Castledown

OPEN ACCESS

Language Education & Assessment
ISSN 2209-3591
https://www.castledown.com/journals/lea/

Language Education & Assessment, 3(1), 13–35 (2020)
https://doi.org/10.29140/lea.v3n1.193

Assessing L2 Listening at a Japanese University: Effects of Input Type and Response Format

KERRY PUSEY [a]

[a] *Nagasaki University, Japan*
kerryjpusey@gmail.com

Abstract

The use of video lectures and authentic listening tasks (e.g., taking notes; responding to short answer questions) is common practice in EAP classrooms. However, many classroom-based tests of L2 listening comprehension continue to employ audio-only listening texts and a multiple-choice response format. The effect of these differences in input type and response format on test-taker performance remains elusive and begs the question as to which is the best option in terms of construct validity. Furthermore, the interaction between these test task characteristics and their potential joint effect on performance has not been sufficiently explored. To address this gap, a study was conducted at a Japanese university which investigated the effect of input type (audio-only vs. video) and response format (multiple-choice vs. short answer) on L2 listening test performance. Participants were divided into four groups to take an academic listening test with one of four combinations of input and response format: (1) audio-only with multiple-choice questions; (2) video with multiple-choice questions; (3) audio-only with short answer questions; and (4) video with short answer questions. Results of a 2 x 2 factorial ANOVA revealed a statistically significant effect of response format on test-taker performance. No significant effect for input type was found and no significant interaction among the variables was detected. Results suggest that visual input and audiovisual literacy need to be more clearly articulated within the construct definition of academic listening, if they are to be included at all.

Keywords: English for academic purposes, L2 listening assessment, test task characteristics, response format, visual input

Introduction

Academic listening is a complex process that involves attending to linguistic and non-verbal aspects of communication (e.g., gestures, facial cues) in real time (Winke & Isbell, 2018), typically while engaging in other tasks (e.g., reviewing information in a handout) and competing with the presence of

Data Availability Statement: *All relevant data are within this paper.*

various conflicting signals and distractions (e.g., classmates chatting about irrelevant topics). To simulate the experience of real-world academic listening, English for academic purposes (EAP) instructors often employ video lectures (Field, 2011; Flowerdew & Miller, 2005) and authentic listening tasks (e.g., taking notes; responding to short answer questions) in the classroom, which correspond to many of the characteristics of the academic target language use (TLU) domain (Bachman & Palmer, 2010). In the absence of access to live lectures in the target language, teachers working in English as a foreign language (EFL) contexts in particular, such as Japan, depend on such materials to help prepare students for studying in an English-medium university context. However, in testing situations, despite the affordances of modern technology, there tends to be an overwhelming reliance on the use of audio-only input and a multiple-choice response format (Cubilo & Winke, 2013), which raises questions of ecological and cognitive validity (Field, 2011; 2013), test task authenticity (Bachman, 1990; Gan, 2012; Li, 2013), and construct representation (Miller, Linn, & Gronlund, 2009).

In response, some researchers have examined the effect of input type (i.e., video vs. audio-only) on second language (L2) listening test-taker performance (e.g., Cubilo & Winke, 2013; Suvorov, 2009; Wagner, 2010; 2013), while others have looked at the effect of response format (i.e., selected vs. constructed response, e.g., Brindley & Slatyer, 2002; In'ami & Koizumi, 2009). In both cases, however, results have been mixed (though the general picture is somewhat clearer for response format; see below). Furthermore, few studies have investigated the potential interaction between input type and response format in tests of L2 listening comprehension. More research is therefore needed in order to better understand the role these variables play in L2 listening tests.

To this end, the current study investigates the effect of input type (video vs. audio-only) and response format (multiple-choice vs. short answer) on the listening comprehension test scores of a group of Japanese students enrolled in an EAP program at a public university.

Literature Review

Although there seems to be a general consensus on the cognitive processes involved in L2 listening comprehension (see models described by Field, 2008; Flowerdew & Miller, 2005; Rost, 2011; Vandergrift & Goh, 2012), researchers agree less about how the construct of listening should be defined within the context of language testing (Cubilo & Winke, 2013, p. 372). According to Buck (2001, p. 112), there are no universally agreed upon "rules" for how to define the construct of listening; rather, the test purpose and the TLU situation should inform how the construct is defined and operationalized in specific language tests. In the absence of a need to specify an alternative definition (e.g., a language for specific purposes test), Buck's (2001) "default listening construct" (pp. 112-115) is a useful benchmark for defining the L2 listening construct and is one that has informed previous studies of L2 listening assessment (e.g., Wagner, 2010). Buck (2001) summarizes his default listening construct as the ability to:

- process extended samples of realistic spoken language, automatically and in real time,
- understand the linguistic information that is unequivocally included in the text, and
- make whatever inferences are unambiguously implicated by the content of the passage. (p. 114)

While Buck's often cited work is a useful starting point for conceptualizing the construct of L2 listening ability, there are alternative views—in particular as related to the role of visual input in L2 listening comprehension.

Visual Input and Listening Comprehension

Visuals, such as video, photographs, and graphic representations of textual information (e.g., charts, graphs, infographics), are commonly used in language classrooms and are generally believed to aid in the teaching and learning of L2 listening (Flowerdew & Miller, 2005; Winke, Gass, & Sydorenko, 2013). Videotexts in particular are considered useful, due to their inclusion of non-verbal semiotic information (e.g., style and appearance of speakers) and kinesic aspects of communication, such as gestures, facial expressions, lip movements, gaze, body positionings, and movement (Gregersen, 2007; Hall, 2019; Kellerman, 1992; Taylor, 2014). Videotexts are also noteworthy for their ability to contextualize listening input (e.g., by including information about the location of communicative events), which helps activate listeners' schemata (Flowerdew & Miller, 2005).

For these reasons, among others (e.g., increasing students' engagement and motivation in listening activities; Brinton, 2014; Parry & Meredith, 1984; Progosh, 1996), many scholars of L2 listening (Feak & Salehzadeh, 2001; Field, 2011; Flowerdew & Miller, 2005; Lynch, 2011) have advocated for the use of videotexts not only in the classroom but also in language testing situations. Lynch (2011) clearly indicates this position, arguing that "it is becoming increasingly difficult to justify academic listening assessment (and research) based on audio-only input, of the type that has been the norm" (p. 86). However, the use of videotexts in tests of L2 listening comprehension has remained controversial, and research that has investigated the effect of visual input on test-taker performance has yielded mixed results.

Studies Investigating the Effect of Visual Input on L2 Listening Comprehension

Some studies have shown positive effects for visual input on test-taker performance. Sueyoshi and Hardison (2005) investigated the effect of visual input (specifically facial cues and gestures) on listening comprehension as measured by a 20-question multiple-choice listening comprehension task. Participants in the study completed the listening task under one of three conditions: audiovisual with gestures and facial cues, audiovisual with facial cues (no gestures), or audio-only. Results indicated significant effects for both visual input groups (regardless of proficiency level) as compared to the audio-only group. No significant difference in comprehension scores were found between the two visual input groups.

One researcher whose findings have consistently shown a positive effect of visual input on listening test performance is Elvis Wagner. In Wagner's (2010) study, a quasi-experimental design was employed to compare the effects of visual input (video vs. audio-only) on listening comprehension test scores of an experimental (video; n = 103) and control (audio-only; n = 99) group. The posttest instrument used in the study included 40 items (18 multiple-choice and 22 short answer questions), and results revealed statistically significantly higher mean scores for the video group as compared with the audio-only group.

