Check for updates

Article

LANGUAGE TEACHING RESEARCH

Language Teaching Research
1–25
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1362168820985367
journals.sagepub.com/home/ltr
SAGE

Adapting multiple-choice comprehension question formats in a test of second language listening comprehension

Stefan O'Grady
Bilkent University, Turkey

Abstract

The current study explores the impact of varying multiple-choice question preview and presentation formats in a test of second language listening proficiency targeting different levels of text comprehension. In a between-participant design, participants completed a 30-item test of listening comprehension featuring implicit and explicit information comprehension questions under one of four multiple-choice question preview and presentation conditions. Interactions between preview, presentation and comprehension in the participants' test scores were analysed using many facet Rasch analysis. The results suggest that the measurement of participants' listening ability was directly influenced by the presentation of multiple-choice questions. Test scores were highest when participants were able to preview the multiple-choice question stems before the sound file and listened to the options after the text had completed. However, interactions between preview and presentation conditions and comprehension level were only statistically significant in an analysis of the low scoring students' item responses, which were more frequently correct when preview of item stems was available for questions targeting comprehension of implicit information. The research underscores the importance of accounting for test design when making inferences about language learners' listening ability and will be of interest to teachers, practitioners and researchers developing listening assessment tasks.

Keywords

English as a second language, listening comprehension, multiple-choice questions, Rasch measurement

Corresponding author:
Stefan O'Grady, School of English Language, Bilkent University, Ankara 06800, Turkey.
Email: ogradystefan@gmail.com

I Introduction

The measurement of prospective students' language ability is a primary responsibility of pre-sessional university language programs (Dimova et al., 2020). For valid inferences about prospective students' readiness to begin undergraduate study in a second language to be made based on the outcome of measurement, the measurement instrument must appropriately reflect language processing and use in situations beyond the second language classroom (O'Sullivan, 2016; O'Sullivan & Weir, 2011; Weir, 2005). Assessment also has a reciprocal relationship with teaching, whereby the results of assessment inform the focus of instruction and the focus of instruction likewise informs the contents of assessment (Green, 2014). The assessment of listening ability is a necessary but notoriously complicated component of language proficiency measurement that has conventionally drawn upon comprehension question formats such as multiple-choice to determine students' listening competence but is also commonly criticized for assessing cognitive operations that are unlikely to be required in situations beyond pedagogical contexts (Badger & Yan, 2009; Field, 2019; Taylor & Geranpayeh, 2011). However, the potential to assess large groups of test candidates simultaneously means that comprehension question formats remain a common feature of listening tests and attention has recently shifted toward increasing the authenticity of the format (Chang & Read, 2013). The research literature suggests that responses to comprehension questions in listening proficiency assessments may be influenced by a series of interrelated factors; these include the opportunity to preview question contents, whether the questions are presented verbally or in print, and the level of textual comprehension required to answer the questions correctly (Field, 2019). However, studies investigating interactions between these factors are rare and have reached conflicting conclusions. Consequently, there are large gaps in the theoretical approach test developers follow when producing measurement instruments of listening proficiency and the interpretation of listening test scores is necessarily clouded. This study attempts to fill this gap by investigating the interaction between these factors in a test of second language listening proficiency.

II Literature review

1 Test texts

It is widely recognized that listening assessment texts should reflect the language that test takers will encounter in the target context (Green, 2014; Weir, 2005). That is, listening tests should aim for situational authenticity (Badger & Yan, 2009). During academic studies, undergraduate students listen in a range of situations that vary intrinsically in terms of organizational and stylistic conventions (Carter & McCarthy, 1997, 2006). To reflect this range, spoken texts used for purposes of language learning and assessment are often categorized according to a continuum ranging from spontaneous and oral to edited and literate (Wagner & Ockey, 2018). Many tests of academic listening proficiency fail to represent the diversity this continuum describes because input sound files are prepared from texts that were originally produced to be processed visually, such as interviews published on websites, and hence bear little resemblance to authentic speech (Buck, 2001; Field, 2019; Wagner & Ockey, 2018). Despite best efforts to sound realistic

in the recording studio, replicating the features of speech in genuine, communicative interaction is difficult and when attempted, unlikely to appear completely natural (Wagner, 2018). The limitation is less evident in tasks designed to assess comprehension of academic monologues such as lectures, which are common in tests of English for academic purposes (EAP) and represent an extensively prepared variety of speech that tends to resemble written language. In contrast, when dialogue tasks involve scripts, test takers are assessed on their ability to follow interaction that is artificial and test scores do not appropriately reflect candidates' ability to comprehend speech in the target situation. The consequences of not including authentic listening materials in assessments are likely to extend to language classrooms where teachers and students may regard authentic spoken texts as superfluous or even irrelevant because they do not appear in the end-of-course test.

Although concerns related to inauthenticity are well attested in the listening assessment and pedagogical literature, scripted material remains ubiquitous largely because of the possibility to adapt the textual organization of scripts to facilitate the production and even-distribution of comprehension questions (Wagner, 2016, 2018). This ensures that sufficient context and support can be provided to answer questions but also detracts from the representativeness of the speech. Researchers have suggested that sufficient textual organization can be achieved without substantially threatening textual authenticity by drawing upon structured improvisation in the recording studio (Clark, 2014; Field, 2019; Wagner & Ockey, 2018). Structured improvisation refers to an approach to recording in which trained voice actors are provided with a general outline of a particular situation rather than a script and 'follow the basic outline of the text while still composing and uttering (virtually) simultaneously' (Wagner & Ockey, 2018, p. 21). This results in a recording that contains features of speech test takers will commonly encounter in the target situation but are difficult to reproduce in scripted interaction. Such speech may be unfamiliar to listeners that do not have sufficient exposure but its inclusion is necessary for decisions to be made about prospective students' ability to participate in undergraduate study. Furthermore, the addition of authentic speech in assessments is likely to produce positive washback in language classrooms where students will be more motivated to complete listening tasks requiring the comprehension of genuine spoken interaction (Buck, 2001). In sum, in addition to scripted monologues, assessing the ability to comprehend spontaneous interaction means that test scores are more likely to be representative of listening ability in the target context.

2 The targeted comprehension level of multiple-choice questions (MCQs)

The use of multiple-choice questions (MCQs) in listening tests has been argued to lack interactional authenticity (Badger & Yan, 2009). However, the format allows for simultaneous and reliable assessment of large numbers of prospective students and those test candidates that are identified as suitably proficient may then sit institutionally resource-demanding, integrated skills tests requiring trained interlocutors and raters that replicate more thoroughly the language demands of undergraduate study. Measurement items assessing receptive language skills may be categorized as requiring comprehension of information that is made explicit in the text and information that is implicit, yet also

openly recoverable from the context (Becker, 2016; Buck, 2001; Kang et al., 2019; Koyama et al., 2016). Examples of information that is explicit in a text may relate to specific details such as the names of characters or of locations in the input. Implicit information is not directly stated in the text and concerns aspects of the social situation such as rhetorical purpose, speaker attitude, and illocutionary intent.

