Routledge
Taylor & Francis Group

International Journal of Listening

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hijl20

INVESTIGATING THE ROLE OF RESPONSE FORMAT IN COMPUTER-BASED LECTURE COMPREHENSION TASKS

Stefan O'Grady

To cite this article: Stefan O'Grady (31 Jan 2024): INVESTIGATING THE ROLE OF RESPONSE FORMAT IN COMPUTER-BASED LECTURE COMPREHENSION TASKS, International Journal of Listening, DOI: 10.1080/10904018.2024.2312272

To link to this article: https://doi.org/10.1080/10904018.2024.2312272

INTERNATIONAL JOURNAL OF LISTENING
https://doi.org/10.1080/10904018.2024.2312272

Routledge
Taylor & Francis Group

INVESTIGATING THE ROLE OF RESPONSE FORMAT IN COMPUTER-BASED LECTURE COMPREHENSION TASKS

Stefan O'Grady

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

ABSTRACT

Language assessment is increasingly computermediated. This development presents opportunities with new task formats and equally a need for renewed scrutiny of established conventions. Recent recommendations to increase integrated skills assessment in lecture comprehension tests is premised on empirical research that demonstrates enhanced construct coverage over conventional selected response formats such as multiple-choice. However, the comparison between response formats is underexplored in computer-mediated assessment and does not consider test item presentation methods that this technology affords. To this end, the present study investigates performance in a computer-mediated lecture comprehension task by examining test taker accounts of task completion involving multiple-choice questions without question preview and integrated response formats. Findings demonstrate overlap between the formats in terms of several core processes but also point to important differences regarding the prioritization of aspects of the lecture, memory and test anxiety. In many respects, participant comments indicate the multiple-choice format measured a more comprehensive construct than the integrated format. The research will be relevant to individuals with interests in computer-mediated assessment and specifically with a responsibility for developing and validating lecture comprehension assessments.

Introduction

The lecture comprehension task is a common component of language tests for English medium university admissions. Increasingly, such tests involve integrated listening and speaking response formats in which test takers are required to describe the lecture to a live examiner or computer interface (Choi, 2022; Khabbazbashi et al., 2022). As integrated tasks require language comprehension, information synthesis, and language production, the format permits language testers "to better reflect the demands students face in tertiary studies" than more conventional independent lecture tasks (Frost et al., 2021, p. 133; Inoue & Lam, 2021; Westbrook, 2023). Integrated task types are argued to promote fairness over independent language production tasks by mitigating the impact of variation in background knowledge through the provision of task content (Weigle, 2004). Correspondingly, researchers have concluded that integrated tasks measure comprehension abilities in a way that is "more authentic than item types such as multiple-choice and matching items, by assessing abilities corresponding to those performed outside testing situations" (Rukthong & Brunfaut, 2020, p. 32; Wei & Zheng, 2017).

CONTACT Stefan O'Grady so59@st-andrews.ac.uk International Education Institute, The University of St Andrews, Kinnesburn, Kennedy Gardens, St Andrews KY16 9DJ, UK

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10904018.2024.2312272

Arguments for integrated response formats in lecture comprehension tasks are further premised on perceived flaws of the alternative. At worst, selected response formats, such as multiple-choice questions (MCQ) permit test-wise strategies to complete assessment tasks, reward shallow listening processes and superficial engagement with lecture content, and underrepresent the complexity of listening in academic contexts (Badger & Yan, 2009; Field, 2019; Rukthong, 2021). However, to date much of the research has taken a narrow view on response formats such as MCQs, which may be constructed and presented in various ways to manage test taker engagement with assessment tasks (Chang & Read, 2013; Yanagawa & Green, 2008). The restriction of MCQ item preview in particular has recently been shown to impact directly on test scores; this impact is attributed to a hypothesized association between preview and test-wise strategies such as lexical matching (O'Grady, 2023; Yeager & Meyer, 2022). This research suggests that removal of item preview may resolve limitations associated with MCQs and generate a reliable measure of lecture comprehension. However, empirical evidence in the form of test taker accounts is required to verify this possibility. To this end, the present case study operationalizes a model of second language listening recently proposed by Aryadoust and Luo (2022) to reexamine construct coverage in integrated and multiple-choice response lecture comprehension tasks and to determine the impact of preview restriction on the measurement of listening skills.

Literature Review

The listening construct

Valid language assessment requires identification and unambiguous definition of target knowledge and skills; construct definition is thus a crucial aspect of language assessment development (Chalhoub-Deville & O'Sullivan, 2020). In a recent systematic review, Aryadoust and Luo (2022) identify three approaches to construct definition that conventionally applied in listening assessment research: process-based, subskills-based, and attribute-based approaches. The authors propose a unified model of the listening construct that synthesizes the three approaches. Applying the model, lecture comprehension assessment operationalizes attributes of the assessment task (relating to features of the input, text type, characteristics of the speech, visual information, and task format); characteristics of the test taker (including experiential knowledge, target language proficiency, demographic information, and affective factors); subskills (knowledge of the L2 sound system, vocabulary and syntax to support local meaning, knowledge of discourse and context to support global meaning, and pragmatic competence); and cognitive processes relating to bottom-up and top-down processing, memory, cognitive and metacognitive strategies. The current study applies the Aryadoust and Luo (2022) model to define the listening construct. The investigation is carried out in a computer-mediated environment as communication is increasingly facilitated by technology (Khabbazbashi et al., 2022), and test takers may display different behavior in pen and paper and computer-mediated tests of listening (Coniam, 2006).