Another study by Wagner (2013) looked at the effect of visual input in conjunction with access to test questions (i.e., allowing test-takers to view the questions while the listening text was played) on test performance. Once again, the video group scored significantly higher than the audio-only group (though the effect size was quite small [partial η^2 = .04]). No significant main effect for access to test questions was found.

In addition to Wagner (2010; 2013), other researchers have also found a positive effect for visual input on listening comprehension (e.g., Parry & Meredith, 1984; Shin, 1998). However, several studies have shown no effect for visual input on test-taker performance, and others have even shown a negative

effect. A study by Cubilo and Winke (2013) examined the effect of visual input (video vs. audio-only with a still picture) on test-taker performance when performing an integrated writing task, as well as input effects on note-taking behavior. The authors found no effect of input type on overall essay scores (though there was a statistically significant difference in scores for the *language use* criterion of the scoring rubric; those who watched a videotext scored higher than the audio-only with still picture group). However, the presence of video texts did have an effect on note-taking behavior: Significantly fewer notes were taken when participants watched a video lecture as compared to those who listened to a lecture with a still picture. Studies by Brett (1997), Coniam (2001), and Londe (2009) have likewise shown no significant effect of visual input (compared with audio-only input) on listening test scores.

Suvorov's (2009) study looked at the effects of input (audio-only, photographs, or video) and text type (dialogues and lectures) on test-taker performance and found that participants scored significantly lower in the video condition. Similarly, a study by Pusey and Lenz (2014) compared two groups' (one with videotexts and one with audio-only texts) performances on a test of listening comprehension and found a significant negative effect of visual input on listening comprehension test scores with a low to moderate effect size ($\eta^2 = .25$).

Overall, the extent to which visual input impacts test-taker performance—whether positively or negatively, and under what conditions—is unclear. One possibility is that in order for test-takers to make productive use of visual input, the test tasks they perform (i.e., the means through which they demonstrate their comprehension) and the characteristics of these tasks must be specifically designed to allow for the test-taker to attend to the visual dimension of the listening input they receive. In part, this is a question of response format characteristics.

Effects of Response Format on L2 Listening Test Performance

The effect of response format on L2 listening test performance has been investigated in a number of studies (e.g., Cheng, 2004; In'ami & Koizumi, 2009; Wei & Zheng, 2017). Unlike the effects of visual input, results appear to be a bit more consistent across studies: Selected response formats (e.g., multiple-choice questions) tend to be easier than open-ended formats (e.g., short answer questions), which indicates that variation in this aspect of test tasks needs to be carefully considered and, ideally, controlled for when assessing L2 listening skills (Brindley & Slatyer, 2002).

In a study by Cheng (2004), participants took a listening test which contained three response formats presented in a balanced test design: traditional multiple-choice, multiple-choice cloze, and open-ended (i.e., short answer questions). Mean scores for each of the three formats were compared and results indicated significantly higher scores on both of the selected response formats, with the multiple-choice cloze format achieving the highest mean scores. A similar study by Teng (1998, as cited in Cheng, 2004) investigated the effect of response format (multiple-choice, cloze, and short answer) and text type on listening test performance and found that test-takers scored highest in the multiple-choice format, while the cloze format yielded the lowest scores.

Brindley and Slatyer's (2002) study looked at the effect of a range of variables (speech rate, text type, number of hearings, input source, and item response format) on task difficulty in the listening section of an assessment for adult immigrants in Australia. The different response formats included sentence completion items, a table completion task, and short answer questions. Analysis of individual items revealed that the interaction among three components of the tasks (the necessary information, the surrounding text, and the stem; see p. 382) appeared to explain differences in performance, rather than any individual variable (e.g., response format). The results of this study demonstrate the complex

interaction among test task characteristics, including their combined effects on performance, and the need to examine these interactions further.

In a validation study of the listening section of the recent Pearson Test of English Academic (PTE), Wei and Zheng (2017) explored whether integrated and independent listening tasks measured the same underlying listening construct and how different item types/response formats (including selected and constructed response formats) performed in terms of difficulty and item type effectiveness (i.e., item discrimination). The analysis of item performance indicated that difficulty was highest for the constructed response item types (which involved listening and writing). However, multiple-choice with multiple answer items were the second most difficult, suggesting that response format alone does not determine difficulty or level of performance. In addition, as was found in Brindley and Slatyer's (2002) study, other test task characteristics, such as whether utilization of co-text was needed to answer a given question, were important predictors of test-taker performance.

Finally, a meta-analysis by In'ami and Koizumi (2009) looked at the effect of multiple-choice and open-ended response formats on test-taker performance in first language (L1) reading, L2 reading, and L2 listening. Results of their study found that multiple-choice formats are easier than open-ended formats for L1 reading and L2 listening, and in the case of L2 listening, the effect of format ranged from medium to large.

The studies discussed here indicate an overall advantage of selected response over constructed response formats, in terms of test-taker performance. Nevertheless, Taylor and Geranpayeh (2011) assert that "a good quality academic listening test will include a range of task types and response formats, rather than rely on a single test method, e.g., 4-option multiple-choice" (p. 95). They go on to note that "choice of response method has major implications not only for the type of cognitive processing that is provoked but also for scoring validity" (Taylor & Geranpayeh, 2011, p. 97; see also Field, 2013, pp. 144-145). Thus, in addition to differences in item difficulty, different response formats may actually elicit different cognitive processes and therefore may measure different latent traits (Hohensinn & Kubinger, 2011, pp. 733-734). The differences in performance noted above and the possibility of measuring different underlying constructs raises important questions about "interactional authenticity" (Bachman, 1991, as cited in Buck, 2001, p. 126) or "cognitive validity" (Field, 2013), as well as construct irrelevant variance and test fairness (Miller, Linn, & Gronlund, 2009).

In sum, the literature indicates that selected response formats are generally easier, more preferable to learners (Cheng, 2004), typically more practical in terms of administration and scoring (Field, 2013, p. 145), and generally more consistent (though not necessarily more reliable; see Kastner & Stangl, 2011). Nevertheless, an important question remains: Is the multiple-choice format the most *valid* means of measuring L2 listening ability—especially in an academic TLU domain? For many (e.g., Winke & Isbell, 2018), the answer is "no." Considering the notions of construct and cognitive validity (Bachman & Palmer, 1996; 2010; Field, 2013), in addition to the greater vulnerability of multiple-choice questions to the use of test-wise strategies (Field, 2013; Suvorov, 2018; but see also Lee & Winke, 2012), it is reasonable to question the validity of inferences made about listening ability based solely on this response format.

Furthermore, one wonders whether there is an optimum combination of response format and input type, such that test-takers are not disadvantaged, and the best possible inferences can be made about their listening ability. Though some researchers have indicated the potential combined effects of test task characteristics on test scores (e.g., Bloomfield et al., 2011; Taylor & Geranpayeh, 2011), to date very few studies have investigated their interaction directly.

The Current Study: Rationale and Research Questions

Based on the literature reviewed above, it is apparent that more research is needed in order to better understand the role that input type and response format play in tests of L2 listening comprehension. Importantly, at the time of writing, no studies could be identified which look specifically at the interaction between input type (video vs. audio-only texts) and response format (multiple-choice vs. short answer), though studies by Brindley and Slatyer (2002), Cubilo and Winke (2013), and Wagner (2013) have made contributions toward this end.

