that the item responses were generally consistent with the exception of item 9, which recorded a negative value. This item was associated with a particularly high outfit mean square statistic indicating that responses to this item were erratic and did not conform with responses to the other 29 items. However, outfit is particularly sensitive to outliers and the infit mean square statistic, sensitive to patterns in the data rather than single item responses, was acceptable (Linacre 2021b). The range of infit mean square values was from .85 to 1.27 indicating that item responses were generally consistent. The mean item facility value for the items on the unscripted tasks was .52 (SD = .14), whereas the corresponding value for the unscripted tasks was .57 (SD = .13), indicating that test takers scored higher when responding to items on the scripted tasks.

To answer the first research question, descriptive statistics for all presentation conditions by item focus and text authenticity are presented in Table 4. The table presents the mean values associated with each interaction. The highest score was recorded under Condition 1 on the items assessing explicit information in the scripted texts. In contrast, the lowest score was recorded under Condition 5 on the items assessing explicit information in the unscripted texts.

A three-way ANOVA was completed to determine the statistical significance and effect size of these interactions. The results are presented in Table 5 and demonstrate that the effect of the three-way interaction between item presentation, item focus and text authenticity on item

Table 2. Order of conditions.

	Questions									
Test takers	1–3	4–6	7–9	10–12	13–15	16–18	19–21	22–24	25–27	28–30
19	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5
11	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1
11	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2
9	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3
11	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4

Table 3. Item analysis statistics.

	Item	Infit mean square	Outfit mean square	Point-biserial correlation	Item facility
Unscripted	1	1.12	1.31	.12	.33
	2	1.22	1.23	.06	.59
	3	1.20	1.30	.02	.41
	4	1.00	1.01	.29	.62
	5	1.00	.93	.32	.84
	6	1.06	1.01	.21	.44
	7	1.02	1.07	.27	.57
	8	1.12	1.16	.13	.34
	9	1.27	1.67	−.11	.38
	10	.87	.83	.43	.41
	11	.92	.87	.41	.69
	12	.96	.93	.36	.64
	13	.97	.92	.30	.39
	14	.99	.94	.31	.54
	15	1.09	1.10	.19	.56
Scripted	16	1.03	1.07	.23	.43
	17	.97	.98	.36	.64
	18	.77	.71	.56	.46
	19	1.02	.97	.26	.48
	20	1.07	1.04	.22	.46
	21	.91	.87	.41	.56
	22	.81	.63	.55	.80
	23	.86	.75	.49	.72
	24	.93	.94	.38	.61
	25	1.11	1.10	.19	.77
	26	.85	.79	.47	.72
	27	.94	.88	.35	.43
	28	.99	.98	.27	.41
	29	.98	.94	.31	.52
	30	.87	.82	.44	.49

responses was not statistically significant $F(4,1830) = 2{,}016$, $p = .09$. However, there was a statistically significant two-way interaction between item focus and text authenticity with a small effect size $F(1,1830) = 3{,}042$, $p = .00$, $\eta^2 = .007$. Moreover, the main effects of presentation condition $F(4,1830) = 9{,}193$, $p = .00$, $\eta^2 = .02$, and of authenticity $F(1,1830) = 5{,}059$, $p = .03$, $\eta^2 = .003$ were statistically

Table 4. Descriptive statistics by text authenticity, item presentation, and item focus.

	Condition	Item focus	Mean	SD
Unscripted texts	1	Implicit	.64	.48
		Explicit	.69	.47
	2	Implicit	.52	.50
		Explicit	.44	.50
	3	Implicit	.61	.49
		Explicit	.53	.50
	4	Implicit	.55	.50
		Explicit	.41	.49
	5	Implicit	.42	.50
		Explicit	.38	.49
Scripted texts	1	Implicit	.55	.50
		Explicit	.73	.44
	2	Implicit	.53	.50
		Explicit	.52	.50
	3	Implicit	.52	.50
		Explicit	.69	.47
	4	Implicit	.48	.50
		Explicit	.71	.46
	5	Implicit	.51	.50
		Explicit	.47	.50

Table 5. Three-way ANOVA results.

Source	Type III Sum of Squares	df	Mean Square	*F*	*p*	η^2
Corrected model	17.758	19	.935	3.875	.000	.039
Intercept	539.627	1	539.627	2237.188	.000	.553
Authenticity	1.220	1	1.220	5.059	.025	.003
Presentation	8.870	4	2.217	9.193	.000	.020
Focus	.241	1	.241	.998	.318	.001
Authenticity*presentation	1.020	4	.255	1.057	.376	.002
Authenticity*focus	3.042	1	3.042	12.613	.000	.007
Presentation*focus	1.672	4	.418	1.733	.140	.004
Authenticity*presentation*focus	1.945	4	.486	2.016	.090	.004
Error	436.586	1810	.241			
Total	991.000	1830				
Corrected Total	454.344	1829				

Note: *R* squared =.039 (Adjusted *R* Squared = .029).