Attributes of the assessment task: multiple-choice formats

Attributes of the assessment task such as response format have been the focus of empirical research investigating interactions between task design and test score utility in listening assessment (Aryadoust & Luo, 2022; Buck, 2001; Field, 2019). For instance, research comparing the effects of single and double play in listening tests concludes that double play increases test scores and elicits a wider range of listening processes than single play (Holzknecht, 2019). Research exploring the impact of audio-visual content on listening comprehension has concluded that the opportunity to view a speaker facilitates comprehension over audio-only formats (Batty, 2021). The level of scriptedness of the sound file has been shown to constitute an important component of listening test design because scripted sound files are associated with higher scores but may misrepresent spoken interaction (Wagner et al., 2020). Response format as a test attribute also has an important bearing on the measurement of listening comprehension (O'Grady,

2023). MCQs are associated with high levels of measurement consistency but have also been argued to cause construct irrelevant variance (score variance that cannot be directly attributed to variation in listening proficiency) by inadvertently assessing reading comprehension (Chang & Read, 2013), and encouraging test takers to seek weaknesses in the format to exploit (Field, 2019; Rukthong, 2021). Further sources of construct irrelevant variance and construct underrepresentation associated with MCQs may involve eliminating options without comprehending the source text, guessing, or lexical matching between the source text and item contents (Holzknecht et al., 2020). , test takers report that they prefer selected response formats such as MCQ owing to reduced anxiety and facilitation of strategies, such as guessing or predicting content (H. Cheng, 2008). Broadly, MCQs may induce a problem-solving approach rather than a text comprehension approach that is unique to assessment contexts (Rupp et al., 2006), and provoke what Field (2012, p. 391) has referred to as "test specific behaviour" that limits test score utility.

Such limitations have motivated research efforts to revise the MCQ format (Chang & Read, 2013; Yanagawa & Green, 2008; Yeager & Meyer, 2022). These include presenting MCQs aurally to reduce the impact of reading ability (Chang & Read, 2013; Yeom, 2016) and manipulating the extent to which test takers may preview MCQ content before the sound file (Koyama et al., 2016; Li et al., 2017; Yanagawa & Green, 2008; Yeager & Meyer, 2022). In a review of this literature, O'Grady (2023) argued that test-wise strategies involving lexical matching in listening tests represented a significant threat to test score utility that was more likely to be observed when MCQs focused on assessing information that was made explicit in the sound file and were simultaneously available for preview. Explicit information questions were defined as those requiring comprehension of information that was directly stated such as times, locations, names, and definitions. In contrast, implicit information questions ask about speaker purpose, illocutionary intent, attitude and application. While it is important to measure test candidate's ability to extract information such as names and dates from speech, this ability represents a basic process in Aryadoust and Luo's (2022) model and reflecting the listening needs in the academic domain requires listening tests to engage a wider range of listening processes (Field, 2019). In O'Grady (2023), test takers completed listening tests featuring both explicit and implicit MCQs under different question preview conditions, whereby the preview was completely withheld, limited to stem only, or full preview was provided in either aural or written modes. Results showed that explicit questions were often completed more accurately when participants were able to preview the question contents in full, whereas the implicit questions were less clearly affected by preview. Preview restriction was thus hypothesized to limit the test-wise strategies employed to answer explicit information questions; however, the hypothesis is yet to be tested using data collection methods designed to investigate test taking processes (Cohen, 2011). This is an important focus for research because test validation should go beyond inferring cognitive processes from test scores (Weir, 2005).

Integrated response formats

Accounts in the literature associate integrated response formats with higher levels of construct representation than selected response tasks (Frost et al., 2021, p. 133; Inoue & Lam, 2021; Westbrook, 2023). Based on stimulated recall data, Rukthong and Brunfaut (2020) report that integrated listening assessments activate a range of lower and higher order listening processes and strategies and conclude that the integrated approach involves strong construct coverage. Rukthong (2021) used stimulated recall to compare cognitive processing in integrated listening-into-speaking summary tasks and selected response tasks. The findings suggested that the selected response tasks were completed by employing word recognition and simple syntactic parsing processes, whereas the integrated tasks involved more higher-level processing of semantic and pragmatic aspects of the sources. Construct irrelevant test-wise strategies such as lexical matching, choice deletion and guessing were also associated with the selected response tasks. Overall, the picture that emerges from the literature is that integrated formats increase construct representation over multiple-choice formats, which create misleading impressions of second language listening ability (He & Jiang, 2020).

Integrated response formats are associated with a wider range of listening processes involving engagement with both explicitly stated (specific details) and implicit (propositional and pragmatic) content and thus demonstrate stronger construct coverage. However, empirical comparisons between integrated response and restricted preview multiple choice formats are yet to be made.

An important reason to make this comparison is that integrated skills assessment is mediated by raters, and is hence prone to rater effects, or computer scoring systems, which may neglect pragmatic aspects of integrated task performance and base scores on numerable linguistic parameters (Zechner & Evanini, 2020). Examiners cannot be entirely sure that omission of important source content in the response is due to preference, the prioritizing of information, or comprehension breakdown (Frost et al., 2021). A related concern is the possibility of verbatim copying from the source text which may erroneously inflate scores by creating false impressions of linguistic ability (Plakans, 2015). Empirical research findings indicate that this is certainly a possibility; Crossley and Kim (2019) report that integrated responses with higher numbers of shared words with the source text tended to score higher, and Frost et al. (2021) found that higher scoring test takers generally reproduced more source content with greater levels of accuracy than the lower scoring test takers. In Crossley et al. (2014), the strongest predictor of scores was the total number of words integrated from source texts, whereas Cumming et al. (2005) report that more proficient test takers use more verbatim phrases from source texts, possibly as a result of stronger comprehension and increased availability of attentional resources during listening. In short, the amount of repetition from the source text in the integrated response seems to influence test scores, with clear implications for construct definition and the role of memory in integrated listening and speaking tasks; test takers with higher working memory capacity may score higher on integrated tasks. Finally, integrated listening and speaking tests have been associated with anxiety. While Huang and Hung (2013) found that anxiety impacted independent and integrated performance equivalently and concluded that anxiety may constitute an important source of construct irrelevant variance that is common regardless of the response format, participants in several studies report speaking directly to a computer is stressful and unnatural (Huang et al., 2016; Kormos et al., 2020; Lee & Winke, 2018).

Research Questions

The effect of variation in response format using integrated responses and restricted preview MCQs in computer-mediated lecture comprehension assessments is underexplored. This is an important gap because the limitations associated with MCQs are commonly cited in support of integrated formats but crucially may be mitigated through preview restriction (O'Grady, 2023). Applying the listening construct outlined in Aryadoust and Luo (2022), the current study investigates the extent to which response format determines test taker interaction with listening stimuli in computer-mediated lecture assessments by seeking to answer the following question and sub questions:

1. Do test takers report different listening processes according to variation in response format in a lecture comprehension task?