Understanding the interaction between these variables in tests of L2 listening comprehension has implications for test design and may inform how the construct of academic listening is defined. Thus, this study seeks to provide evidence that can be used to answer the question 'should audiovisual literacy be included in the construct definition of L2 listening?'. This information may help test developers to better identify sources of construct relevant variance and avoid construct underrepresentation (Miller, Linn, & Gronlund, 2009). By doing so, listening tests stand to have a greater degree of validity, authenticity, fairness, and positive washback effects on classroom teaching.

With these observations in mind, the present study seeks to address the following research questions:

> RQ1: To what extent does the presence of visual input (video vs. audio-only) affect performance on a test of L2 listening comprehension?
> RQ2: To what extent does response format (multiple-choice vs. short answer) affect performance on a test of L2 listening comprehension?
> RQ3: To what extent does the interaction between the presence of visual input and response format affect L2 listening comprehension test scores?

Methods

Design

The present study employed a quasi-experimental, between-groups design (Hatch & Lazaraton, 1991) with one dependent variable (scores on an academic listening comprehension test) and two independent variables, each with two levels: input type (video vs. audio-only) and response format (short answer vs. multiple-choice). Four mixed ability groups took an academic listening test corresponding to one of four combinations of task and input type. Average scores of each group were then compared in order to determine if there was a statistically significant difference in performance among the groups. The data was also analyzed in order to reveal if there was an interaction effect among the variables.

Participants

Sixty-eight Japanese undergraduate students enrolled in an academic English program at a public university in the south of Japan participated in this study. Participants comprised a convenience sample recruited from four intact classes. The academic English program (henceforth, the AEP—a pseudonym for the actual program) was a special course of classes taken in lieu of the university's general English program. AEP classes conferred the same credits as general English classes; however, they required a much greater time commitment and heavier work load. Because membership in the AEP was voluntary (interested students self-selected to participate in the Program), it could be inferred that these students possessed a high level of intrinsic motivation to learn English for academic purposes.

At the time of the study, students had received approximately one semester of EAP instruction in the Program. They possessed mixed proficiency levels, with overall TOEFL PBT (Test of English as a Foreign Language, Paper-Based Test) scores ranging from 377 to 583, and an average score of 481 at the beginning of the semester. As a point of reference, a TOEFL PBT score of 377 corresponds to the A2 level on the Common European Framework of Reference for Languages (CEFR); 583 corresponds to the CEFR B2 level, and 481 corresponds to CEFR B1 (ETS, 2019). In terms of CEFR proficiency levels, these TOEFL scores would classify approximately 33% of learners as A2 (range = 337-459), 59% as B1 (range = 460-542), and 8% as B2 (range = 543-627) (ETS, 2019). This information is summarized in Table 1.

Table 1 *Distribution of Participants' Overall TOEFL Scores in Terms of CEFR Levels*

TOEFL Score Range	CEFR Level	*n*	%
337-459	A2	22	33
460-542	B1	40	59
543-627	B2	6	8

Note. Score range = 377-583. Information adapted from ETS (2019).

Among the participants, nearly all were aged 18 to 19 (with the exception of one student, who was 26 years old). Twenty-eight students were males and 40 were female.

Materials

Two versions of a 26-item academic listening test (henceforth, the ALT)—one with multiple-choice questions and one with short answer questions—were created by the researcher for use in the study. A pilot version of the test was first administered independently to two highly proficient (CEFR C2 level) Japanese nonnative speakers of English and two native speakers of English in order to determine that all questions, answer choices, directions, and visual information (i.e., any content visuals included from the video; see below) were clearly indicated and unambiguous in both versions of the test. Any discrepancies were discussed and changes to the test were made accordingly.

The ALT contained four listening texts, each with five to eight comprehension items based on Buck's (2001) default listening construct (see Appendix A). Multiple-choice and short answer versions of the test featured stem-equivalent items (In'ami & Koizumi, 2009), which theoretically required test-takers to listen for and comprehend the same information but demonstrate their comprehension by either selecting or constructing a response. Similar to the test used by Apostolou (2010), the multiple-choice items contained a stem and three answer choices (one key and two distractors). The directions for the short answer versions of the ALT informed students "you *do not* need to write complete sentences, but you must directly answer the question" (emphasis in the original directions; see also Appendix B). These directions were intended to simplify the task demands and lessen the cognitive load of constructing a response while listening and (potentially) attending to the visual input of the listening text (see Cross, 2011; Cubilo & Winke, 2013).

In an effort to narrowly operationalize "visual input" as the kinesic, contextual, and/or embodied aspects of communication present in the videotexts (e.g., facial cues, gestures, gaze, spatial arrangements of the classroom setting, and other embodied action; see examples in Appendix C), all content visuals (i.e., any images in the video [other than the speaker] that conveyed specific meaning related to the aural text, e.g., pictures, written words; see Cross, 2011; Ginther, 2002) from the videotexts were captured via computer screenshot, then copied and pasted into all test booklets (see Appendix D). Thus, any unique advantageous effect these visuals might otherwise have for the video groups were controlled for.

The four listening passages were "lecturettes" (i.e., short, videotaped academic lectures for the purpose of instruction; cf. Wagner, 2007) on topics in psychology, business, and sociology, delivered by three different speakers (one of the speakers was featured in two different videotexts). Though there were a total of four different testing conditions, the audio input was identical across the conditions; in the case of the audio-only groups, the video monitors were turned off when the texts were played (see Data Collection Procedures below). In the videotexts, speakers were shown mostly from the waist up, speaking in a typical university classroom setting. The length of the texts ranged from 01:35 to 02:48 (see Appendix A), with an average time of 02:06. The video listening texts were taken from the EAP textbook series *Lecture Ready*, books *1*, *2*, and *3* (Frazier, & Leeming, 2013; Sarosy & Sherak, 2013a; Sarosy & Sherak, 2013b) and thus had an academic TLU domain (Bachman & Palmer, 2010). Though the texts were scripted and performed by actors, they were believed to realistically simulate an academic lecture, as they included features of unplanned discourse, such as false starts and were spoken in a formal academic register (Biber & Conrad, 2009).

Data Collection Procedures

Students took the ALT in the last week of the semester during normal class time as part of their Listening and Speaking course. To take the test, participants were placed into one of four experimental groups, each corresponding to a different combination of task and input type: (1) audio-only with multiple-choice questions [AO, MC]; (2) video with multiple-choice questions [VID, MC]; (3) audio-only with short answer questions [AO, SA]; and (4) video with short answer questions [VID, SA]. Testing groups were formed using stratified random assignment based on students' TOEFL PBT listening subsection scores (obtained prior to the beginning of the semester). These groupings, along with descriptive statistics of each group's TOEFL listening scores, are summarized in Table 2.

Table 2 *Testing Group Assignment and Associated TOEFL PBT Listening Scores*

Testing Group	*n*	*M*	*SD*	95%CI
Group 1 [AO, MC]	16	46.79	5.94	[43.36, 50.21]
Group 2 [VID, MC]	15	49.00	5.62	[46.00, 52.00]
Group 3 [AO, SA]	18	48.69	5.23	[45.53, 51.86]
Group 4 [VID, SA]	19	48.94	5.51	[46.00, 51.87]

Note. Score range = 31-68.