Table 6. Text authenticity and item focus interactions.

		Mean	SD
Unscripted	Implicit	.55	.50
	Explicit	.49	.50
Scripted	Implicit	.51	.50
	Explicit	.62	.49

significant. The interaction effect between presentation condition and text authenticity was not statistically significant *F*(4,1830) = 1,057, *p* = .38.

Descriptive statistics describing the effect of the interaction between item focus and text authenticity on item responses are presented in Table 6. The figures demonstrate that scores on the implicit information items were higher when the text was not scripted. The difference in mean scores between the scripted and unscripted tasks was .04. In contrast, scores on the explicit information items were higher when the text was scripted and the difference in mean scores was .13.

To summarize the results, the effect of the interaction between item focus, presentation and text authenticity was not statistically significant. However, a statistically significant interaction between item focus and text authenticity was observed and this is discussed in the following section.

Discussion

The study examined interactions between item presentation, item focus and text authenticity in a test of second language listening. In language testing, it is important to establish that assessment tasks are free from construct irrelevant variance; 'variables not directly related to the listening construct must be removed' (Aryadoust 2012, 41). To this end, the study presented comprehension questions as part of a sound file and as printed text in a question booklet and compared the results. This method represents an attempt to investigate a 'pure' test of listening ability that is not contaminated by variation in reading ability (Chang and Read 2013, 575). It is also important to establish that the construct is not underrepresented in the assessment instrument and the use of unscripted texts containing features of connected speech is a central aspect of this requirement in listening assessments (Wagner, Liao, and Wagner 2020)

Research has shown that test takers score consistently higher when comprehension questions are presented in writing rather than in speech (Chang and Read 2013; Kim 2015; Yeom 2016). This is despite accounts of the adverse effects of dividing attention between the visual and aural modes in listening assessments (Field 2019). In addition, test takers conventionally record higher scores on scripted sound files (Read 2002; Wagner 2018; Wagner, Liao, and Wagner 2020). However, an overall interaction between item presentation, item focus and text authenticity was not observed, indicating that test takers reacted in the same way to variation in item presentation when completing the explicit and implicit comprehension items regardless of whether the assessment text was scripted. This was an unexpected finding as it was anticipated that test takers would require the extra support of preview to comprehend the connected speech. It may be the case that test takers did not find the distinction between unscripted and scripted texts as marked as expected. Alternatively, test takers may have engaged similar test taking strategies to respond to the multiple-choice questions irrespective of text authenticity. This suggests that the common limitations attributed to the multiple-choice format such as lexical matching, successfully eliminating options using skills not specifically related to the target construct, and guessing are just as likely in scripted and unscripted texts (Haladyna, Rodriguez, and Stevens 2019; Holzknecht et al. 2021; Rukthong 2021). However, it should be noted that the absence of a clear difference in overall scores between the scripted and unscripted texts is also strong justification to include authentic sound files in listening assessments: test representativeness increases, and test takers do not seem to be adversely affected by this modification to the test.

A significant interaction between item focus and text authenticity was observed. Comprehension of explicit information involving dates, names and explicitly stated details was markedly higher on the scripted tasks. This suggests that the spoken characteristics in the scripted texts were more conducive to matching information between the sound file and the item contents. As the literature review makes clear, processing of small units of information requires limited levels of listening ability (Field 2013; Shohamy and Inbar 1991; Wang and Treffers-Daller 2017). In contrast, the implicit information items were more frequently correct on the unscripted texts. This indicates that test takers were better able to answer questions requiring the identification of speaker purpose and attitude, main ideas and inferencing when the text featured connected speech. In short, test takers generate a better overall sense of the communicative event when the sound file is authentic.

The results of this study indicate that the impact of using unscripted speech may be greater on certain item types (Read 2002; Wagner 2018; Wagner, Liao, and Wagner 2020). Whereas scripted speech is conducive to answering information about specific information, unscripted speech may promote higher order processing of the sound file (Field 2019). Evidence of higher order processing is reflected in the results of implicit information items, which were designed to assess various aspects of the test takers' pragmatic competence (Ishihara and Cohen 2022). To summarize, scripted speech may lead to higher scores in listening tests but this effect is likely to be dependent on the number of explicit information questions the test features. This finding represents an important contribution to the listening assessment literature and may be regarded as further evidence for the necessity of using unscripted speech in listening assessments, particularly when the test-taking population includes advanced level learners.