The distinction between comprehension of explicitly stated and implicit information in listening test development is not a trivial one. To compensate for limited automaticity of processing, inexperienced listeners most typically seek to engage simple lexical matching strategies to complete second language listening assessments (Field, 2019). Consequently, inexperienced listeners may only successfully comprehend short, isolated extracts of the input text (Field, 2013; Shohamy & Inbar, 1991; Wang & Treffers-Daller, 2017). As listeners develop their knowledge of the second language, listening comprehension becomes progressively automatic and attentional resources become available for the parsing and storing of larger units of information in the discourse (Aryadoust et al., 2012). The ability to interpret elusive rhetorical features of the discourse and the subtle implications they have on socially contextualized interaction also becomes more available with increasing automaticity of processing owing to increasingly rapid lexical search and parsing mechanisms (Bardovi-Harlig, 2013; Becker, 2016; Buck; 2001; Chikalanga, 1992; Rost, 2011). Based on this account of the development of listening proficiency, successful responses to questions requiring comprehension of implicit information may be indicative of advanced ability in second language listening.

The research literature supports this interpretation. Becker (2016) compared scores on items requiring comprehension of explicit and implicit information and found greater discrimination between the participants using the implicit information questions. Taguchi (2008) reports large correlations between TOEFL scores and implicit information questions in a listening test. In a test of listening featuring both implicit and explicit information questions, Batty (2018) reports that implicit information questions were completed more successfully when participants were able to view an audio-visual recording of the input text in relation to audio only, whereas scores on the explicit information items were less clearly impacted by the presence of visual information. This finding was accredited to the extra support provided by the presence of visuals cues in the video recording.

The potential to distinguish between failure to comprehend explicitly stated information and implicit information in listening texts is likely to benefit language teachers seeking to target instruction through diagnostic activities. Whereas failure to respond correctly to items requiring comprehension of explicitly stated information may be indicative of a need to introduce the spoken forms of relevant lexis and grammar in connected speech at the pre-listening stage of a pedagogical activity, inaccurate responses to implicit information questions may signal that more time should be devoted to the activation of schemata (Wallace, 2020). Including both explicit and implicit information questions in diagnostic listening activities is pedagogically appropriate.

3 The preview of MCQs

Listening assessment tasks typically involve a brief amount of time before the sound file begins that permits test candidates to preview the contents of the MCQs. This feature of

the test facilitates the setting of goals, provides information about the topic of the input texts, and allows candidates to target attention and listen selectively (Buck, 2001; Wagner, 2013; Yanagawa & Green, 2008). An unintended consequence of this preview, however is that the opportunity to engage lexical matching strategies from the contents of the MCQs to the sound file increases (Field, 2019). Successful lexical matching clouds the inferences that may be made based on test results because scores may be more representative of basic recognition skills than the target construct of academic listening ability.

Despite this, the research findings on the facilitative impact of question preview are inconsistent. Li et al. (2015) report no difference in scores between preview of stem and options, stem only and options only. Wagner (2013) investigated the impact of variation in question preview in listening tests featuring audio-visual sound files and also found no statistically significant difference between the preview conditions. In contrast, in a study exploring preview of MCQs, Sherman (1997) reported higher scores when question preview was available. In a study of video-comprehension, Koyama et al. (2016) investigated the impact of varying the amount of information available during preview by presenting questions with preview of question stem only, full question and option preview, and no preview. The research findings showed that test candidates achieved higher scores when some form of preview was included in the task. Analysis of the question contents and contents of the sound files revealed substantial overlap and repetition of important words and phrases between the questions and the input text and thus test takers were able to engage lexical matching strategies successfully in the preview condition. An important implication of this finding is that preview of MCQ options may prove especially facilitative in the completion of explicit information questions rather than implicit information questions owing to likely overlaps and the presence of close synonyms between the input text and question contents. Yanagawa and Green (2008) took the exploration of preview of question contests further by comparing test taker scores recorded after preview of stem only, and preview of options only and contrasted the results with scores on a full stem and option only preview. The results demonstrated that scores were higher in tasks featuring preview of the question stem and the question stem and options. Scores were lowest in the option-only preview condition. The researchers' interpretation of these findings was that preview of stems serves to contextualize the sound file, whereas option preview may in fact lead to unfavorable lexical matching strategies that may be erroneously followed when the contents of the distracters, rather than the key, overlap with the contents of the sound file.

4 The presentation of MCQs

In a study investigating the construct validity of a listening test, Yi'an (1998) speculates that the use of written MCQs in tests of listening comprehension results in a situation in which score variation may be directly attributable to variation in reading ability. Traditional MCQs require reading comprehension skills and test scores inevitably, to some extent, reflect reading ability. However, a larger, related threat to validity, traditional MCQ formats predispose the test candidate to divide attention concurrently between the sound file and printed text (i.e. the contents of the comprehension questions) in such a way that would be unnatural in most communicative situations. The cognitive

burden associated with this assessment format cannot be said to reflect the kind of processing required at the undergraduate level and thus this test method effect essentially reduces claims about the representativeness and fairness of listening test tasks (Field, 2019; Green, 2014; Swain, 1985).

Chang and Read (2013, p. 575) make the case for what they refer to as a "pure" test of listening. The researchers contend that test scores would be more representative of listening ability if the listening task contained MCQs that were delivered verbally. To test this claim, the researchers developed a test of listening comprehension featuring spoken MCQs and compared the scores on this test with a traditional version of the test featuring conventional MCQs. The research findings demonstrated that test scores did not vary between the two versions of the test. Examining interactions between scores on the tests and the results of general English proficiency test, the researchers found that test takers recording low scores on the general English proficiency test were able to score highly on the listening test when conventional comprehension questions were involved. This indicated that the less able test takers were reliant upon the extra support provided by the written contents of the listening test task and were essentially able to boost their scores under this condition. In later research, Yeom (2016) reported that lower ability listeners were able to increase their scores substantially when comprehension questions were presented in writing rather than verbally. Kim (2015) also reports higher scores using written comprehension questions but further discusses correlations between listening test scores using spoken MCQs and the results of a conceptual span task to measure working memory capacity, indicating that spoken MCQ score variation may also be attributable to variation in working memory capacity. When test scores are used as the basis for pedagogical decisions such as placement to a course of study or for targeting instruction, there can be little doubt about what the score represents. Scores that are attributable to variation in working memory capacity or reading ability rather than listening ability may result in misplacement of test takers into courses of study or nonessential instruction. Efforts should be made to minimize construct irrelevant influences on listening test scores.