It is important to note that prior to taking the ALT, students had had multiple opportunities throughout the semester to practice listening and responding to questions in each of the four testing conditions. They had also received explicit instruction in a variety of listening strategies, including how to utilize visual input to support listening and how to respond appropriately to short answer questions based on the specific directions given. The ALT was the final of three graded listening quizzes; previous quizzes featured different combinations of task and input type, such that students would have experience with all possible combinations of these test task characteristics. Each previous quiz was followed by a review session where answers were explained and strategies were given for dealing with the different test task formats—all of which was intended to have a positive washback effect on students' development of academic listening skills, regardless of their testing group.

The test was administered in students' normal classrooms, which were nearly identical across test administrations. The rooms contained a high-quality projector, large projector screen, and speakers. For the video groups, the videotexts were projected on the projector screen and audio was played (simultaneously) through the classroom speakers. For the audio-only group, the exact same listening texts were played, but the video monitor was turned off, allowing only the audio to be transmitted. Each test administrator (the three teachers in the AEP, including the researcher) followed a protocol

that explained the exact procedures of test administration in order to ensure consistency across the groups. Students were told to take a seat and put away all materials except pens, pencils, and erasers. The video groups were instructed to sit within an appropriate distance to the screen in order to facilitate viewing of the visual content. The test administrators then distributed the test booklets and informed students to put their name and student identification number on the cover page and refrain from opening the booklet until instructed to do so. Students were informed about the general format and procedures of the test (i.e., number of texts, number of questions in total and per listening, question type, and approximate total length of the test) and were encouraged to put forth their best effort. Students were then told to open their test booklets to begin the test.

Each of the listening texts was played twice and included one minute to read the directions and preview the questions before listening, 30 seconds to review questions and answers between the first and second listening, and 10 seconds to finish answering questions after the second time listening to a given text. This procedure was repeated for each of the four listening texts. The total time for the test took approximately 25 minutes.

All test items were scored dichotomously (correct answers received one point; incorrect answers received zero points). Short answer tests were scored independently by two raters; agreement across all judgments was 92.8% and the correlation between students' total scores from each rater was $r = .76$. Scoring discrepancies were resolved through discussion in order to arrive at final scores for analysis.

Data Analysis

To answer the three research questions in this study, a 2 x 2 factorial ANOVA was used to analyze the effects of input type and response format (the independent variables) on test scores (the dependent variable), and to determine if there was an interaction among these variables. Analyses were performed using R (R Core Team, 2019). Prior to running the ANOVA, homogeneity of ALT score variance across groups was checked. The results of a Levene's test, $F(3, 64) = 2.06$, $p = .11$, suggested that the difference in group variances was not statistically significant. Thus, the assumptions required for the ANOVA procedure were met (Hatch & Lazaraton, 1991, p. 384).

Results

The ALT contained 26 items, each worth one point, for a total maximum score of 26 points. Internal consistency was calculated for each version of the ALT ([AO, MC; $\alpha = .26$]; [VID, MC; $\alpha = .59$]; [AO, SA; $\alpha = .66$]; [VID, SA; $\alpha = .65$]). Table 3 presents summary statistics for each experimental group. The average scores of Groups 1 and 2 were nearly equal, while Groups 3 and 4 were lower. A visual representation of these group means is given in Figure 1. Table 4 displays descriptive statistics for input modality and response format groupings. As can be seen in Table 4, the mean scores of each input group are nearly equal; however, the mean scores for the multiple-choice group are higher than the short answer group.

Table 3 *Results from the Academic Listening Test*

Testing Group	*n*	*M*	*SD*	95%CI
Group 1 [AO, MC]	16	22.62	1.86	[21.64, 23.61]
Group 2 [VID, MC]	18	22.94	2.26	[21.82, 24.07]
Group 3 [AO, SA]	15	17.20	3.49	[15.27, 19.13]
Group 4 [VID, SA]	19	16.42	3.40	[14.78, 18.06]

Table 4 *Results from the Academic Listening Test by Input Modality and Response Format Groupings*

	Results by Input Modality			
Testing Group	*n*	*M*	*SD*	95%CI
Audio-Only	31	20.00	3.87	[18.58, 21.42]
Video	37	19.59	4.37	[18.13, 21.05]
	Results by Response Format			
Testing Group	*n*	*M*	*SD*	95%CI
Multiple-Choice	34	22.79	2.06	[22.07, 23.51]
Short Answer	34	16.76	3.41	[15.57, 17.95]

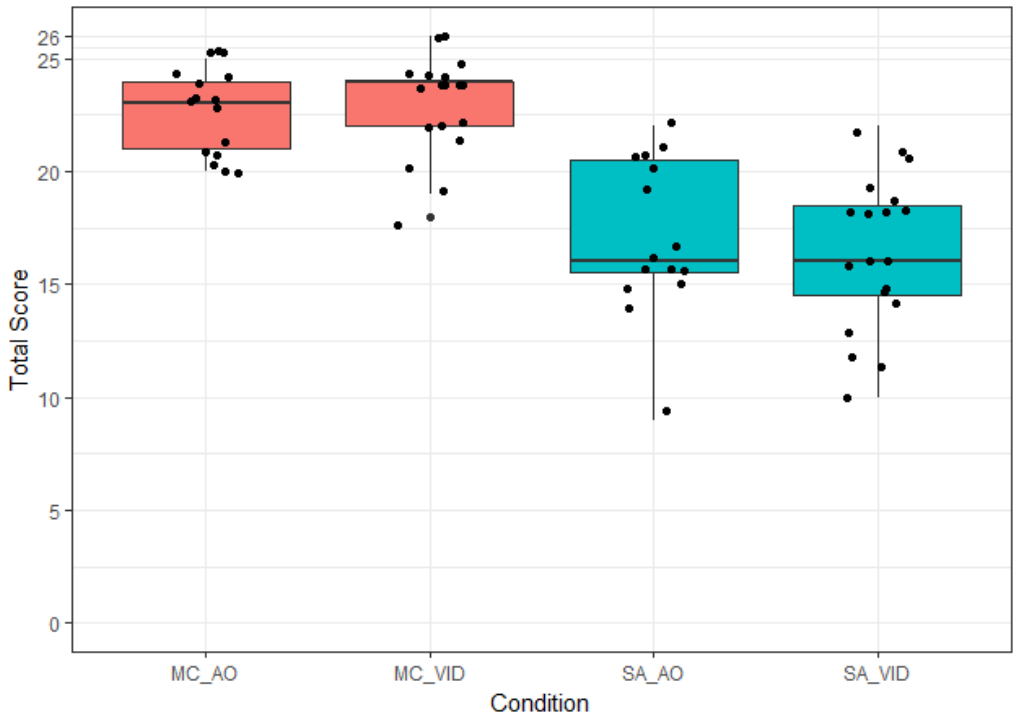

Figure 1 *Group means on the Academic Listening Test (ALT)*

Results of the 2 x 2 ANOVA, which was used to analyze the effects of input type (video vs. audio-only) and response format (multiple-choice vs. short answer) on students' ALT scores, are shown in Table 5. As can be seen in the table, there was a statistically significant effect of the main effect of *format*, $F(1, 64) = 76.16$, $p < .001$, and the effect size was large (partial $\eta^2 = .54$). Tukey HSD post hoc tests revealed that students who took a multiple-choice version of the ALT scored significantly higher ($M = 22.79$, $SD = 2.06$, $n = 34$) than students who took a short answer version ($M = 16.76$, $SD = 3.41$, $n = 34$). There was no statistically significant effect for the main effect of *input* and there was no statistically significant interaction found between input and format.