1a. To what extent do test takers demonstrate a range of listening subskills and cognitive processes in MCQ and integrated response tasks?

1b. To what extent is the MCQ and the integrated response format associated with factors that cause construct irrelevant variance?

Methodology

Participants

Five participants enrolled in postgraduate courses of study at a university in the United Kingdom were recruited to take part in the study. Four participants were from China and spoke Mandarin as their

first language, and one was from Japan and spoke Japanese. Three were female and two were male and ages ranged from 22 to 38. Participants were requested to report the language test and the result they had attained to enter the university; two reported IELTs scores of 7 and the other two of 7.5. One participant had completed a bachelor's degree at a university in the UK and was thus not required to submit a language test score as part of the application to the university.

Materials

Lectures were created using Microsoft PowerPoint and recorded as MP4 files. To create the lectures required information on two general academic topics that would be comprehensible to test takers regardless of their academic background, and not unduly benefit those with background knowledge (Huang et al., 2016). The researcher identified two appropriate topics by searching articles on the *Nature* website. The first article discussed the relationship between collaboration in academic work and diversity (Diversity; Freeman & Huang, 2014), and the second discussed reproducibility of academic research (Reproducibility; Baker, 2016). In addition, Google Scholar was used to identify another text on each topic to provide alternative perspectives on the subject and to facilitate comparisons that could be made in the lecture. In each of the four texts, the researchers reported the results of an empirical study.

During the following stage, the researcher read the texts and made notes of the key content. The notes were logically organized and used to create PowerPoint slides to introduce the topic, provide an overview of the area, report the research findings, make comparisons with the second study, and conclude. Using the PowerPoint file, the researcher recorded unscripted lectures on the topics that lasted for approximately 5.30 minutes (Reproducibility was 5.37 and Diversity was 5.34). The lectures were unscripted to include features of connected speech and spoken grammar, which are difficult to replicate when following a script (O'Grady, 2023; Wagner, 2018).

Aryadoust and Luo (2022) emphasize the importance of describing test related attributes including features of the text or aural input as part of construct definition. It is also important to establish equivalence of language assessment tasks if they are to be used for the same purposes (Inoue, 2013). To this end, an analysis of the recordings was completed. The analysis demonstrated characteristics of connected speech including false starts, pauses and hesitations, contracted forms, deixis, ellipsis, assimilation, liaison, and anaphoric reference. To establish the speech speed, speech rate was calculated for each recording as the number of words divided by the number of seconds, and pruned speech rate as the same calculation with filled hesitations, false starts and repetitions removed from the transcript (De Jong, 2016). The pruned transcriptions were also used to calculate the proportion of words the lectures contained from the most common 1000-word (K1), 2000-word (K2), and academic word list (AWL) frequency bands using VocabProfile (Cobb, 2022b). The results are presented in Table 1, which indicates that speech rate was broadly equivalent and that the range of lexis in the two lectures was comparable.

The MCQs were created directly from the MP4. The questions were designed to measure comprehension of specific details, main ideas, purpose, lecturer attitudes and opinions, organizational structure, and illocutionary intent (O'Grady, 2023). The questions consisted of one question stem

TABLE 1. Analysis of the recordings.

	Reproducibility	Diversity
Speech rate in seconds	2.34	2.27
Speech rate in minutes	140.65	135.99
Pruned speech rate in seconds	2.25	2.08
Pruned speech rate in minutes	135.96	124.67
K1	80.00%	77.12%
K2	4.29%	1.69%
AWL	9.35%	12.43%
Off-list	6.36%	8.76%

and three options (Rodriguez, 2005). Twelve questions were developed for each lecture and an answer key was prepared. Accuracy of the answer key was confirmed by having an English teacher-trainer with over 20 years' experience at the local English for speakers of other languages (ESOL) college to complete the tasks. Several distracters were revised to decrease their attractiveness and after a second comparison, agreement with the proposed answer key was 100%. The questions were also categorized as those assessing implicit and explicit information by the teacher-trainer and agreement with the intended categorization was 100%. To enhance the potential for comparisons between the current study and the findings in the literature, the integrated response task was adapted from Rukthong (2021, p. 5) and involved the following instructions; You will have two minutes to summarize the lecture by discussing the main points and specific details.

Procedure

Participants completed the tasks on a PC with screen and audio capture and a webcam to record the participant and their interaction with the computer using Panopto (https://www.panopto.com/). The order of lectures and response format was counterbalanced between the participants (see Table 2). To complete the MCQ task, participants watched the recording twice. During the first run, test takers watched the lecture without interruption. During the second run, pauses were inserted into the recording at regular intervals not exceeding one minute, and questions about the section were presented on the screen. This design feature was developed to prevent preview of the MCQs and to restrict test-wise strategies that have been reported in the literature (O'Grady, 2023; Rukthong, 2021). To complete the integrated response task, participants were asked to watch the lecture twice and provided a two-minute summary of the lecture contents by speaking directly into the computer. The effect of double play has been a focus of research for some time with researchers reporting that it increases comprehension (Field, 2019; Holzknecht, 2019). In the current study, the decision to permit double play was taken to enhance domain representation; if students are following a lecture on a computer interface, they are free to view the recording as much as they please.

After each task, the participant completed a stimulated recall interview using the Panopto recording as stimulus. During the interviews, participants were presented with the audio-visual recording of themselves completing the task immediately after they had finished. During the stimulated recall interviews, the recording was paused by the researcher and the student was asked to recall their processing. Participants were encouraged to pause the recording independently if they recalled anything during viewing. The stimulated recall interviews were recorded as MP3 files.