5 Summary

To summarize, the literature indicates that comprehension questions that are based on implicit information require higher automaticity of listening processes and may be completed less successfully than explicit information questions by less advanced language learners. Further, although the role of written comprehension questions in invalidly inflating listening scores is widely assumed, the empirical evidence of this effect is inconsistent. Full written preview of question stems and options may permit test candidates to record higher scores than would be possible without preview. The literature also indicates that scores generated through written MCQs may confound listening and reading and this assessment format may induce a level of cognitive burden that is uncommon in the target situation. Presenting MCQs verbally may be regarded as a pure test of listening that is not contaminated by reading proficiency. However, the presentation of MCQs may interact with question preview in a way that determines the extent to which test candidates engage lexical matching strategies to complete the test. Without a permanent record of the MCQs, storing the contents of aurally processed MCQs may place an extra cognitive burden on

test candidates that affects their ability to complete the test. For this reason, the length of MCQs should be limited. Question preview and question comprehension level may interact because preview of options will likely prove amenable to answering questions requiring comprehension of explicitly stated information. Responses to implicit information questions are less likely to be impacted by the preview of options as there is little chance for lexical overlap between the sound file and the MCQ contents to occur.

III Research questions

Variation in MCQ presentation, preview, and comprehension level may have a decisive impact on the outcome of listening tests. To examine the potential interaction between these test features, the following research questions were developed.

1. Do participants' scores vary according to the MCQ presentation and preview condition they complete the test under?
2. Do MCQ presentation and preview conditions interact with the question comprehension level?
3. Do interaction effects vary according to participants' listening proficiency levels and item difficulty levels as measured by the listening test?

IV Method

1 Participants

Participants were 160 students enrolled in English preparatory courses at a Turkish university. The participants' ages were between 18 and 21 years old and all had Turkish as their first language. At the time of the study, most participants had been studying English for approximately nine years. In this context, English is taught as a foreign language and the opportunity to communicate in English is largely limited to language classrooms and 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).

2 Materials

a The listening test. The listening test was developed to reflect the communicative contexts participants were likely to encounter during their undergraduate studies. It was deemed important that the test include both academic monologues, e.g. in the form of presentations and lectures, and dialogues, e.g. in the form of seminar discussions and group project discussions. The listening test therefore involved five monologue texts and five dialogue texts. 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. To reflect the stylistic conventions in the target context, a script was written for the monologues texts and improvised role plays were used for the dialogue tasks. The text topics were designed to be accessible and equally familiar to all

participants. However, when developing L2 assessments it is important to acknowledge that variation between the levels of background knowledge test takers have is inevitable (Wallace, 2020). It is the test developer's responsibility to minimize the effects of this variation by selecting topics that do not require specialist knowledge and efforts were made to ensure that no such knowledge was required to complete the tasks used in this study. The monologue texts were written by the researcher and comprised the following situations: a lecture about history, a student presentation about advertising, a voice message about a group project, the itinerary for a university trip, and a lecture about statistics. To produce the dialogues, a social context (a seminar discussion, preparing a presentation, selecting an elective course, discussing a lecture, consulting a family member about an assignment) was described to two teachers who each assumed a role to play in the discussion. The different approaches to the production of the sound files resulted in spoken texts that varied structurally and stylistically. Whereas the monologue sound files were organized and featured clearly defined boundaries between words and between clauses, the dialogue sound files contained various features of interactional connected speech, such as overlap between speakers, false starts, contracted forms, situational ellipsis and hesitations. The sound files included in the test may thus be regarded as reflecting the continuum of spoken texts discussed at length in the literature review (Wagner & Ockey, 2018). The texts were recorded using Audacity (2.0.6, 29 September 2014, http://audacity.sourceforge.net) and all sound files lasted between one and two minutes in duration. The MCQs were produced by the researcher from the sound files. Explicit information comprehension questions ($n = 15$) were based on explicitly stated details (Becker, 2016; Koyama et al., 2016), e.g.

How do the speakers describe the lecture?

A. interesting

B. amusing

C. offensive

Implicit information questions were developed to assess participants' ability to recognize attitude and purpose, main ideas and make inferences, e.g.

What was the subject of the lecture?

A. philosophy and law

B. politics and media

C. government and education

The MCQs were recorded by a speaker that was unfamiliar with the texts. The literature review indicated that it was important to limit the amount of information in the MCQs to prevent overloading working memory in the aural conditions. Each MCQ therefore

Table 1. Test booklet by condition.

Condition	Description	Booklet sample
1	Full written preview of question stems and options	*1. What can we understand about the late book?* *A. It is expensive.* *B. It is new.* *C. It is damaged.*
2	Written preview of question stems Options are presented in speech as part of the sound file after the input text	*1. What can we understand about the late book?* *A.* *B.* *C.*
3	Full spoken preview of question stems and options	*1.* *A.* *B.* *C.*
4	No preview of question stems or options. Questions are spoken as part of the sound file and are presented after each sound file is complete.	*1.* *A.* *B.* *C*

contained three options involving a maximum of three words. Five seconds silence was added between each spoken question and ten seconds silence between each sound file. After the test had been constructed, the categorization of MCQs as assessing implicit and explicit information and the answer key were evaluated by two teachers in the institution and agreement was 100%. A pilot study involving 30 participants studying in classes targeting the same proficiency level was conducted and results indicated that the test recorded a reliability statistic of $\alpha = .85$.

b Conditions. In a between-participants design, participants completed the test under one of four conditions. Four different test booklets were developed for the conditions (see Table 1). Condition 1 represents the format that is conventionally followed in listening tests. This condition involved a full written preview with all 30 MCQ stems and options printed in text. Condition 2 involved a written preview of MCQ stems and aural processing of options after each sound file finished. Condition 3 involved aural preview of MCQ stems and options, which were also repeated once each sound file was complete. Condition 4 did not involve any form of preview and participants listened for the question stems and options which were presented after each sound file had finished.

c Proficiency test. Participants were placed into the same ability level (B1) using an internally produced language test that is benchmarked to the CEFR at the B2 level; a passing grade indicates that the test taker has attained the B2 level of ability (Council of Europe, 2001). The test featured 100 multiple-choice items assessing reading and listening comprehension, lexical and grammatical ability, and integrated reading into writing,

independent writing and independent speaking skills. The participants received similar scores on each component of the test. The participants may therefore be regarded as having very similar levels of listening ability and also comparable levels of lexical knowledge. It was important to establish that participants shared comparable levels lexical knowledge because the literature indicates test takers are likely to attempt to engage lexical matching strategies to complete listening tests and an imbalance in terms of lexical ability would likely skew results.

3 Procedure

The participants were assigned to the four conditions evenly so that there would be equal numbers of test samples in each condition. The final breakdown of participants ($n = 160$) by condition was condition 1 ($n = 40$), condition 2 ($n = 40$), condition 3 ($n = 40$), condition 4 ($n = 40$). The test was administered in small groups in the participants' classrooms using a speaker system. The participants completed the test by marking their responses onto the test booklet.