Table 5 *Results of the Factorial Analysis of Variance*

Source	**df**	***SS***	***MS***	**F**	**P**
Input	1	2.8	2.8	0.343	0.56
Format	1	616.1	616.1	76.16	< .001*
Input x Format	1	5.1	5.1	0.63	0.43
Error	64	517.7	8.1		

$*p < .001$.
Note. SS = Sum of Squares. MS = Mean Square.

In sum, input type (video vs. audio-only) did not have a significant effect on test-taker performance. However, response format did have a significant effect: Test-takers who took a multiple-choice version of the ALT scored higher than those who took a short answer version of the test. The large effect size (**η^2** = .539, partial η^2 = .543) of this factor suggests that approximately 54% of the observed variation in scores was due to the difference in response format. As noted above, no interaction between the independent variables (input type and response format) was found.

Discussion

Building on findings and lingering questions from previous investigations (e.g., Pusey & Lenz, 2014), this study sought to advance knowledge in the field of language testing in regards to how L2 listening test task characteristics affect test-taker performance. Given the essential role of visual input in nearly all forms of verbal communication, including the TLU domain of academic listening, it was hypothesized that the video group would attain higher scores on the ALT than the audio-only group. To test this hypothesis, the first research question (RQ1) asked 'To what extent does the presence of visual input (video vs. audio-only) affect performance on a test of L2 listening comprehension?'. Results showed that test-takers in the video group (n = 37, M = 19.59, SD = 4.37) did not score significantly differently than the audio-only group (n = 31, M = 20.00, SD = 3.87). In fact, the average score was slightly lower for the video group.

The results of RQ1 are thus similar to findings from several other studies (e.g., Brett, 1997; Coniam, 2001; Cubilo & Winke, 2013; Londe, 2009) in which no significant main effect for visual input on listening comprehension test scores was found. The findings contrast, however, with those of Wagner (2010; 2013), Sueyoshi and Hardison (2005), and others (e.g., Parry & Meredith, 1984; Shin, 1998), who found a significant facilitative effect of visual input on test performance. Furthermore, though not statistically significant, the fact that test-takers in the video group obtained slightly lower scores than those in the audio-only group suggests that the video input may have inhibited the performance of some test-takers, as was the case in Suvorov (2009) and Pusey and Lenz (2014). These findings therefore seem to reflect the cautionary remarks of Buck (2001), who observes, "We do not know how video-texts [*sic*] affect listening comprehension, nor whether tests with video-texts [*sic*] are in any significant way different from audio-texts" (p. 253).

Based on the literature (e.g., Buck, 2001; Coniam, 2001; Gruba, 1993; Suvorov, 2009) and the findings of the current study, it may be reasonable to suggest that, in many testing situations, one mode of input (i.e., video or audio-only) is not inherently superior to the other. Rather, as suggested in a number of studies (e.g., Cubilo & Winke, 2013; Ockey, 2007; Wagner, 2007; 2008), the utility of visual input may be a matter of individual differences and natural variation in test-taking behavior (e.g., perceived value of visual input, strategic use of visuals to aid in listening comprehension, familiarity with multimodal genres of L2 listening).

The question thus arises: What *could* video lectures featuring speakers shown from the waist-up and minimal content visuals (as in the present study) add to "default construct" (cf. Buck, 2001) listening comprehension? Possible answers to this question are suggested in a general sense in the Literature Review. However, in regards to the present study specifically, qualitative analysis of the videotexts (which did not constitute part of the primary data analysis) reveals a variety of non-verbal, multimodal, and contextual information that may help "frame" (Goffman, 1974) the speech event (a classroom lecture), activate schemata, and aid the listener in interpreting the speakers' intended meanings. As Hall (2019) notes:

> *Teachers must calibrate their language, facial expressions, gestures, body positions, and even the use of material artifacts such as a textbook or smart pad such that the pedagogical project is advanced, the shared attention of students is maintained, and individual students' participation in promoted. (p. 47)*

Such embodied action (Matsumoto, 2018, 2019; Matsumoto & Dobs, 2017) thus typifies classroom interaction, is thought to enhance communication, and is observable in the videotexts used in the present study. For example, in the videotext "Neuromarketing" (described in Appendix A), the speaker uses hand gestures throughout her lecture to reinforce the semantic content of her message, such as making a downward motion with her hands as she says *under the surface* (see Appendix C). In the videotext "Staycations," the speaker combines verbal and non-verbal communicative behaviors, including word stress, raised eyebrows, a smiling facial expression, upraised hands, and shifting gaze (i.e., panning across the classroom) to create a sense of irony as she defines "staycation": ***Staycation****... is the new term for remaining* ***at home****... during vacation time* (words in bold indicate word stress; see also Appendix C). These examples demonstrate the potential for the non-verbal, kinesic aspects of communication present in the videotexts to amplify the speakers' messages. Coupled with the other semiotic resources in the videotexts, it could be expected that these visual elements would provide a relative advantage to test-takers in the video groups. Yet, this advantage was not borne out in the test results.

Cultural differences may play some role as well. As Wagner (2007) notes, "It could be that test-takers from a particular cultural group might be more inclined to orient to the video monitor compared with test-takers from another cultural group" (p. 78). If this particular group of test-takers were 'less inclined to orient to the video monitor,' even if the video had had a potentially facilitative effect, it would not have been reflected in the quantitative data. (This question is taken up in a forthcoming, complementary qualitative study by the author.)

Beyond any speculation about cultural differences, it could be expected that some learners—in any cultural context, whether L1 or multilingual users of English—prefer to watch the video while others prefer to look at their test booklets, especially if there is no explicitly stated requirement to utilize the visual channel (see Conclusions and Implications below). Thus, what is observed in the testing literature (e.g., Ockey, 2007; Wagner, 2007; 2008) is not unlike what one might observe in real life: In actual lectures, students often look down—sometimes for large portions of the lecture—in order to take copious notes. Unless otherwise prompted, a given individual need not necessarily visually attend to the [primarily] aural input that one is receiving. Nevertheless, the findings of RQ1, though similar to others in the literature, and perhaps expected based on individual differences, are still somewhat puzzling.

Response format, on the other hand, did confer a distinct performance advantage for test-takers: Those who took a multiple-choice version of the test scored significantly higher than test-takers who took a short answer version of the test. Thus, the answer to RQ2 'To what extent does response format (multiple-choice vs. short answer) affect performance on a test of L2 listening comprehension?' was quite clear: Results showed not only a significant difference in test scores between the two groups (the multiple-choice group [M = 22.80, or 88%] scored approximately 24% higher than the short answer group [M = 16.76, or 64%]), but also a fairly large effect size ($\eta^2 = .539$).

These results corroborate what has been found in previous studies (e.g., Cheng, 2004; In'ami & Koizumi, 2009), which may be related to several factors. The number of distractors used in the multiple-choice items (in this case, three), for example, could have made the test questions relatively easier than may have been the case with four or five answer choices (Lee & Winke, 2012). However,

there is no strict consensus in the literature on the number of distractors that should be used (see Buck, 2001, pp. 142-143). Furthermore, as long as the alternatives are plausible, it has been claimed that "there is little difference in difficulty, discrimination, and test score reliability among items containing two, three, and four distractors" (Brame, 2013). In developing the ALT, three distractors were used in order to minimize the cognitive demands of discriminating among multiple answer choices (Lee & Winke, 2012), as well as to reduce the introduction of construct-irrelevant variance related to reading ability.