Analysis

To determine the range of listening processes test takers engaged during the tasks required analysis of task responses to identify evidence of processing of explicit and implicit content and analysis of the contents of the stimulated recall. Responses to the MCQs were collated and compared with the answer key. This process was completed to determine whether students have provided evidence of successful comprehension of both explicit and implicit information. The integrated responses were transcribed and analyzed for speech rate and instances of overlap with the input text using Text Lex Compare (Cobb, 2022a). The software calculates the combined total

TABLE 2. Order of lecture and response mode.

Participant	Input and Response	
1, 5	Lecture 1 MCQ	Lecture 2 Spoken
2	Lecture 1 Spoken	Lecture 2 MCQ
3	Lecture 2 MCQ	Lecture 1 Spoken
4	Lecture 2 Spoken	Lecture 1 MCQ

number of tokens, word families and three-word strings in the lectures and the integrated responses and calculates the proportion that is common to both texts. Overlap at the token and word family level is expected as the response is a summary task and participants are likely to use similar words to compete the task, however the presence of common three-word strings may indicate efforts to memorize and replicate language use from the lecture, which does not necessarily entail comprehension (Cumming et al., 2005). In addition, the transcripts were analyzed for evidence of comprehension of both implicit and explicit information by both the researcher and teacher-trainer. This involved counting the number of ideas expressed in each response and categorizing the idea as explicit or implicit (for a similar approach to content analysis see Knoch et al., 2014). Explicit information was categorized as discussion of specific details, whereas implicit was classified as discussion of implications and inferences, lecturer attitudes, and drawing conclusions. Categorization was verified by an external coder with 100% agreement.

The stimulated recall interviews were transcribed and coded according to Aryadoust and Luo's (2022, p. 19) categories of task and listener related attributes, processes and subskills:

Attributes	
Listener Related Attributes	Task Related Attributes
Task Knowledge	Features of the input
Proficiency	Visual contents
Affective	Task format
Processes	
Bottom-up processing	
Top-down processing	
Memory	
Cognitive and metacognitive strategies	
Subskills	
Knowledge of the sound system	
Understanding local linguistic meanings (vocabulary and syntax)	
Understanding global meanings or inferred meanings	
Communicative listening ability	

An external coder coded 32% of the total transcripts (2808 words of 8867) and agreement with the first coder was calculated as the total number of opportunities for agreement divided by instances of agreement. The agreement figure was 78%. After calculating the agreement statistic, the two coders discussed discrepancies; a disagreement about the differences between local linguistic meanings and bottom-up processing, and understanding global meanings and top-down processing could not be resolved. In Aryadoust and Luo's (2022) figure, overlap between these categories is clear, e.g. making inferences appears as both a top-down process and as a subskill involving understanding global and inferred meanings. It was therefore agreed to combine the categories (local & bottom up/global & top-down). However, the subcategories "*Knowledge of the sound system*" and "*Communicative listening ability*" were retained because there were comments that were specifically identified as belonging to these categories. Once these revisions had been made, the remaining content was coded according to the revised coding plan. After coding was completed, the frequency of categories was tallied per task and the results were compared.

TABLE 3. MCQ responses.

	Lecture 1 (Diversity)				Lecture 2 (Reproducibility)		
Item	Type	Participant 1	Participant 4	Participant 5	Type	Participant 2	Participant 3
1	Implicit	1	0	1	Explicit	1	1
2	Explicit	1	1	1	Explicit	0	1
3	Explicit	0	1	0	Explicit	1	1
4	Implicit	0	1	1	Implicit	0	1
5	Implicit	1	1	1	Explicit	0	1
6	Explicit	1	1	1	Explicit	0	1
7	Implicit	0	1	1	Explicit	1	1
8	Explicit	0	1	1	Explicit	0	0
9	Explicit	0	1	0	Explicit	1	1
10	Explicit	1	1	0	Explicit	1	1
11	Implicit	1	1	1	Implicit	1	1
12	Implicit	1	1	1	Implicit	0	1
Total		7	11	9		6	11

TABLE 4. Speech rate and pruned speech rate.

Participant	MCQ Score	Total Number of words	Speech Rate (words per minute)	Pruned speech rate (words per minute without repetitions)
1	7	143	71.5	61
2	6	272	136	131
3	11	133	66.5	60.5
4	11	200	100	91
5	9	186	93	87.5

Results

The MCQ responses were scored and are presented in Table 3. Lecture 1 yielded a stronger balance of implicit and explicit questions than lecture 2, which contains three implicit information questions. Nonetheless, the responses in each task provide some evidence of successful and unsuccessful processing of implicit and explicit information.

To investigate the amount of speech produced and the relative levels of fluency involved in the integrated responses, the total number of words produced was calculated and this was used to establish speech rate, expressed as words per minute, and pruned speech rate with repetitions and false starts removed, expressed as words per minute. The results are presented in Table 4. The total number of words produced ranged substantially between participants. Participant two produced a particularly high number of 272 words in two minutes, whereas participant three produced less than half of this number. However, a contradictory pattern was observed in the MCQ task, with participant three providing almost twice as many correct responses as participant two.

To investigate the levels of overlap between the input and the students' responses, the transcripts were analyzed using Text Lex Compare and results are presented in Table 5. The table indicates the proportion of overlap at the token, word family and three-word string level as well as common examples. High levels of overlap were observed at the token level and word family level. Instances of overlap in three-word strings were less commonly observed although participant five included 25 instances of three-word strings that overlapped with the source text.

The participants' spoken responses were searched for evidence of processing of explicit and implicit content. The average number of ideas produced in each response was 8.8 (SD 2.28). Evidence of processing of explicit information was immediately clear in all responses as participants were able to discuss lecture content with accuracy. However, close reading of the responses demonstrated that the participants had merely reported on the contents of the lectures without discussing implicit information. The second coder was likewise unable to identify any evidence of implicit processing.

TABLE 5. Overlap between source texts and integrated responses.