The contributions of these findings relate most clearly to language test development. For example, in an analysis of IELTS listening test content, Aryadoust (2012) found that test items primarily focused on comprehension of explicit information and suggested that the test had represented the listening construct too narrowly and required more questions targeting comprehension of implicit information, compelling test takers to consider the function and illocutionary intent of the input texts. The results of this study provide further insight in this discussion and indicate that listening tests featuring both explicit and implicit information comprehension questions are likely to be affected by the decision to use scripted material. Test developers need to be made aware of this potential test method effect.

It is important to acknowledge a number of limitations of the research. Firstly, by using spoken comprehension questions, construct irrelevant variance associated with reading ability is removed. However, the method introduces an additional confounding element relating to spoken question comprehension. For this reason, scores on spoken comprehension questions reflect both text comprehension and question comprehension. Although there was no statistically significant interaction between item presentation, focus and text authenticity, the conflation of text and question comprehension may constitute a source of construct irrelevant variance and may be investigated in future research involving qualitative methods. Examining test taker accounts of test performance may have provided information about the interaction between item preview and presentation, focus and text authenticity with the effect that a more detailed construct definition would be possible for the listening section of the university entrance test (Rukthong 2021). This would be an important focus for future validation research. Finally, the use of audio-visual based texts may have provided a source of valuable contextual support; this is especially the case in the dialogic texts, which were set in a more socially oriented setting where interlocutors often make use of non-verbal information to relay and receive information (Batty 2015, 2018). Future research may examine interactions between audio-visual texts and item characteristics to identify potential avenues for test development.

Conclusion

The current study investigated the interaction between multiple-choice question focus, question stem and option preview and presentation mode on scores in a test of second language listening for university admissions. The primary objective of the research was to identify an appropriate format for the listening section of the university entrance test that would enhance construct coverage (Field 2013). The findings indicated that an appropriate direction for this test is to use unscripted material and focus assessment on the comprehension of implicit information. This format is likely to enhance the representativeness of the test and assess a richer listening construct. Furthermore, it is widely accepted that language assessments should be representative of language use in the target context (Weir 2005) but also bias for the best performance (Swain 1985). To this end, institutions involved in test development must account for likely interactions between preview, processing mode, item focus and text authenticity when designing listening assessment tasks and investigate the effect of this interaction on the processes test takers engage during the test. As language assessments progressively fulfill a role of regulating access to opportunity, such innovation in test design will increasingly become a necessity for language testers.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Stefan O'Grady is an Associate Lecturer in Academic English and TESOL at the University of St Andrews. He has worked in English language teaching and assessment for over fifteen years in a variety of national contexts including Turkey, France, China and Kuwait. He received his PhD in 2018 from CRELLA, at the University of Bedfordshire. His main research interests are in language test development and validation.

ORCID

Stefan O'Grady http://orcid.org/0000-0003-3810-713X

References

Aryadoust, V. 2012. "Differential Item Functioning in While-Listening Performance Tests: The Case of the International English Language Testing System (IELTS) Listening Module." *International Journal of Listening* 26 (1): 40–60. doi:10.1080/10904018.2012.639649.

Aryadoust, V., C. C. M. Goh, and L. O. Kim. 2012. "Developing and Validating an Academic Listening Questionnaire." *Psychological Test and Assessment Modeling* 54 (3): 227–256.

Bardovi-Harlig, K. 2013. "Developing L2 Pragmatics." *Language Learning* 63 (1): 68–86. doi:10.1111/j.1467-9922.2012.00738.x.

Batty, A. O. 2015. “A Comparison of Video- and Audio-Mediated Listening Tests with Many-Facet Rasch Modeling and Differential Distractor Functioning.” *Language Testing* 32 (1): 3–20. doi:10.1177/0265532214531254.

Batty, A. O. 2018. “Investigating the Impact of Nonverbal Communication Cues on Listening Item Types.” In *Assessing L2 Listening Moving Toward Authenticity*, edited by G. Ockey and E. Wagner, 161–179. Amsterdam: John Benjamins.

Becker, A. 2016. “L2 Students’ Performance on Listening Comprehension Items Targeting Local and Global Comprehension.” *Journal of English for Specific Purposes* 24: 1–13. doi:10.1016/j.jeap.2016.07.004.

British Council. 2015. *The State of English in Higher Education in Turkey*. www.britishcouncil.org.tr/sites/default/files/he_baseline_study_book_web_-_son.pdf.