The number of times the text is played (Wagner, 2007, pp. 71-72) as well as the availability of question preview (Koyama, Sun, & Ockey, 2016; Wagner, 2013) are also factors that could have advantaged test-takers in the multiple-choice group in particular. Allowing for question preview gives test-takers the chance to look at the answer choices, possibly annotate their test booklet (e.g., underlining keywords) and often grasp a general idea of the listening text. Playing the text multiple times (for the ALT, each text was played twice) would further enable students to use test-wise strategies, such as the process of elimination, to successfully answer the questions. Those in the short answer group would not benefit to the same extent from these procedural factors due to the limited input in the question stems; thus, their performance may have been unaffected.

The combined effects of the limited number of distractors (in the case of the multiple-choice versions of the test), the number of times each listening was played, as well as the availability of question preview may have further contributed to the low internal consistency of each version of the ALT ([AO, MC; $\alpha = .26$]; [VID, MC; $\alpha = .59$]; [AO, SA; $\alpha = .66$]; [VID, SA; $\alpha = .65$]). The reliability for the audio-only multiple-choice [AO, MC] version of the ALT is particularly poor and is a limitation of this study (see Limitations below). This version of the test resulted in very little variation among test-takers and was apparently easy for this group overall. This may have produced a ceiling effect, as many students in this testing condition clustered near the maximum possible score (see Figure 1).

Although the results of RQ2 were perhaps to be expected based on previous findings (e.g., In'ami & Kozumi, 2009), they once again ran counter to what was originally hypothesized. Given that "MC items… are usually longer than short answer questions, and thus require more attentional resources of the test-taker while the text is playing" (Wagner, 2013, p. 191), it was thought that short answer questions might better allow students to attend to the aural (or visual) input, thus increasing their overall ability to decode and construct meaning from the text (Field, 2008). However, this was not the case, which may have been due to a complex combined effect of input, response format, and possibly other test-task characteristics.

Aptly, RQ3 aimed to identify whether there was an interaction between input type and response format on L2 listening comprehension test scores. In line with Brindley and Slatyer's (2002) observation that "particular combinations of item characteristics appear either to accentuate or attenuate the effect on difficulty" (p. 387), it was hypothesized that there may be an interaction effect among the independent variables. Specifically, it was thought that the short answer response format, by virtue of its relative simplicity, might allow test-takers more freedom to attend to the visual input than would multiple-choice questions, which may in turn lead to better comprehension and thus higher test scores. However, there was no statistically significant interaction between the variables.

The results are thus similar, to some extent, to those of Wei and Zheng (2017), as well as Brindley and Slatyer (2002), who point out that "any conclusions regarding the effect of *any single task or item characteristic* [emphasis added] on difficulty…need to be carefully qualified in the light of what is known about its interaction with other variables" (p. 387). Considering the range of variables that characterize and potentially intervene in testing situations, studies such as this one are needed to help

clarify how and under what circumstances these task and item characteristics interact with one another, and how to plan accordingly.

Limitations

There are a number of limitations to this study. First, the participants in this study constituted a culturally homogenous group and intact classes were used. Thus, claims of generalizability are limited and the findings of the study should be interpreted as such. Another limitation was that the listening materials used in the study were not authentic. Although the texts chosen for the study were believed to exhibit fairly realistically the linguistic and situational characteristics of a spoken academic register (Biber & Conrad, 2009), the scripted nature of the texts may have had some effect on test-taker performance, which is a potential threat to validity. Finally, the low internal consistency (reliability) of the four tests—particularly the multiple-choice audio-only version—was a notable shortcoming of the present study. A more extensive piloting and test refinement procedure should be used in future studies.

Conclusions and Implications

This study found that use of the short answer response format led to significantly lower performance on a test of L2 listening comprehension in comparison with the multiple-choice format. Test developers, researchers, and teachers need to be cognizant of the response format they utilize in the tests they create, and the differences in performance that each format is likely to bring about. The presence of visual input, on the other hand, did not significantly affect test scores and did not interact with response format to advantage or disadvantage any particular group. It is possible that the task demands of the ALT (whether writing short answer responses or reading and selecting among answer choices) or individual differences in test-taking behavior may have diminished the usefulness of video input for many test-takers (cf. Ockey, 2007; Wagner, 2008). Indeed, Wagner (2013) asserts that "providing audiovisual texts allows the test-taker to *choose* [emphasis added] whether he or she wants to attend to that visual input, ignore the input, or attend to the visual input only part of the time" (p. 193). Thus, unless explicitly required for task completion, individual differences in test-taking behavior may ultimately determine how, when, and whether (or not) test-takers utilize visual input in L2 listening tests.

In recognition of the prominent role that visual information plays in real-world listening, it may be appropriate to advance alternative forms of listening assessment that explicitly target audiovisual literacy as complementary to—but distinct from—Buck's (2001) "default listening construct" (cf. Ockey, 2007). Such assessments would need to explicitly define the audiovisual literacy skills to be measured within the construct definition of listening ability. Utilization and comprehension of predetermined visual aspects of the audiovisual texts would need to be reflected in scoring criteria, thus constituting a dimension of the expected response (Bachman & Palmer, 2010). On a practical note, these audiovisual literacy skills would need to be explicitly stated in the test directions so that test-takers would be unambiguously aware of the need to attend to the visual input of audiovisual texts in order to locate and provide the necessary information (Cross, 2011) for the expected responses.

Beyond the implications for L2 listening assessment, this study has several implications for teaching. If communicative competence is understood, in part, as the competent manipulation and interpretation of semiotic resources in use (Atkinson, 2011; Douglas Fir Group, 2016), then teachers need to guide students in how and when to utilize the variety of semiotic resources available to them (e.g., textual and visual information on PowerPoint slides, diagrams, charts, handouts, gestures, facial cues, intonation)—whether they are taking a test, communicating in and out of the classroom, or completing

some other language use task. However, given that listening tests do not typically include many of the semiotic resources (or "mediation tools"; Lantolf, 2011) that are normally present in the TLU domain of academic listening (e.g., questioning the instructor or peers, accessing the Internet, consulting dictionaries or translation devices), and additionally introduce situational factors that may negatively impact performance (e.g., time pressure, anxiety), students need to learn strategies for coping with resource-deficient listening conditions. In regards to the results of the present study specifically, it seems that students may require extensive practice with different response formats (particularly short answer questions) in order to prepare them for both real-world and test-specific listening.

Further research is needed in order to better understand the relationship among different combinations of test task characteristics and their effects on L2 listening performance. One promising area of research is the use of eye-tracking technology (Suvorov, 2015; Winke, Gass, & Sydorenko, 2013), which could be used to investigate how test-takers utilize visual input in video-based tests of listening comprehension. For example, this technology could be used to correlate viewing behavior with task "interactiveness" (Bachman & Palmer, 1996; Buck, 2001; Li, 2013), as well as identify specific attributes of visual input that either facilitate or detract from test performance (Cross, 2011). Research on response formats that do not rely on literacy skills (i.e., reading and writing ability) is also needed. Looking into interactional listening, for example, may strengthen the theoretical grounding for including visual input within a construct definition of listening ability (Flowerdew & Miller, 2005; Lynch; 2011; Rost, 2011) and would allow for alternative response formats for gauging listening ability. Finally, experimenting with live lectures and authentic listening tasks (see Field, 2011), including qualitative investigations of L1 and L2 lecture listening behavior, may provide valuable insight into the myriad ways that students listen, respond to, and interact with multimodal input in academic contexts.

Hazards and Human or Animal Subjects Statement

The procedures involved in conducting this research were performed in compliance with relevant laws and institutional guidelines.