Participant	Lecture	Tokens	Word family	Three-word strings	Examples of three-word strings*
1	2	54.84%	18.50%	.89%	001. this kind of 4 002. are two ways 2 003. there are two 2
2	1	68.14%	25.38%	1.85%	001. and diversity and 2 002. by freeman and 2 003. collaboration and diversity 2 004. freeman and huang 2 005. location where they 2 006. more citations and 2 007. quality of the 2 008. the quality of 2 009. they found that 2
3	1	53.18%	17.69%	1.80%	001. benefits of collaboration 3 002. the benefits of 3 003. and so they 2 004. collaboration and diversity 2 005. people tend to 2 006. tend to work 2 007. of collaboration and 1
4	2	63.98%	23.96%	1.36%	001. the academic fields 4 002. happened in the 2 003. readers of the 2 004. to see how 2 005. the readers of 1
5	2	62.08%	22.51%	6.90%	001. there are two 4 002. this kind of 4 003. are two ways 3 004. the data and 3 005. about reproducibility and 2 006. and replicability in 2 007. basis of the 2 008. findings for example 2 009. is to request 2 010. on the basis 2 011. original data and 2 012. reach the same 2 013. replicability in academic 2 014. reproducibility and replicability 2 015. request the original 2 016. the basis of 2 017. the original data 2 018. the same findings 2 019. to readers of 2 020. to request the 2 021. and there are 1 022. data and the 1 023. findings this kind 1 024. same findings this 1

*the number indicates the number of times the overlap occurs.

Stimulated recall

The results of the coding of the stimulated recall interviews are presented in Table 6. Based on these figures, there are clear distinctions between the response types in terms of the number of comments about task knowledge, knowledge of the sound system and communicative listening ability. However, more generally the stimulated recall contents were equivalently distributed between the two response formats. The following section presents examples of the comments made about the MCQ and integrated task to facilitate comparisons (categories are represented by numbers, subcategories by letters, and instances by Roman numerals).

TABLE 6. Coding of the stimulated recall transcripts.

Category	Subcategory	MCQ Response	Integrated Response	Total
1 Task related attributes	(a) *Task format*	8	9	17
	(b) *Features of the input*	0	0	0
	(c) *Visual contents*	7	9	16
2 Listener related attributes	(a) *Affective*	6	5	11
	(b) *Proficiency*	0	0	0
	(c) *Task Knowledge*	0	2	2
3 Processes and Subskills	(a) *Bottom-up processing & understanding local linguistic meanings (vocabulary and syntax)*	14	14	28
	(b) *Knowledge of the sound system*	0	3	3
	(c) *Top-down processing & understanding global meanings or inferred meanings*	10	10	20
	(d) *Communicative listening ability*	7	1	8
	(e) *Memory*	10	11	21
	(f) *Cognitive and metacognitive strategies*	12	16	28

Task Related Attributes. The first category is task related attributes relating to task format, features of the input and visual contents. Comments about task format described the process of answering MCQs (*1.a.i*), with one participant noting that MCQs provide a monitor of comprehension and success during the task (*1.a.ii*). Comments about the integrated response discussed planning the summary and the difficulty of the task (*1.a.iii*). All participants commented on the effect of the visual contents, describing the divided attention effect of simultaneously processing reading and listening (*1.c.i*), and less frequently the effect of seeing the lecturer (*1.c.ii*). In the integrated response, it was common for participants to discuss the role of the slides in the summary (*1.c.iii* & *1.c.iv*).

1.a.i. I struggled with these two I think I finally settled with the marriage and ethnicity are unrelated because I think first and third option it's not exactly the right thing so I choose the second one (Participant 1: MCQ)

1.a.ii. At first I thought the question was very easy so like direct and quick answers and then I look at the question I thought ok this question designed like a little bit tricky and then I think I need to pay more attention (Participant 3: MCQ)

1.a.i. I was recalling how I structure my summarisation . . . I know I understand but I know it would be difficult to summarise (Participant 5: integrated)

1.c.i. I cannot separate reading from listening. I have this two thing in mind working together reading and listening (Participant 5: integrated)

1.c.ii. I can see your face I can see your body language that actually helped for my understanding (Participant 1: MCQ)

1.c.iii. I'm just trying to speak whatever I remember specific detail points the first one more from the slide because it's quite brief and short so I mainly rely on my screen shot in my head of the slides (Participant 2: integrated)

1.c.iv. I think I was more concentrated on the text reading because I was thinking about summarising so as a preparation for that I was trying to memorise as much information as possible . . . About the first slide I thought I was ok and in the middle I realised I skipped the second slide so I went back to the second slide and then move on to the third slide (Participant 3: integrated)

Listener Related Attributes. The second category is listener related attributes involving affective, proficiency, and task knowledge. All students commented on their affective state during the tasks. Examples in the MCQ task included confusion and surprise at the MCQ options. In the integrated task, comments coded as affective most frequently express anxiety at the prospect of speaking (*2.a.i*), which was something all participants described. Participant 1 also described task knowledge as experiences of completing integrated tasks (*2.c.i*).

2.a.i. I'm quite nervous there because when you asked me to summarise the whole lecture I feel like I would panic (Participant 1: integrated)

2.c.i. actually I did this kind of test before in my undergrad by I always feel like it is difficult if you forgot what is the beginning of the lecture video (Participant 1: integrated)

Processes and Subskills. The third category is processes and subskills relating to bottom-up processing & understanding local linguistic meanings (vocabulary and syntax), knowledge of the sound system, top-down processing & understanding global meanings or inferred meanings, communicative listening ability, memory, cognitive and metacognitive strategies. There was an equal distribution of codes between response types within the bottom-up processing & understanding local linguistic meanings category. An interesting trend was for participants to purposefully engage bottom-up processing during the second listening to understand specific details (*3.a.i*). There was also a common focus on "key words" to use in the summary, which may account for the overlap in lexis observed between the lecture and the responses (see Table 5; *3.a.ii*). Participants also expressed a perceived need to pay attention to detail in the lecture to later summarize (*3.a.iii*). One participant reported simultaneously engaging bottom-up processing of lecturer's speech and the lecture slides during the second listen as an attempt to resolve uncertainty experienced during the first listen (*3.a.iv*). Within the knowledge of the sound system subcategory, participants described a perceived need to focus on the pronunciation of novel vocabulary to include in the summary (*3.b.i*).