4 Analysis

Responses to the MCQs were examined using *Facets*, computer software for conducting many facet Rasch measurement (3.71.4, 18 January 2014, www.winsteps.com). *Facets* calibrates the variables of interest (i.e. the test facets) to a single interval scale of the latent trait, which is referred to as the logit scale (Linacre, 2019). The result of this calibration is that all test facets share the same measurement and each individual element within the facets (e.g. an individual participant, a presentation condition) is assigned a logit measure value for comparisons with other elements. In the current study, four facets were examined in the analysis: participants' listening ability as measured by their responses to the test items, the presentation and preview of MCQs, the comprehension level required to answer the MCQ, and 30 MCQs. To produce an unambiguous assessment of the participants' listening ability, it was important to prevent the other test facets (i.e. the presentation and preview and comprehension conditions) from influencing the measurement. For this reason, the presentation and preview variable and the comprehension level variable were entered in to the model as dummy facets, meaning that these facets were anchored at 0 logits. Dummy facets are useful for investigating interaction effects in the data without distorting the measurement of test takers (Engelhard, 2009). Comparisons between the logit measure values associated with these facets were conducted to ascertain the extent to which scores on explicit and implicit information questions were determined by the presentation and preview of the MCQs. In addition, *Facets* calculates Welch's (1951) *t*-tests to test the statistical significance of interactions. The significance level was set at $p = .05$ to avoid type two error and Cohen's (1988) *d* values were calculated to establish effect sizes. To derive measure values to make comparisons between the four preview and presentation conditions, a second model was specified in which this facet was unanchored and a second analysis was completed using Facets. Comparisons between the measure values were made using Welch's *t*-tests and effect sizes were calculated.

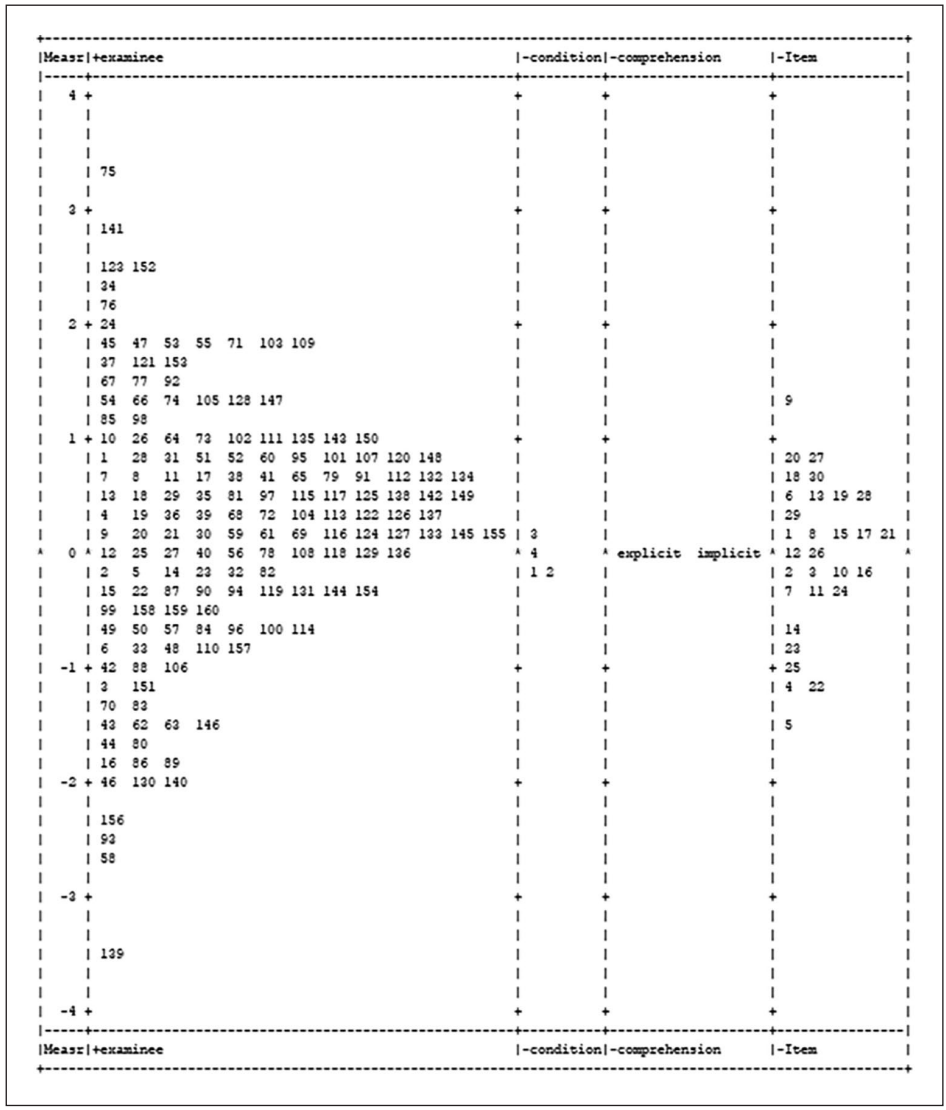

Figure 1. Wright map.

V Results

The test results are first presented in a Wright map (see Figure 1). The map rank orders the individual elements of each facet by placing high scoring participants, the difficult conditions and comprehension level and challenging items toward the top of the figure. Items and participants that appear in the same horizontal position on the Wright map are well matched, indicating that the participant has a 50% chance of responding correctly to the

Table 2. Item and participant measure, fit and reliability and statistics.

Measures	Item	Participant
Mean	86.9	16.3
SD	19.3	6.4
SE	.18	.44
RMSE	.18	.45
Adjusted (True) SD	.62	1.06
Infit MS:		
Mean	1.00	1.00
SD	.11	.11
Outfit MS:		
Mean	1.00	1.00
SD	.17	.18
Separation	3.37	2.35
Strata	4.83	3.47
Reliability of separation	.92	.85

item (Linacre, 2019). The spread of the participant measure values indicates that there was a wide range of scores on the test. Furthermore, the ordering of conditions implies that there was a difference in terms of the challenge associated with each test condition. In contrast, the two elements of the comprehension level facet are positioned identically in the map suggesting that there was little difference in overall difficulty between explicit and implicit questions. Finally, there is a spread of item measure values in the map indicating that the items differed in terms of the level of challenge they posed. Based on the distribution of participant ability measures and item difficulty measure, the range of ability in the test taking population exceeded the range of challenge in the test items.

Facets generates model fit and reliability statistics that may be consulted to establish whether the test is suitably reliable to generate an accurate account of test takers' language ability. Separation and strata values may be assessed to gauge the number of ability levels and the number of difficulty levels participants and items were placed into on the basis of test results. The strata statistic records extreme performances (e.g. participants that record very high or low scores on the test) as representative of true variation in the test taking population and results in a higher estimate than the separation statistic, which explains values at the far boundaries of the logit scale as a result of measurement error (Linacre, 2019).

Table 2 presents the overall test reliability and fit statistics and Table 3 reports on the individual items. Overall, fit statistics that fall within a range of .70 to 1.30 fit the Rasch model expectations and are considered productive for measurement. Item infit statistics ranged from .81 to 1.23 (mean = 1.00, SD = .11). Regarding the participant infit statistics, the mean infit statistic was 1.00 and the standard deviation was .11. The data can thus be regarded as meeting the expectations of the Rasch measurement model. Reliability statistics confirmed the classification of participants and items into different levels was reliable. Item separation and strata statistics suggest that the items were categorized into three distinct levels of challenge (strata = 4.83). In contrast, participants were separated

Table 3. Item focus, difficulty, standard error and fit statistics.