Declaration of Conflicting Interests

The author declares no financial or other substantive conflict of interest with respect to the results or interpretation of this manuscript.

References

Apostolou, E. (2010). Comparing perceived and actual task and text difficulty in the assessment of listening comprehension. In *Lancaster University Postgraduate Conference in Linguistics & Language Teaching* (pp. 26-47). Lancaster: Department of Linguistics and English Language (LAEL), Lancaster University.

Atkinson, D. (2011). Introduction: Cognitivism and second language acquisition. In D. Atkinson (Ed.), *Alternative approaches to second language acquisition* (pp. 1-23). Oxford: Routledge.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice.* New York, NY: Oxford University Press.

Biber, D., & Conrad, S. (2009). *Register, genre, and style* [Cambridge Textbooks in Linguistics]. Cambridge: Cambridge University Press.

Bloomfield, A., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2011). *What makes listening difficult? Factors affecting second language listening comprehension* (Technical Report TTO 81434 E.3.1). College Park, MD: University of Maryland Center for Advanced Study of Language.

Brame, C. (2013). *Writing good multiple choice test questions*. Retrieved from https://cft.vanderbilt.edu/guides-sub-pages/writing-good-multiple-choice-test-questions/

Brett, P. (1997). A comparative study of the effects of the use of multimedia on listening comprehension. *System, 25*, 39-53. http://dx.doi.org/10.1016/S0346-251X(96)00059-0

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

Brinton, D. M. (2014). Tools and techniques of effective second/foreign language teaching. In M. Celce-Murcia, D. M. Brinton, & M. A. Snow (Eds.), *Teaching English as a second or foreign language* (4th ed., pp. 341-361). Boston, MA: Heinle Cengage.

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

Cheng, H. F. (2004). 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

Coniam, D. (2001). The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: A case study. *System*, *29*, 1-14. http://dx.doi.org/10.1016/S0346-251X(00)00057-9

Cross, J. (2011). Comprehending news videotexts: The influence of the visual content. *Language Learning & Technology*, *15*(2), 44-68. Retrieved from http://llt.msu.edu/issues/june2011/cross.pdf

Cubilo, J. & Winke, P. (2013) Redefining the L2 listening construct within an integrated writing task: Considering the impacts of visual-cue interpretation and note-taking. *Language Assessment Quarterly 10*(4), 371-397. https://doi.org/10.1080/15434303.2013.824972

Douglas Fir Group. (2016). A transdisciplinary framework for SLA in a multilingual world. *Modern Language Journal*, *100* (Supplement 2016), 19-47. https://doi.org/10.1111/modl.12301

ETS. (2019). Research: CEFR mapping study. Retrieved from https://www.ets.org/toefl_itp/research

Feak, C. B., & Salehzadeh, J. (2001). Challenges and issues in developing an EAP video listening placement assessment: A view from one program. *English for Specific Purposes*, *20*, 477-493. http://doi.org/10.1016/S0889-4906(01)00021-7

Field, J. (2008). *Listening in the language classroom*. Cambridge: Cambridge University Press.

Field, J. (2011). Into the mind of the academic listener. *Journal of English for Academic Purposes*, *10*(2), 102-112. https://doi.org/10.1016/j.jeap.2011.04.002

Field, J. (2013). Cognitive validity. In A. Geranpayeh, & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language* (pp. 77-151). Cambridge: Cambridge University Press.

Flowerdew, J., & Miller, L. (2005). *Second language listening: Theory and practice*. New York, NY: Cambridge University Press.

Frazier, L., & Leeming, S. (2013). *Lecture ready 3: Strategies for academic listening and speaking* (2nd ed.). New York: Oxford University Press.

Gan, Z. (2012). Test-task authenticity: The multiple perspectives. *Changing English 19*(2), 237-247. https://doi.org/10.1080/1358684X.2012.680765

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing, 19*, 133-167. https://doi.org/10.1191/0265532202lt225oa

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Cambridge, MA, US: Harvard University Press.

Gregersen, T. S. (2007). Language learning beyond words: Incorporating body language into classroom activities. *Reflections on English Language Teaching, 6*(1), 51-64. Retrieved from

http://www.nus.edu.sg/celc/research/books/relt/vol6/no1/51-64gregersen.pdf

Gruba, P. (1993). A comparison study of audio and video in language testing. *JALT Journal, 15*, 85-88.

Hall, J. K. (2019). *Essentials of SLA for L2 teachers: A transdisciplinary framework*. New York, NY: Routledge.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics.* Boston, MA: Heinle & Heinle Publishers.

Hohensinn, C., & Kubinger, K. D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement*, *71*(4), 732-746. https://doi.org/10.1177/0013164410390032

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing, 26*(2), 219-244. https://doi.org/10.1177/0265532208101006

Kastner, M., & Stangla, B. (2011). Multiple choice and constructed response tests: Do test format and scoring matter? *Procedia-Social and Behavioral Sciences*, *12*, 263-273. https://doi.org/10.1016/j.sbspro.2011.02.035

Kellerman, S. (1992). I see what you mean: The role of kinesic behaviour in listening and the implications for foreign and second language learning. *Applied Linguistics, 13*(3), 239-258. http://doi.org/10.1093/applin/13.3.239

Koyama, D., Sun, A., & Ockey, G. J. (2016). The effects of item preview on video-based multiple-choice listening assessments. *Language Learning & Technology*, *20*(1), 148-165. Retrieved from http://llt.msu.edu/issues/february2016/koyamasunockey.pdf

Lantolf, J. P. (2011). The sociocultural approach to second language acquisition: Sociocultural theory, second language acquisition, and artificial L2 development. In D. Atkinson (Ed.), *Alternative approaches to second language acquisition* (pp. 24-47). Oxford: Routledge.

Lee, H., & Winke, P. (2012). The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test. *Language Testing, 30*(1), 99-123. https://doi.org/10.1177/0265532212451235

Li, Z. (2013). The issues of construct definition and assessment authenticity in videobased listening comprehension tests: Using an argument-based validation approach. *International Journal of Language Studies*, *7*(2), 61-82.

Londe, Z. C. (2009). The effects of video media in English as a second language listening comprehension tests. *Issues in Applied Linguistics*, *17*, 41-50.

Lynch, T. (2011). Academic listening in the 21st century: Reviewing a decade of research. *Journal of English for Academic Purposes*, *10*(2), 79-88. https://doi.org/10.1016/j.jeap.2011.03.001

Matsumoto, Y. (2018). Challenging moments as opportunities to learn: The role of nonverbal interactional resources in dealing with conflicts in English as a lingua franca classroom interactions. *Linguistics and Education*, *48*, 35-51. https://doi.org/10.1016/j.linged.2018.08.007

Matsumoto, Y. (2019). Embodied actions and gestures as interactional resources for teaching in a second language writing classroom. In J. K. Hall, & S. D. Looney (Eds.), *The embodied work of teaching* (pp. 181-197). UK: Multilingual Matters.

Matsumoto, Y., & Dobs, A. (2017). Pedagogical gestures as interactional resources for learning tense and aspect in the ESL grammar classroom. *Language Learning*, *67*(1), 7-42. https://doi.org/10.1111/lang.12181

Miller, M. D., Linn, R., & Gronlund, N. (2009). *Measurement and evaluation in teaching.* (10th Edition). Upper Saddle River, NJ: Merrill, Prentice Hall.