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

Chang, A., and J. Read. 2013. “Investigating the Effects of Multiple-Choice Listening Test Items in the Oral Versus Written Mode on L2 Listeners’ Performance and Perceptions.” *System* 41 (3): 575–586. doi:10.1016/j.system.2013.06.001.

Chikalanga, I. 1992. “A Suggested Taxonomy of Inferences for the Reading Teacher.” *Reading in a Foreign Language* 8: 697–709.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Field, J. 2013. “Cognitive Validity.” In *Examining Listening: Research and Practice in Assessing Second Language Listening Studies in Language Testing 35*, edited by A. Geranpayeh and L. Taylor, 77–151. Cambridge: Cambridge University Press.

Field, J. 2019. *Rethinking the Second Language Listening Test from Theory to Practice*. Sheffield: Equinox.

Green, A. 2014. *Exploring Language Assessment and Testing: Language in Action*. London: Routledge.

Green, R. 2017. *Designing Listening Tests A Practical Approach*. London: Palgrave Macmillan.

Haladyna, T., M. Rodriguez, and C. Stevens. 2019. “Are Multiple-Choice Items Too Fat?” *Applied Measurement in Education* 32 (4): 350–364. doi:10.1080/08957347.2019.1660348.

Harding, L. 2012. “Accent, Listening Assessment and the Shared-L1 Advantage: A DIF Perspective.” *Language Testing* 29 (2): 163–180. doi:10.1177/0265532211421161.

Holzknecht, F., G. McCray, K. Eberharter, B. Kremmel, M. Zehentner, R. Spiby, and J. Dunlea. 2021. “The Effect of Response Order on Candidate Viewing Behaviour and Item Difficulty in a Multiple-Choice Listening Test.” *Language Testing* 38 (1): 41–61. doi:10.1177/0265532220917316.

Ishihara, N., and A. Cohen. 2022. *Teaching and Learning Pragmatics Where Language and Culture Meet*. 2nd ed. London: Routledge.

Kang, T., M. Arvizu, P. Chaipuapae, and R. Lesnov. 2019. “Reviews of Academic English Listening Tests for Non-Native Speakers.” *International Journal of Listening* 33 (1): 1–38. doi:10.1080/10904018.2016.1185210.

Kantarcioglu, E., C. Thomas, J. O’Dwyer, and B. O’Sullivan. 2010. “Benchmarking a High-Stakes Proficiency Exam: The COPE Linking Project.” In *Aligning Tests with the CEFR: Reflections on Using the Council of Europe’s Draft Manual*, edited by W. Martyniuk, 102–118. Cambridge: Cambridge University Press.

Kim, B. 2015. “The Effects of Working Memory Span on Listening Tests Without Preview Questions.” *Language Research* 51 (2): 403–420.

Knoch, U., and C. A. Chapelle. 2018. “Validation of Rating Processes Within an Argument-Based Framework.” *Language Testing* 35 (4): 477–499. doi:10.1177/0265532217710049.

Koyama, D., A. Sun, and G. Ockey. 2016. “The Effects of Item Preview on Video-Based Multiple-Choice Listening Assessments.” *Language Learning and Technology* 20 (1): 148–165.

Levelt, M. 1989. *Speaking: From Intention to Articulation*. Cambridge: MIT Press.

Li, C., C. Chen, M. Wu, Y. Kuo, Y. Tseng, S. Tsai, and H. Shih. 2017. “The Effects of Cultural Familiarity and Question Preview Type on the Listening Comprehension of L2 Learners at the Secondary Level.” *International Journal of Listening* 31 (2): 98–112. doi:10.1080/10904018.2015.1058165.

Linacre, J. M. 2021a. *Winsteps® (Version 5.0.0)* [Computer Software]. Beaverton, OR: Winsteps.com. Accessed January 1, 2021. https://www.winsteps.com/.

Linacre, J. M. 2021b. *Winsteps® Rasch Measurement Computer Program User’s Guide*. Beaverton, OR: Winsteps.com.

O’Dwyer, J., and H. Atlı. 2018. “ESP/EAP in University Programs in a Non-Target Language Community – Issues and Challenges.” In *Key Issues in English for Specific Purposes in Higher Education*, edited by Y. Kırkgöz and K. Dikilitaş, 291–304. Cham: Springer.

O’Grady, Stefan. 2021. “Adapting Multiple-Choice Comprehension Question Formats in a Test of Second Language Listening Comprehension.” *Language Teaching Research*. http://dx.doi.org/10.1177/1362168820985367.