Number	Type	Measure	SE	Infit *MS*	Outfit *MS*	Total Score	Separation
5	Explicit	–1.4	0.21	1.08	1.09	126	
4	Implicit	–1.22	0.21	1.19	1.27	122	
22	Explicit	–1.1	0.2	0.8	0.73	119	
25	Explicit	–0.9	0.2	0.91	0.94	114	
23	Explicit	–0.79	0.19	0.84	0.79	111	Level 1
14	Implicit	–0.72	0.19	0.94	0.86	109	
11	Implicit	–0.37	0.18	0.98	0.98	99	
7	Explicit	–0.34	0.18	1.04	0.99	98	
24	Implicit	–0.31	0.18	0.88	0.83	97	
10	Explicit	–0.18	0.18	1.01	1.06	93	Level 2
2	Implicit	–0.14	0.18	1.05	1.06	92	
16	Implicit	–0.14	0.18	1.05	1.08	92	
3	Explicit	–0.11	0.18	1.23	1.44	91	
26	Implicit	–0.11	0.18	0.92	0.89	91	
12	Explicit	0.11	0.18	1.02	1.1	84	
21	Implicit	0.11	0.18	0.91	0.88	84	
15	Implicit	0.17	0.18	0.97	0.91	82	
8	Explicit	0.24	0.18	1.08	1.12	80	
17	Explicit	0.24	0.18	1.11	1.15	80	
1	Explicit	0.27	0.18	1.23	1.39	79	
29	Explicit	0.3	0.18	0.85	0.81	78	Level 3
6	Implicit	0.39	0.18	1.04	1.03	75	
19	Explicit	0.49	0.18	1.05	0.97	72	
13	Explicit	0.61	0.18	1.07	1.12	68	
28	Explicit	0.64	0.18	0.88	0.86	67	
30	Implicit	0.64	0.18	0.91	0.9	67	
18	Implicit	0.71	0.18	1.03	1.04	65	
20	Implicit	0.8	0.18	0.99	0.94	62	
27	Implicit	0.8	0.18	0.81	0.73	62	
9	Implicit	1.32	0.19	1.04	1.07	47	

into two ability levels (strata = 3.47). To identify the items and participants that are grouped together using the separation statistic, the root-mean-square error (RMSE) is multiplied by three (items: .18 × 3 = .54, participants: .45 × 3 = 1.35) and the resulting value defines the range of the statistically distinct level on the logit scale (Wright & Masters, 2002). Applying this calculation, 90 participants were categorized as high scoring and 56 were categorized as low scoring. The remaining 14 participants recorded measure values that did not fall into the range described by the separation statistic. The items were separated into a simple level (level one, $n = 5$), a mid-level (level two, $n = 11$), and a challenging level (level three, $n = 9$; see Table 3). The remaining five items were associated with extreme measure values and did not amount to one statistically distinct level of separation.

Table 4. Comparisons between condition measure values.

Condition	Measure value	Cohen's *d*			
		1	2	3	4
1	−.09	×	.35	.78*	.39
2	−.23	.35	×	1.10*	.71*
3	.25	.78*	1.10*	×	.39
4	.08	.39	.71*	.39	×

Note. * significant at $p = .05$.

Table 3 presents the spread of item difficulty by reporting measure values indicating the position of each item on the logit scale. Item 5 was the least challenging item and recorded a measure value of -1.40 whereas item 9 recorded a value of 1.32 and was the most challenging. The implicit MCQs recorded a mean measure value of -.07, corresponding to a total score of 1,285 out of 2,400. The explicit information MCQ mean measure value was .07, which represents a total score of 1,321. This suggests that the difference in challenge between the explicit and implicit information items was relatively minor.

To answer research question one, condition 2 resulted in the highest mean score of 17.9 ($SD = 7.5$), followed by condition 1 (mean = 17.4, $SD = 4.4$), condition 4 (mean = 15.7, $SD = 6.7$) and condition 3 (mean = 14.2, $SD = 5.9$). Table 4 reports the results of the comparison between the four condition measure values on the logit scale to determine which test condition was associated with the highest performance. The results of the Welch's *t*-test indicate the difference in measure values between conditions 2 and 3 was statistically significant at ($t(78) = 4.8487$) $p < .01$, with a large effect size. Comparisons between conditions 2 and 4 and 1 and 3 were also statistically significant at ($t(78) = 3.1315$) $p < .01$ and ($t(78) = 3.4345$) $p < .01$ respectively, with large effect sizes. The remaining comparisons between conditions 1 and 2 ($t(78) = 1.5185$, $p = .13$), conditions 1 and 4 ($t(78) = 1.7173$, $p = .089$), and 3 and 4 ($t(78) = 1.7173$, $p = .089$) did not reach statistical significance.

To answer research question two, the comparison of within conditions explicit and implicit logit measure values is presented in Table 5. The table presents the implicit and explicit mean measure values recorded under each test condition and the associated standard errors. The size of the difference between implicit and explicit scores within the conditions can best be understood by examining the contrast column, which presents the difference between mean measure values. The largest difference of -.27 was between implicit and explicit measure values recorded under condition 2. The smallest difference of -.01 was recorded under condition 3. The following columns present the results of the Welch *t*-tests and effect sizes. The results of the Welch *t*-tests did not reach statistical significance indicating that there was no difference between explicit and implicit scores within test conditions.

The between conditions analysis of implicit and explicit logit measure values is presented in Table 6. The table contrasts the mean measure values recorded for both explicit and implicit items under the four conditions and reports the results of the Welch *t*-tests.

Table 5. Comparisons between comprehension level measure values within conditions.

	Implicit		Explicit		Contrast	SE	Welch's *t*-test			*d*
	Measure	SE	Measure	SE			*t*	*df*	*p*	
C1	.02	.09	−.03	.09	.05	.13	.39	1197	.70	.02
C2	−.13	.10	.13	.10	−.27	.14	−1.88	1197	.06	.10
C3	−.01	.09	.01	.09	−.01	.13	−.09	1197	.93	.01
C4	−.05	.10	.05	.10	−.11	.13	−.80	1197	.42	.04

The contrast statistic summarizes the difference between implicit and explicit mean measure values under the different conditions. For instance, the largest difference was between explicit measure values recorded under conditions 1 and 2, whereas the smallest difference was between implicit measure values recorded under conditions 1 and 3. The results of the Welch *t*-tests were not significant indicating that implicit and explicit scores did not vary between the conditions.

The test separated participants into two statistically distinct levels of listening ability and test items into three statistically distinct levels of challenge. This separation presented an opportunity to investigate potential interactions between participant ability (as estimated by the test), the test condition and the item separation level, and secondly the item challenge level, the test condition and the item comprehension level (see research question three). To investigate these interactions a series of analyses were completed in which the measurement model was altered to include only relevant data (e.g. the participants belonging to the higher ability level were deleted to analyse the lower scoring participant measure values). This generated measure values for each facet under investigation and comparisons between the values were made using Welch *t*-tests.