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, *24*, 517-537. http://dx.doi.org/10.1177/0265532207080771

Parry, T., & Meredith, R. (1984). *Videotape vs. audiotape for listening comprehension tests: An experiment*. (ERIC Document Reproduction Services ED 254 107).

Progosh, D. (1996). Using video for listening assessment opinions of test-takers. *TESL Canada Journal*, *14*, 34-43.

Pusey, K., & Lenz, K. (2014). Investigating the interaction of visual input, working memory, and listening comprehension. *Language Education in Asia*, *5*(1), 66-80. doi:10.5746/LEiA/14/V5/I1/A06/Pusey_Lenz.

R Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Rost, M. (2011). *Teaching and researching listening* (2nd ed.). New York, NY. Routledge.

Sarosy, P., & Sherak, K. (2013). *Lecture ready 1: Strategies for academic listening and speaking* (2nd ed.). New York: Oxford University Press.

Sarosy, P., & Sherak, K. (2013). *Lecture ready 2: Strategies for academic listening and speaking* (2nd ed.). New York: Oxford University Press.

Shin, D. (1998). Using videotaped lectures for testing academic language. *International Journal of Listening*, *12*, 56-79. https://doi.org/10.1080/10904018.1998.10499019

Sueyoshi, A., & Hardison, D. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*, 661-699. https://doi.org/10.1111/j.00238333.2005.00320.x

Suvorov, R. (2009). Context visuals in L2 listening tests: The effects of photographs and video vs. audio-only format. In C. A. Chapelle, H. G. Jun, & I. Katz (Eds.), *Developing and evaluating language learning materials* (pp. 53-68). Ames, IA: Iowa State University.

Suvorov, R. (2015). The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos. *Language Testing*, *32*(4), 463-483. https://doi.org/10.1177/0265532214562099

Suvorov, R. (2018, March). *Investigation of test-taking strategies via eye tracking and cued retrospective reporting*. Paper presented at the American Association for Applied Linguistics (AAAL) Conference, Chicago, IL.

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. https://doi.org/10.1016/j.jeap.2011.03.002

Taylor, R. (2014). Meaning between, in and around words, gestures and postures: Multimodal meaning-making in children's classroom discourse. *Language and Education*, *28*, 401-420. doi:10.1080/09500782.2014.885038

Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. New York, NY: Routledge.

Wagner, E. (2007). Are they watching? An investigation of test-taker viewing behavior during an L2 video listening test. *Language Learning and Technology*, *11*(1), 67-86. Retrieved from: http://llt.msu.edu/vol11num1/pdf/wagner.pdf

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

Wagner, E. (2010). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing*, *27*(4), 493-513. http://doi.org/10.1177/0265532209355668

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

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

Winke, P., Gass, S., & Sydorenko, T. (2013). Factors influencing the use of captions by foreign language learners: An eye-tracking study. *The Modern Language Journal*, *97*(1), 254-275. https://doi.org/10.1111/j.1540-4781.2013.01432.x

Winke, P., & Isbell, D. (2018). Construct of listening. In J. I. Liontas (Ed.), *TESOL encyclopedia of English language teaching*. Hoboken, NJ: Wiley. https://doi.org/10.1002/9781118784235.eelt0618

Author biodata

Kerry Pusey received his MA-TESL from Northern Arizona University. He is a former U.S. Department of State sponsored English Language Fellow, and has taught English for academic purposes in the United States, Macau, Colombia, Japan, Thailand, and Brazil. His research interests include teacher education, curriculum development, language assessment and program evaluation, and multilingual/multicultural education. He will begin doctoral studies in Educational Linguistics at the University of Pennsylvania Graduate School of Education in the fall of 2020.

Appendix A

The Academic Listening Test (ALT): Table of Specifications

Sub-constructs		**Primary Construct: Academic Listening Comprehension**					
		Main Ideas	**Explicit Details**	**Implicit Details**	**Vocab in Context**		
Input		**Obj. 1**	**Obj. 2**	**Obj. 3**	**Obj. 4**	**Total Items**	**%**
Text	**Characteristics**						
LR1_Test01_U01_Video01 **Intro to psychology**	1) Topic: Psychology 2) Genre: Lecture 3) # of speakers: 1 4) Rate of speech: Slow 5) Length: 01:35 (x 2)	2	3, 4, 5	1	6	6	23%
LR2_Test02_U02_Video01 **Staycations**	1) Topic: Sociology 2) Genre: Lecture 3) # of speakers: 1 4) Rate of speech: Medium 5) Length: 02:05 (x 2)	7	9, 11	13	8, 10, 12	7	27%
LR3_Ch01_Lecture2 **Neuromarketing**	1) Topic: Business 2) Genre: Lecture 3) # of speakers: 1 4) Rate of speech: Fast 5) Length: 02:48 (x 2)	14, 19, 21	16, 17, 20	15	18	8	31%
LR1_Test02_U02_Video01 **Polaroid Cameras**	1) Topic: Psychology 2) Genre: Lecture 3) # of speakers: 1 4) Rate of speech: Slow 5) Length: 01:59 (x 2)	22	24	25, 26	23	5	19%
# of items		6	9	5	6	26	100%
Total points/obj		6	9	5	6	26	
Obj. %		23%	35%	19%	23%	100%	

Primary construct, operationalizations, and criteria for correctness	Indicate listening comprehension by responding correctly to a series of multiple-choice (selecting the appropriate answer choice from a selection of three possible answers) or short answer (constructing the requested information in a limited [between 1 - 10 words, on average] written response) main idea, detail, and inference questions based on short, monologic aural texts of an academic nature. Multiple-choice questions require the test-taker to select the appropriate answer choice from a selection of three possible answers. Short answer questions require the test-taker to construct (write) the requested information in a limited (i.e., between 1 - 10 words, on average) written response. For the short answer questions, grammatical (e.g., article use, subject-verb agreement) and mechanical (i.e., spelling) accuracy are not tested, and complete sentences are not required; however, any errors in grammar or mechanics must not interfere with meaning, and written answers must fully respond to the question; no essential information may be left out of the response. All questions, regardless of response format, are worth one point, and are scored dichotomously (1 or 0).
Sub-constructs	
Obj. 1	Comprehend main ideas in a spoken academic lecture, as derived from local and discourse-level text comprehension.
Obj. 2	Comprehend explicitly stated details in a spoken academic lecture.
Obj. 3	Comprehend implicitly stated details in a spoken academic lecture.
Obj. 4	Infer the meaning of vocabulary unequivocally implied in the context of a spoken academic lecture.

Appendix B

Examples of Test Task Directions and Stem-Equivalent Items

1.2 Lecture

Directions: Listen to a lecture from a sociology class and circle the best answer for each question below. You may take notes while you listen. You will hear the lecture twice.

7. What is the main topic of the lecture?

A. The costs of vacations.
B. A growing leisure trend.
C. The benefits of travel.

8. According to the lecture, **"staycation"** means _____.

A. staying in someone else's home
B. vacationing in your local community
C. remaining at home for vacation

9. When did staycations become popular?

A. 2006 – 2007
B. 2007 – 2008
C. 2008 – 2009

1.2 Video Lecture

Directions: Listen to a lecture from a sociology class write a short answer response to the questions below. You *do not* need to write complete sentences, but you must directly answer the question. You will hear the lecture twice.

7. What is the main topic of the lecture?

_____

8. According to the lecture, **"staycation"** means…

_____

9. When did staycations become popular?

_____

Appendix C

Examples of Visual Input

Appendix D

Examples of Content Visuals