3.a.i. before that question I always get the main idea of everything you are saying and then after that question I tried to focus on every exact word that you were talking about (Participant 1: MCQ)

3.a.ii. I tried to remember what the lecturer was saying I tried to grab the key words he was saying (Participant 4: integrated)

3.a.iii. (pay attention to) 2.5 million surnames (in general not specific authors) and then U S and then surname to integrate ethnicity because I summarise it accurately (Participant 2: integrated)

3.a.iv. In the beginning I was wondering what does it mean? homophily the word but you already explained it helps when you see the word while listening I was reading here to see whether what you are saying can be reflected here (Participant 5: MCQ)

3.b.i. I was thinking do I need to remember that how to pronounce it in my summarising . . . I'm trying to remember how you pronounce it I'm going to use it later (Participant 2: integrated)

Comments relating to top-down processing and understanding global meanings or inferred meanings described understanding main ideas (*3.c.i*), and synthesizing information and understanding relationships (*3.c.ii*). The communicative language ability subcategory involved comments about the MCQ task and the need to understand the speaker's purpose and opinions to respond to a question (*3.d.i* & *3.d.ii*). Crucially, the integrated response comments did not refer to communicative language ability except for one comment that signaled an order effect in the presentation of the tasks (*3.d.iii*).

3.c.i. I'm trying to get the general idea of it so I actually I would draw a mind map (Participant 1: integrated)

3.c.ii. You didn't mention them together you mentioned Chinese and Korean together and English later much later . . . I realised that the first two questions more specific and then it suddenly changed and there's more about higher order thinking skills you connect things together (Participant 5: MCQ)

3.d.i At first the lecturer clearly explained the research content and the benefits and the limitations so I think it is very objective . . . I choose the purpose of this lecture is to demonstrate that's its beneficial to promote diversity (Participant 4: MCQ)

3.d.ii. did you show your personal opinion did you this is really convincing like this kind of thing did you mention this kind of thing or are you just stating the facts? (Participant 5: MCQ)

3.d.iii. The first time you see there are two findings and the second time you focus on what these two findings are really talking about also from the last video whether the lecturer had a view on it or not (Participant 5: integrated)

Participants discussed the need for memory to complete the tasks with an equal distribution between the response formats. Whereas in the MCQ task, participants described comparing memory

with MCQ options, in the integrated task, comments often concerned the need to commit details to memory to recall during the summary, and the role of the slides in this process (*3.e.i*).

3.e.i. I was not sure if I could memorise what I heard . . . I found it interesting the content of this lecture but still I was thinking about if I managed to memorise everything . . . I think I was more concentrated on the text reading because I was thinking about summarising so as a preparation for that I was trying to memorise as much information as possible (Participant 3: integrated)

The participants discussed the cognitive and metacognitive strategies they had applied to complete the tasks. Participants reported evaluating understanding, prioritizing information during the second listen, planning the summary while listening to the text (*3.f.i*), rehearsing the summary while listening (*3.f.ii*), and efforts to exploit the MCQ response format using test-wise strategies (*3.f.iii*).

3.f.i. I was thinking about summarising again while listening I thought like which word to use (Participant 3: Integrated)

3.f.ii The second time listening I tried to reproduce my mouth I tried to reproduce the words because I know the final goal is to speak the words like reproduce it (Participant 4: Integrated)

3.f.iii. I thought none of them applies to what I heard but I thought I have to choose one so which one could be the right answer just eliminate the option which is less likely to be correct (Participant 2: MCQ)

Discussion

The study set out to determine the impact of response format on computer-mediated lecture comprehension tasks. This is an increasingly important focus as computer-mediated assessment becomes more common and language testers need to be aware of the impact of response type on constructs in this environment (Khabbazbashi et al., 2022). The primary finding was that the integrated response and MCQ formats overlapped and diverged in ways that impacted on the measurement of the assessment construct. Participant comments indicate that the integrated format did not comprehensively engage communicative listening ability in the way that MCQs explicitly requiring comprehension of implicit information did. This finding contrasts with conclusions reached in research, which tend to associate selected response tasks with construct underrepresentation and construct irrelevant variance (Rukthong, 2021). This effect may be attributed to assessment task design and underscores the necessity of including both explicit and implicit items and restricting item preview in lecture comprehension tasks (O'Grady, 2023). Integrated response instructions may also need to explicitly state that tests takers focus on attitudes and speaker purpose in their responses otherwise test developers run the risk of eliciting simple summary responses that demonstrate comprehension of explicitly stated information only. Communicative listening ability as defined by Aryadoust and Luo (2022) is associated with increased language proficiency, and in the present study only the MCQ format created the necessary conditions for test takers to display this competence.

An important difference between the formats was the prioritization of pronunciation of new vocabulary in the integrated response. Participants were concerned with learning the pronunciation of new words to include in their summaries. This finding might be interpreted in two ways: primarily, an argument could be made that prioritizing the pronunciation of new words represents a test method effect that is not associated with listening comprehension and hence a source of construct irrelevant variance. Alternatively, there is a possibility that the integrated format may be engaging the ability to utilize new vocabulary; ability to learn is a skill that is often targeted and valued in academic language tests (Cohen & Upton, 2006). Further research may be required to investigate the potential to measure ability to learn with integrated response tasks.

Response format was shown to interact with several core listener attributes. Primarily, the integrated format was associated with increased test anxiety (Kormos et al., 2020; Lee & Winke, 2018). Anxiety is a significant source of construct irrelevant variance that affects task performance; "negative self-preoccupations and self-ruminative thoughts act as an extra load on the cognitive system, leading to cognitive deficits during cognitive processes necessary for performance: Attention, memory, and retrieval are reduced" (L. Cheng & Zheng, 2020, p. 180). References to

attention, memory and retreival were present in the data: to establish the accuracy of MCQ options, and in the integrated task to store and retrieve main ideas and specific details to accurately summarize the lecture. This was evidenced by the various lexical and phrasal borrowings from the source texts in the integrated responses. However, borrowing was typically limited to single token or word families rather than verbatim copying from the sources. While the analysis did indicate instances of overlap at the three-word level, these were typically common expressions such as "this kind of," or "they found that." Verbatim copying may thus be less of a concern in integrated listening and speaking assessment than integrated writing assessment; in the speaking task integrating language creates cohesion with the source rather than inflating grades (Plakans, 2015).