To begin with the results of the analysis investigating interactions between participant ability, test condition and item comprehension, results showed test condition and item comprehension interacted at the lower ability level only. Low scoring participants completing the test under condition 2 recorded a measure value of -.35 on the implicit information questions and .35 on the explicit information questions and the Welch *t*-test demonstrated that the difference of .70 on the logit scale was statistically significant at ($t(267) = -2.52$) $p = .05$, with a small effect size of .31. The remaining interactions were not statistically significant (for results, see Appendices 1 and 2). To summarize this result, lower scoring participants recorded higher scores on implicit items than explicit items when completing the test under condition 2.

Turning to the results of the analysis of interactions between the item separation level, the test condition and the comprehension level, participants that completed mid-level items (the second separation level) under condition 2 recorded a measure value of -.30 on the implicit items and a measure value of .19 the explicit items. The difference amounted to .49 on the logit scale and the result of the Welch *t*-test was statistically significant at ($t(341) = -2.00$) $p = .05$, with a small effect size of .22. The remaining interactions did not reach statistical significance (for results, see Appendices 3 and 4). To summarize this result, participants completing the second separation level items under test condition 2 scored higher on implicit items that explicit items.

Table 6. Comparisons of implicit and explicit measure values between conditions.

Comprehension	Condition	Measure	SE	Condition	Measure	SE	Target contrast	Joint SE	*t*	*df*	*p*
Explicit	1	−.03	.09	2	.13	.10	−.16	.14	−1.18	1195	.24
Implicit	2	−.13	.10	3	−.01	.09	−.13	.14	−.93	1196	.35
Explicit	1	−.03	.09	4	.05	.10	−.08	.13	−.60	1197	.55
Implicit	2	−.13	.10	4	−.05	.10	−.08	.14	−.57	1197	.57
Explicit	3	.01	.09	4	.05	.10	−.05	.13	−.36	1197	.72
Explicit	1	−.03	.09	3	.01	.09	−.03	.13	−.24	1197	.81
Implicit	1	.02	.09	3	−.01	.09	.03	.13	.23	1197	.82
Implicit	3	−.01	.09	4	−.05	.10	.05	.13	.36	1197	.72
Implicit	1	.02	.09	4	−.05	.10	.08	.13	.60	1197	.55
Explicit	2	.13	.10	4	.05	.10	.08	.14	.58	1197	.56
Explicit	2	.13	.10	3	.01	.09	.13	.14	.94	1196	.35
Implicit	1	.02	.09	2	−.13	.10	.16	.13	1.17	1194	.24

VI Discussion

This study set out to establish the effect of adopting alternative MCQ presentation and preview formats in a test of second language listening ability. The primary motivating factor for the investigation was to create a listening test that imitated the language processing listeners are likely to engage in the target context more closely than conventional assessment formats by reducing the reading component involved in the test and simultaneously maintaining the ability to provide test takers with a contextualized focus through preview of MCQ contents. Variation in the MCQ format along presentation and preview lines was predicted to interact with the comprehension level required by the items by reducing the potential for test takers to engage lexical matching strategies to complete the test. Results confirmed that variation in the preview and presentation of MCQs impacted test scores and participants recorded the highest scores when a written preview of question stems was included in the test. There was also an interaction between presentation and preview variables and item comprehension level in the low scoring participants' item responses whereby scores were higher on implicit information questions than explicit information questions under written preview of MCQ stems and verbal presentation of options.

The finding that the highest scores were associated with preview of the question stems supports conclusions reached by Yanagawa and Green (2008) that this format may provide a facilitative contextualizing focus. The beneficial effect of this focus is observable in responses to both the explicit and implicit information questions. In the former listeners are made aware of specific details to listen out for and in the latter the implicit aspects of the text they will be tested on. The facilitative effect of preview is most evident when the preview does not extend to MCQ options and is printed in the test booklet. This effect may be explained by instances of lexical overlap between the distracters in the MCQ options and the contents of the sound files. Participants relying on basic lexical matching strategies heard the contents of distracters in addition to the answer key in the input text and may have selected the option that was most prominent in the text. For instance, during one of the dialogue tasks, participants listened to two students discussing the elective foreign language courses available at a university and the benefits of learning Spanish, Chinese, and Japanese. The text was an improvised, informal discussion and involved various features of connected speech that test takers may have found unfamiliar and difficult to process (Wagner, 2018). Based on this input, participants were required to answer the following question:

Which language did the speaker study in high school?

A. Spanish

B. Chinese

C. Japanese

For test participants completing this item under condition 1, the number of operations they were required to simultaneously perform involved processing the ongoing speech, monitoring comprehension, anticipating the answer to following questions in the text,

and concurrently dividing attention between the written MCQs and the sound file. This may have imposed a substantial cognitive burden especially if participants were encouraged to attempt bottom-up processing of the sound files by seeking to match the contents of the options with the contents of the texts (Field, 2013). In contrast, the need to retain the MCQ contents in working memory and concurrently devote attentional resources to the ongoing speech in condition 3 may have proved too demanding and resulted in the particularly low scores under this condition. Incorrect responses to items collected under these conditions provide little diagnostic feedback for score users aiming to measure listening ability because it is not possible to disentangle the response from the various cognitive operations required to provide it. In contrast, it seems the provision of written stems only under condition 2 may have encouraged participants to adopt a top-down approach that was sufficient to focus attention on the relevant parts of the text without draining resources by dividing attention between the written MCQs and continuously processing the sound file (Vandergrift, 2007). Scores generated under this condition are more likely to be attributable to second language listening ability than irrelevant factors and will thus provide more relevant feedback to assessment score users.

The effect of varying MCQ formats is clear in items categorized as level two (i.e. the mid-level of item challenge) in the analysis. These items were distributed within a range of -.18 to .27 on the logit scale (the mean value is 0 logits) and are hence well matched to the average ability level in the test taking population (the mean value is .21 logits; see Figure 1). The effect of stem preview does not influence participants in the same way as they respond to less challenging items (separation level 1), which may have generally been completed correctly regardless of the support provided by the presentation and preview of the item. By the same token, items that are difficult (separation level 3) for participants to complete do not seem to be affected by variation in the MCQ format, perhaps because the relevant section of the sound file involving the information was not successfully processed or contained language that was unfamiliar. For these reasons, items that are well suited to the test taking population seem to be most susceptible to variation in preview and presentation.

In conventional tests of listening, score variation may partially be attributable to variation in reading ability and for this reason reading proficiency may be considered a source of construct irrelevant variance. Reducing the impact of construct irrelevant variance is an important aspect of test development because it increases the likelihood that test scores are representative of the construct of interest (Aryadoust, 2012). This means score users such as teachers and students can have more trust that decisions based on test scores are appropriate. Despite claims made in the literature, the findings of this study show that eliminating reading completely from a listening test leads to low overall scores. Although Chang and Read (2013) argue that the verbal presentation of MCQs results in a pure form of listening assessment, it would be difficult to support and defend this form of assessment to test stakeholders in the knowledge that it results in the lowest scores on the test (Swain, 1985). Nonetheless, the results of this study indicate that the amount of reading required in a listening test can be reduced substantially to short question stems only without making the test unnecessarily challenging. Indeed, the findings suggest that presenting stems only may elicit the best performance from test takers.