O’Sullivan, B. 2016. “Validity: What Is It and Who Is It for?” In *Epoch Making in English Teaching and Learning: Evolution, Innovation, and Revolution*, edited by Y. Leung, 201–222. Taipei: Crane Publishing.

O’Sullivan, B., and C. Weir. 2011. “Language Testing and Validation.” In *Language Testing Theory and Practice*, edited by B. O’Sullivan, 13–32. London: Palgrave.

Read, J. 2002. “The Use of Interactive Input in EAP Listening Assessment.” *Journal of English for Academic Purposes* 1 (2): 105–119. doi:10.1016/S1475-1585(02)00018-8.

Rossi, O., and T. Brunfaut. 2021. “Text Authenticity in Listening Assessment: Can Item Writers Be Trained to Produce Authentic-Sounding Texts?” *Language Assessment Quarterly* 18 (4): 398–418. doi:10.1080/15434303.2021.1895162.

Rost, M. 2011. *Teaching and Researching Listening*. Harlow: Pearson.

Rukthong, A. 2021. "MC Listening Questions vs. Integrated Listening-to-Summarize Tasks: What Listening Abilities Do They Assess?" *System* 97. doi:10.1016/j.system.2020.102439.

Sherman, J. 1997. "The Effect of Question Preview in Listening Comprehension Tests." *Language Testing* 14 (2): 185–213. doi:10.1177/026553229701400204.

Shohamy, E., and O. Inbar. 1991. "Validation of Listening Comprehension Tests: The Effect of Text and Question Type." *Language Testing* 8 (1): 23–40. doi:10.1177/026553229100800103.

Swain, M. 1985. "Large Scale Communicative Testing: A Case Study." In *New Directions in Language Testing*, edited by Y. Lee, C. Fok, R. Lord, and G. Low, 35–46. Hong Kong: Pergamon Press.

Taguchi, N. 2008. "The Effect of Working Memory, Semantic Access, and Listening Abilities on the Comprehension of Conversational Implicatures in L2 English." *Pragmatics & Cognition* 16 (3): 517–539. doi:10.1075/pc.16.3.05tag.

Taylor, L., and A. Geranpayeh. 2011. "Assessing Listening for Academic Purposes: Defining and Operationalizing the Test Construct." *Journal of English for Specific Purposes* 10 (2): 89–101. doi:10.1016/j.jeap.2011.03.002.

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.

Wagner, E. 2016. "Authentic Texts in the Assessment of L2 Listening Ability." In *Contemporary Second Language Assessment*, edited by J. Banerjee and D. Tsagari, 103–123. London: Bloomsbury Academic.

Wagner, E. 2018. "A Comparison of Listening Performance on Tests with Scripted or Authenticated Spoken Texts." In *Assessing L2 Listening Moving Toward Authenticity*, edited by G. Ockey and E. Wagner, 29–44. Amsterdam: John Benjamins.

Wagner, E., Y. Liao, and S. Wagner. 2020. "Authenticated Spoken Texts for L2 Listening Tests." *Language Assessment Quarterly* 18 (3). doi:10.1080/15434303.2020.1860057.

Wagner, E., and G. Ockey. 2018. "An Overview of the Use of Authentic, Real-World Spoken Texts on L2 Listening Tests." In *Assessing L2 Listening Moving Toward Authenticity*, edited by G. Ockey and E. Wagner, 13–28. Amsterdam: John Benjamins.

Wagner, E., and P. D. Toth. 2014. "Teaching and Testing L2 Spanish Listening Using Scripted vs. Unscripted Texts." *Foreign Language Annals* 47 (3): 404–422. doi:10.1111/flan.12091.

Wang, Y., and J. Treffers-Daller. 2017. "Explaining Listening Comprehension among L2 Learners of English: The Contribution of General Language Proficiency, Vocabulary Knowledge and Metacognitive Awareness." *System* 65: 139–150. doi:10.1016/j.system.2016.12.013.

Weir, C. 2005. *Language Testing and Validation*. London: Palgrave Macmillan.

Yanagawa, K., and A. Green. 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 (1): 107–122. doi:10.1016/j.system.2007.12.003.

Yeom, S. 2016. "The Effects of Presentation Mode and Item Type on L2 Learners' Listening Test Performance and Perception." *English Teaching* 71 (4): 27–54. doi:10.15858/engtea.71.4.201612.27.

Yi'an, W. 1998. "What Do Tests of Listening Comprehension Test? A Retrospection Study of EFL Test-Takers Performing a Multiple-Choice Task." *Language Testing* 15 (1): 21–44. doi:10.1177/026553229801500102.