Participants reported that lecture comprehension divided attention between listening and reading, as well as processing the lecturer's body language, facial expressions, and gestures. At times, reading was often prioritized in instances when listening had been unsuccessful. This would suggest that the lecture comprehension construct in this task involved cognitive processes involved in listening (Field, 2019) and reading (Khalifa & Weir, 2009), which may have resolved ambiguities of miscomprehension. Crucially, it is the slides that the participants most frequently discussed when recalling memories of the speaking component of the integrated task. Memory, retrieval, and reading comprehension are thus clear components of the lecture comprehension construct, with increased demands on resources associated with the integrated reponse format. Scores on multimedia lecture comprehension tasks should thus be interpreted as representative of both reading and listening processes. These findings contribute to the strong theoretical basis for the inclusion of audio-visual content in tests of lecture comprehension because listeners rarely engage audio only processes without simultaneously processing visual stimuli and seem to rely on text to complete integrated tasks (Suvorov & He, 2022; Wagner, 2021). It stands to reason that seeking to develop broader interpretations about students' readiness to begin education in a second language requires the use of multimodal input and computer-based assessment is particularly well-suited to this purpose (Aryadoust, 2022).

Conclusion

The purpose of the present study was to investigate the extent to which response type impacted on task performance in a computer-mediated lecture comprehension test. Participants completed two tasks developed specifically for the study involving implicit and explicit MCQs without question preview and an integrated listening and speaking task and completed stimulated recall interviews. The findings demonstrate substantial overlap between the response formats with several notable exceptions related to knowledge of the sound system, communicative listening ability, memory and test anxiety. Though the study provides a detailed account of variation in response format in lecture comprehension tasks, there are clear limitations. Primarily, the small sample size permitted depth of analysis but also has implications for the generalizability of the findings. For this reason, research involving these response formats with larger and more diverse groups of individuals, for example in terms of language proficiency and L1 background, would be informative. Future research may therefore explore the impact of response format in tests with similar tasks using larger samples, focussing on the scoring related aspects of the assessment such as discriminatory power (Yeager & Meyer, 2022). Assessment research may also adopt a longitudinal approach to determine the most appropriate format to measure achievement. From the perspective of classroom teachers with responsibility for applying test data to inform future instruction, it is important to establish which of these response formats provides the best information about the students' ability to follow lectures. This study indicates that students may have a predisposition to report the contents of the lecture in the integrated tasks rather than discussing information obtained through higher order processing. Identifying potential alterations to the integrated format to obtain evidence of communicative listening ability may thus prove a fruitful avenue for future research.

Disclosure statement

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

ORCID

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

References

Aryadoust, V. (2022). The known and unknown about the nature and assessment of L2 listening. *International Journal of Listening*, *36*(2), 69–79. https://doi.org/10.1080/10904018.2022.2042951

Aryadoust, V., & Luo, L. (2022). The typology of second language listening constructs: A systematic review. *Language Testing*, *40*(2), 375–409. https://doi.org/10.1177/02655322221126604

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

Baker, M. (2016). Scientists lift the lid on reproducibility. *Nature*, *1*(7604), 500. https://doi.org/10.1038/533452a

Batty, A. O. (2021). An eye-tracking study of attention to visual cues in L2 listening tests. *Language Testing*, *38*(4), 511–535. https://doi.org/10.1177/0265532220951504

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

Chalhoub-Deville, M., & O'Sullivan, B. (2020). *Validity theoretical development and integrated arguments*. Equinox.

Chang, A. C. S., & 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*(3), 575–586. https://doi.org/10.1016/j.system.2013.06.001

Cheng, H. (2008). A comparison of multiple-choice and open-ended response formats for the assessment of listening proficiency in English. *Foreign Language Annals*, *37*(4), 544–553. https://doi.org/10.1111/j.1944-9720.2004.tb02421.x

Cheng, L., & Zheng, Y. (2020). Measuring anxiety. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 177–187). Routledge.

Choi, Y. D. (2022). Validity of score interpretations on an online English placement writing test. *Language Testing in Asia*, *12*(1), 12. https://doi.org/10.1186/s40468-022-00187-0

Cobb, T. (2022a) Text Lex Compare v.5 [computer program]. https://www.lextutor.ca/cgi-bin/tl_compare/

Cobb, T. (2022b) Web vocabprofile [computer program]. http://www.lextutor.ca/vp/

Cohen, A. D. (2011). *Strategies in learning and using a second language*. Routledge.

Cohen, A. D., & Upton, T. A. (2006). Strategies in responding to the new TOEFL reading tasks. *ETS Research Report Series*, *2006*(1), i–162. https://doi.org/10.1002/j.2333-8504.2006.tb02012.x

Coniam, D. (2006). Evaluating computer-based and paper-based versions of an English-language listening test. *ReCALL*, *18*(2), 193–211. https://doi.org/10.1017/S0958344006000425

Crossley, S. A., Clevinger, A., & Kim, Y. (2014). The role of lexical properties and cohesive devices in text integration and their effect on human ratings of speaking proficiency. *Language Assessment Quarterly*, *11*(3), 250–270. https://doi.org/10.1080/15434303.2014.926905

Crossley, S. A., & Kim, Y. (2019). Text integration and speaking proficiency: Linguistic, individual differences, and strategy use considerations. *Language Assessment Quarterly*, *16*(2), 217–235. https://doi.org/10.1080/15434303.2019.1628239

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & Jamse, M. (2005). Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL®. *ETS Research Report Series*, *2005*(1), i–77. https://doi.org/10.1002/j.2333-8504.2005.tb01990.x

Field, J. (2012). The cognitive validity of the lecture-based question in the IELTS listening paper. In L. Taylor & C. Weir (Eds.), *IELTS collected paper 2: Research in reading and listening assessment* (pp. 391–453). Cambridge University Press.

Field, J. (2019). *Rethinking the second language listening test from theory to practice*. Equinox.