The literature indicates that lexical matching between the MCQ options and the sound file is common in test of second language listening and that scores on explicit information questions might be highest when preview of the options is available. Indeed, this is a presumption that is commonly made by developers of listening assessments and researchers working in the field (Field, 2019). In the current study, participants were categorized according to their measure values into different levels of ability and results indicated that interactions between condition and comprehension level were evident only in the responses provided by the lower scoring participants. At this level of ability, participants may have benefitted from lexical matching when responding to the explicit information items. This finding supports conclusions reached by Chang and Read (2013) that less proficient test takers may benefit from having written comprehension questions. However, under condition 2 this opportunity was removed and participants did not achieve high scores on the explicit information questions. In contrast, when completing the implicit information questions, the removal of the MCQ options may have freed up attentional resources that would otherwise have been consumed processing options and low scoring participants could devote more attention to the sound file. Under this condition, participants were provided with a contextualizing focus and could concentrate on extracting relevant information from the text without simultaneously processing MCQ options. To generalize from this, low achieving students may benefit from preview of item stems only when completing listening tasks requiring comprehension of implicit information. This finding has pedagogical implications for teachers developing listening comprehension tasks because underperforming students may be more motivated to complete the alternative task format exemplified in condition 2 (Chow et al. 2018).

The findings of this study offer possible directions for future research. Having established the impact on scores of varying MCQ formats, it is vital to validate new assessment formats, and especially with regards to condition 2, by collecting test taker accounts of their processing of the listening test under the different conditions. This kind of validity evidence may be gathered using verbal protocol approaches such as stimulated recall and could be used to investigate the extent to which MCQ presentation and preview impacts upon the test takers' reported strategy use (Badger & Yan, 2009). It would also be worthwhile to investigate the effect of variation in MCQ formats with more proficient learners of English for whom many of the cognitive operations instigated by the test are relatively automatic and hence may have extra working memory capacity available to allocate during the test (Field, 2019).

VII Conclusions

The current study examined the impact of variation in the presentation of MCQ formats and potential interactions with item comprehension level in a test of second language listening ability. The results are likely to be relevant to developers of language tests tasked with assessing prospective undergraduate students in similar contexts (Dimova et al., 2020). However, the results are also generalizable to classroom contexts where the MCQ format is commonly applied in formative assessment to inform low-stakes decisions relating to language learners' comprehension of listening passages and to gauge their progress as listeners. To that end, teachers should be aware that a student's failure

to provide a correct response to MCQs about a listening passage may not be directly attributable to comprehension failure but may instead represent a test method effect. A more reliable gauge of students' ability to comprehend explicitly stated and implicit information in a listening passage using MCQs may be to present options verbally. Furthermore, the high scores associated with condition 2 entail that completing classroom listening tasks using this format may lead to a positive washback effect by decreasing listening anxiety and increasing student motivation to participate in listening tasks (Buck, 2001). Finally, it is broadly recognized that all forms of assessment should aim to elicit the best possible performance from test takers (Swain, 1985). In second language listening assessments, this may be achieved by presenting only question stems for preview in the test booklet and presenting response options verbally as part of the sound file once the relevant information has been presented.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Stefan O'Grady https://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*, 40–60.

Aryadoust, V., Goh, C.C.M., & Kim, L.O. (2012). Developing and validating an academic listening questionnaire. *Psychological Test and Assessment Modeling*, *54*, 227–256.

Audacity Team. (2014). *Audacity(R): Free audio editor and recorder: Version 2.0.6* [computer program]. Available at: http://www.audacityteam.org/download (Version 2.4.2; accessed January 2021).

Badger, R., & Yan, X. (2009). The use of tactics and strategies by Chinese students in the Listening component of IELTS. In Thompson, P. (Ed.), *International English Language Testing System (IELTS) Research Reports 2009: Volume 9* (pp. 67–98). Canberra: British Council and IELTS Australia. Available at: https://search.informit.com.au/documentSummary;dn=070356543696560;res=IELHSS (accessed January 2021).

Bardovi-Harlig, K. (2013). Developing L2 pragmatics. *Language Learning*, *63*, 68–86.

Batty, A.O. (2018). Investigating the impact of nonverbal communication cues on listening item types. In Ockey, G., & E. Wagner (Eds.), *Assessing L2 listening moving toward authenticity* (pp. 161–179). Philadelphia, PA: John Benjamins.

Becker, A. (2016). L2 students' performance on listening comprehension items targeting local and global comprehension. *Journal of English for Academic Purposes*, *24*, 1–13.

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

Carter, R., & McCarthy, M. (1997). *Exploring spoken English*. Cambridge: Cambridge University Press.

Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide: Spoken and written English grammar and usage*. Cambridge: Cambridge University Press.

Chang, A., & Read, J. (2013). Investigating the effects of multiple-choice listening test items in the oral versus written mode on L2 listeners' performance and perceptions. *System*, *41*, 575–586.

Chikalanga, I. (1992). A suggested taxonomy of inferences for the reading teacher. *Reading in a Foreign Language*, *8*, 697–709.
Chow, B.W.Y., Chiu, H.T., & Wong, S.W.L. (2018). Anxiety in reading and listening English as a foreign language in Chinese undergraduate students. *Language Teaching Research*, *22*, 719–738.
Clark, M. (2014). The use of semi-scripted speech in a listening placement test for university students. *Papers in Language Testing and Assessment*, *3*, 1–26.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.
Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
Dimova, S., Yan, X., & Ginther, A. (2020). *Local language testing: Design, implementation, and development*. London: Routledge.
Engelhard, G. (2009). Using item response theory and model-data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement*, *69*, 585–602.
Field, J. (2013). Cognitive validity. In Geranpayeh, A., & L. Taylor (Eds.), *Examining listening: research and practice in assessing second language listening studies in language testing 35* (pp. 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.
Kang, T., Arvizu, M., Chaipuapae, P., & Lesnov, R. (2019). Reviews of academic English listening tests for non-native speakers. *International Journal of Listening*, *33*, 1–38.
Kim, B. (2015). The effects of working memory span on listening tests without preview questions. *Language Research*, *51*, 403–420.
Koyama, D., Sun, A., & Ockey, G. (2016). The effects of item preview on video-based multiple-choice listening assessments. *Language Learning and Technology*, *20*, 148–165.
Li, C., Chen, C., Wu, M., et al. (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*, 98–112.
Linacre, J.M. (2019). *Winsteps Rasch measurement computer program user's guide*. Beaverton, OR: Winsteps.com
Linacre, J.M. (2019). *Winsteps: Version 4.4.6* [computer software]. Beaverton, OR: Winsteps.com. Available at: https://www.winsteps.com (accessed January 2021).
O'Sullivan, B. (2016). Validity: What is it and who is it for? In Leung, Y. (Ed.), *Epoch making in English teaching and learning: Evolution, innovation, and revolution* (pp. 201–222). Taipei: Crane Publishing.
O'Sullivan, B., & Weir, C. (2011). Language testing and validation. In O'Sullivan, B. (Ed.), *Language testing theory and practice* (pp. 13–32). Oxford: Palgrave.
Rost, M. (2011). *Teaching and researching listening*. Harlow: Pearson.
Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question type. *Language Testing*, *8*, 23–40.
Sherman, J. (1997). The effect of question preview in listening comprehension tests. *Language Testing*, *14*, 185–213.
Swain, M. (1985). Large scale communicative testing: A case study. In Lee, Y., Fok, C., Lord, R., & G. Low (Eds.), *New directions in language testing* (pp. 35–46). Hong Kong: Pergamon.
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*, 517–539.