Freeman, R. B., & Huang, W. (2014). Collaboration: Strength in diversity. *Nature*, *513*(7518), 305–305. https://doi.org/10.1038/513305a

Frost, K., Wigglesworth, G., & Clothier, J. (2021). Relationships between comprehension, strategic behaviours and content-related aspects of test performances in integrated speaking tasks. *Language Assessment Quarterly*, *18*(2), 133–153. https://doi.org/10.1080/15434303.2020.1835918

He, L., & Jiang, Z. (2020). Assessing second language listening over the past twenty years: A review within the socio-cognitive framework. *Frontiers in Psychology*, 11. https://doi.org/10.3389/fpsyg.2020.02123

Holzknecht, F. (2019). *Double Play in Listening Assessment* [Doctoral Thesis]. Lancaster University. https://doi.org/10.17635/lancaster/thesis/812

Holzknecht, F., McCray, G., Eberharter, K., Kremmel, B., Zehentner, M., Spiby, R., & Dunlea, J. (2020). The effect of response order on candidate viewing behaviour and item difficulty in a multiple-choice listening test. *Language Testing*, *38*(1), 41–61. https://doi.org/10.1177/0265532220917316

Huang, H. T. D., & Hung, S. T. A. (2013). Comparing the effects of test anxiety on independent and integrated speaking test performance. *TESOL Quarterly*, *47*(2), 244–269. https://doi.org/10.1002/tesq.69

Huang, H. T. D., Hung, S. T. A., & Hong, H. T. V. (2016). Test-taker characteristics and integrated speaking test performance: A path-analytic study. *Language Assessment Quarterly*, *13*(4), 283–301. https://doi.org/10.1080/15434303.2016.1236111

Inoue, C. (2013). *Task equivalence in speaking tasks*. Peter Lang.

Inoue, C., & Lam, D. M. (2021). The effects of extended planning time on candidates' performance, processes, and strategy use in the lecture listening-into-speaking tasks of the TOEFL iBT® test. *ETS Research Report Series*, *2021*(1), 1–32. https://doi.org/10.1002/ets2.12322

Jong, N. (2016). Fluency in second language assessment. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 203–218). Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9781614513827-015

Khabbazbashi, N., Chan, S., & Clark, T. (2022). Towards the new construct of academic English in the digital age. *ELT Journal*, *77*(2), 207–216. https://doi.org/10.1093/elt/ccac010

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge University Press.

Knoch, U., Macqueen, S., & O'Hagan, S. (2014). An investigation of the effect of task type on the discourse produced by students at various score levels in the TOEFL iBT® writing test. *ETS Research Report Series*, *2014*(2), 1–74. https://doi.org/10.1002/ets2.12038

Kormos, J., Brunfaut, T., & Michel, M. (2020). Motivational factors in computer-administered integrated skills tasks: A study of young learners. *Language Assessment Quarterly*, *17*(1), 43–59. https://doi.org/10.1080/15434303.2019.1664551

Koyama, D., Sun, A., & Ockey, G. (2016). The effects of item preview on video-based multiple- choice listening assessments. *Language Learning and Technology*, *20*(1), 148–165. http://dx.doi.org/10125/44450

Lee, S., & Winke, P. (2018). Young learners' response processes when taking computerized tasks for speaking assessment. *Language Testing*, *35*(2), 239–269. https://doi.org/10.1177/0265532217704009

Li, C., Chen, C., Wu, M., Kuo, Y.-C., Tseng, Y.-T., Tsai, S.-Y., & Shih, H.-C. (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. https://doi.org/10.1080/10904018.2015.1058165

O'Grady, S. (2023). Adapting multiple-choice comprehension question formats in a test of second language listening comprehension. *Language Teaching Research*, *27*(6), 1431–1455. https://doi.org/10.1177/1362168820985367

Plakans, L. (2015). Integrated second language writing assessment: Why? what? how? *Language and Linguistics Compass*, *9*(4), 159–167. https://doi.org/10.1111/lnc3.12124

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement Issues & Practice*, *24*(2), 3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Rukthong, A. (2021). MC Listening questions vs. integrated listening-to-summarize tasks: What listening abilities do they assess? *System*, *97*. https://doi.org/10.1016/j.system.2020.102439

Rukthong, A., & Brunfaut, T. (2020). Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks. *Language Testing*, *37*(1), 31–53. https://doi.org/10.1177/0265532219871470

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, *23*(4), 441–474. https://doi.org/10.1191/0265532206lt337oa

Suvorov, R., & He, S. (2022). Visuals in the assessment and testing of second language listening: A methodological synthesis. *International Journal of Listening*, *36*(2), 80–99. https://doi.org/10.1080/10904018.2021.1941028

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

Wagner, E. (2021). Assessing Listening. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed., pp. 223–235). Routledge.

Wagner, E., Liao, Y., & Wagner, S. (2020). Authenticated spoken texts for L2 listening tests. *Language Assessment Quarterly*, *18*(3), 205–227. https://doi.org/10.1080/15434303.2020.1860057

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing*, *9*(1), 27–55. https://doi.org/10.1016/j.asw.2004.01.002

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

Wei, W., & Zheng, Y. (2017). An investigation of integrative and independent listening test tasks in a computerised academic English test. *Computer Assisted Language Learning*, *30*(8), 864–883. https://doi.org/10.1080/09588221.2017.1373131

Westbrook, C. (2023). The impact of input format on written performance in a listening-into-writing assessment. *Journal of English for Academic Purposes*, *61*. https://doi.org/10.1016/j.jeap.2022.101190

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*(1), 107–122. https://doi.org/10.1016/j.system.2007.12.003

Yeager, R., & Meyer, Z. (2022). Question preview in English for academic purposes listening assessment: The effect of stem preview on difficulty, item type, and discrimination. *International Journal of Listening*, *36*(3), 299–324. https://doi.org/10.1080/10904018.2022.2029705

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. https://doi.org/10.15858/engtea.71.4.201612.27

Zechner, K., & Evanini, K. (2020). *Automated speaking assessment using language technologies to score spontaneous speech*. Routledge.