Taylor, L., & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalizing the test construct. *Journal of English for Academic Purposes*, *10*, 89–101.

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

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*, 178–195.

Wagner, E. (2016). Authentic texts in the assessment of L2 listening ability. In Banerjee, J. V., & D. Tsagari (Eds.), *Contemporary Second Language Assessment* (pp. 103–123). London: Blommsbury Academic.

Wagner, E. (2018). A comparison of listening performance on tests with scripted or authenticated spoken texts. In Ockey, G., & E. Wagner (Eds.), *Assessing L2 listening moving toward authenticity* (pp. 29–44). Amsterdam: John Benjamins.

Wagner, E., & Ockey, G. (2018). An overview of the use of authentic, real-world spoken texts on L2 listening tests. In Ockey, G. & E. Wagner (Eds.), *Assessing L2 listening moving toward authenticity* (pp. 13–28). Amsterdam: John Benjamins.

Wallace, M.P. (2020). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language Learning*. Epub ahead of print 8 July 2020. DOI: 10.1111/lang.12424.

Wang, Y., & Treffers-Daller, J. (2017). Explaining listening comprehension among L2 learners of English: The contribution of general language proficiency, vocabulary knowledge and metacognitive awareness. *System*, *65*, 139–150.

Weir, C. (2005). *Language testing and validation*. Basingstoke: Palgrave Macmillan.

Welch, B. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*, 330–336.

Wright, B.D., & Masters, G.N. (2002). *Rasch measurement transactions: Number of person or item strata: (4*Separation+1)/3. 2002*. Available at: http://www.rasch.org/rmt/rmt163f.htm (accessed January 2021).

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

Yeom, S. (2016) The effects of presentation mode and item type on L2 learners' listening test performance and perception. *English Teaching*, *71*, 27–54.

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*, 21–44.

Appendix 1. Interactions between proficiency, condition and comprehension measure values.

Condition	Measure value	Cohen's *d*			
		1	2	3	4
Higher scores: Implicit:					
1	.06	×	.04	−.08	.05
2	−.03	.04	×	.04	.01
3	−.13	.08	.04	×	.03
4	−.06	.05	.01	.03	×
Higher scores: Explicit:					
1	−.07	×	.02	.12	.05
2	.02	.04	×	.04	.01
3	.12	.09	.04	×	.03
4	.05	.05	.01	.03	×
Lower scores: Implicit:					
1	−.01	×	.16	.05	.02
2	−.35	.16	×	.21	.15
3	.10	.05	.21	×	.06
4	−.04	.02	.15	.06	×
Lower scores: Explicit:					
1	.01	×	.16	.04	.02
2	.35	.16	×	.21	.14
3	−.08	.04	.21	×	.06
4	.05	.02	.14	.06	×

Note. * significant at $p = .05$.

Appendix 2. Comparisons between comprehension level measure values within conditions by proficiency.

	Implicit		Explicit		Contrast	SE	Welch's *t*-test			*d*
	Measure	SE	Measure	SE			*t*	*df*	*p*	
Higher scores:										
C1	.06	.11	−.07	.12	.13	.17	.78	747	.44	.06
C2	−.03	.12	.02	.12	−.05	.17	−.29	807	.77	.02
C3	−.13	.14	.12	.14	−.25	.19	−1.30	537	.20	.11
C4	−.06	.14	.05	.13	−.11	.19	−.58	597	.56	.05
Lower scores:										
C1	−.01	.15	.01	.15	−.02	.21	−.08	417	.94	.01
C2	−.35	.19	.35	.20	−.70	.28	−2.52	267	.01	.31
C3	.10	.14	−.08	.13	.18	.19	.94	537	.35	.08
C4	−.04	.14	.05	.15	−.10	.21	−.47	447	.64	.05

Appendix 3. Interactions between item level, condition and comprehension measure values.

Condition	Measure value	Cohen's *d*			
		1	2	3	4
Level 1: Implicit:					
1	−.07	×	.07	.09	.12
2	−.23	.07	×	.16	.18
3	.13	.09	.16	×	.03
4	.18	.12	.18	.03	×
Level 1: Explicit:					
1	.11	×	.10	.14	.10
2	.33	.10	×	.24	.21
3	−.20	.14	.24	×	.04
4	−.12	.10	.21	.04	×
Level 2: Implicit:					
1	.17	×	.21	.10	.13
2	−.30	.21	×	.11	.09
3	−.06	.10	.11	×	.02
4	−.11	.13	.09	.02	×
Level 2: Explicit:					
1	−.11	×	.14	.07	.17
2	.19	.14	×	.07	.03
3	.04	.07	.07	×	.10
4	.26	.17	.03	.10	×
1	−.11	×	.01	.06	.01
2	−.09	.01	×	.05	.01
3	.02	.06	.05	×	.07
4	−.13	.01	.01	.07	×
1	.13	×	.01	.07	.01
2	.11	.01	×	.06	.02
3	−.02	.07	.06	×	.06
4	.10	.01	.00	.06	×

Note. * significant at $p = .05$.

Appendix 4. Comparisons between comprehension level measure values within conditions by item level.

	Implicit		Explicit		Contrast	SE	Welch's *t*-test			*d*
	Measure	SE	Measure	SE			*t*	*df*	*p*	
Level 1:										
C1	−.07	.24	.11	.30	−.18	.39	−.48	135	.64	.08
C2	−.23	.27	.33	.33	−.56	.43	−1.32	112	.19	.25
C3	.13	.22	−.20	.28	.33	.36	.92	151	.36	.15
C4	.18	.30	−.12	.24	.30	.38	.78	129	.43	.14
Level 2:										
C1	.17	.17	−.11	.14	.28	.23	1.25	365	.21	.13
C2	−.30	.19	.19	.15	−.49	.24	−2.00	341	.05	.22
C3	−.06	.18	.04	.14	−.10	.23	−.43	363	.66	.05
C4	-.11	.13	.26	.20	−.37	.24	−1.52	275	.13	.18
Level 3:										
C1	−.11	.16	.13	.18	−.24	.24	−1.01	347	.32	.11
C2	−.09	.18	.11	.20	−.20	.28	−.73	269	.47	.09
C3	.02	.17	-.02	.19	.04	.26	.16	313	.87	.02
C4	−.13	.19	.10	.17	−.23	.24	−.91	304	.37	.10