THE ROLE OF CONTENT-RICH VISUALS IN THE L2 ACADEMIC LISTENING ASSESSMENT CONSTRUCT: PILOT STUDY

Roman Lesnov

Northern Arizona University

Author Note

Roman O. Lesnov, Department of English, Northern Arizona University.

This research was supported by Northern Arizona University.

Correspondence concerning this report should be addressed to Roman Lesnov, 1916 E 6th Ave Apt 5, Arizona, USA, 86004.

Contact: roman.lesnov@nau.edu

Abstract

There has been no research into the role of content-rich visuals, such as graphs or images in a lecture, in the second language (L2) academic listening construct. The primary purpose of this study was to pilot an L2 academic listening test by comparing test-takers' lecture comprehension in the audio-only versus video-based mode, with the latter affording the possibility to watch content-rich videos of the lectures. The four pre-recorded videos in the study contained content-related audio-congruent cues for about 60% of the video lengths, making them visually content-rich. The secondary purpose of the study was to pilot an L2 listening proficiency test, labeled as the anchor test. Finally, the study piloted a questionnaire eliciting test-takers' opinions about the effect of content-rich videos on academic listening. Twenty-nine English learners took the tests and the questionnaire online as part of one assessment battery. Test items performed as expected in general. The results of the Rasch analysis showed no significant impact of mode at the testlet level and a limited facilitative effect at the item level. Similarly, test-takers' questionnaire responses lent limited support for the inclusion of content-rich videos in the L2 academic listening construct. Taken collectively, the findings largely failed to confirm the researcher's theory-driven expectations of a substantial mode effect on both test-takers' performance and perceptions. However, some findings and trends pointed to a promising potential of the future dissertation study to confirm these hypotheses.

Content-Rich Visuals in the L2 Academic Listening Assessment Construct: Pilot Study

Background

Modern technology has dramatically altered the way second language (L2) academic listening is taught worldwide. Reflecting visually rich characteristics of corresponding authentic contexts, academic L2 listening classes nowadays are filled with different kinds of *new media*, such as videos and power point presentations (Lievrouw, 2011; Lynch, 2011). New media are now ubiquitous in L2 education practices such that they change the nature of L2 education, making it more interactive and multimodal (Royce, 2007).

While the multimodal nature of the L2 listening competence is mainly accepted by L2 scholars, the field of the L2 listening assessment keeps operationalizing L2 listening as an exclusively auditory skill. For instance, existing standardized high-stakes tests of academic English proficiency are mainly visual-free (Kang, Gutierrez Arvizu, Chaipuapae, & Lesnov, 2016). To eliminate this mismatch, a growing number of researchers have advocated for the inclusion of visuals in L2 listening tests (e.g., Ockey, 2007; Suvorov, 2015, Wagner, 2008). Researchers' arguments stemmed from (a) the effects videos on the difficulty of a listening message and (b) learners' perceptions of this difficulty.

Studies investigating the effects of videos on L2 listening comprehension have produced inconclusive results. These studies mostly focused on the comparison of participants' scores on listening tests under audio-only versus video-based conditions. Some researchers found a facilitative effect of video-based visual information on L2 test-takers' listening comprehension (e.g., Sueyoshi & Hardison, 2005; Wagner, 2010b). Other studies reported either no effect (e.g., Batty, 2015; Gruba, 1993; Cubilo & Winke, 2013) or a detrimental effect (e.g., Suvorov, 2009;

Wagner, 2010a), leaving the question about the role of visual information in an L2 listening construct largely unanswered.

In the area of L2 listening assessment, learners' perceptions have been investigated in relation to listening difficulty, motivation, and authenticity. L2 learners tend to agree that visual information decreases listening difficulty and increases listeners' motivation (e.g., Ockey, 2007; Wagner, 2008; 2010a). Research into authenticity perceptions is much less abundant and less conclusive (Coniam, 2001; Cubilo & Winke, 2013), calling for more investigations in this area.

There are three major gaps in the research into video effects on L2 listening comprehension and perception. First, attempts to unravel the role of content-rich videos in the L2 listening comprehension construct have been scarce. New ways to account for content richness in videos are needed, which would help future studies to better control for video type in their investigations of delivery mode effects. Second, previous research has barely attempted to investigate whether individual test items tested any information presented via the video channel. Batty (2015) seem to be the only study that investigated the semantic relationship between the items used in the test and the visual content of videos. Finally, more research is needed that would investigate effects of content-rich videos on learners' perceptions of listening difficulty, motivation, and authenticity by video type.

Research Questions

The study was governed by the following research questions.

1. Do content-rich videos affect L2 academic listening comprehension difficulty?
2. Do test-takers' perceptions lend support for using content-rich videos in the L2 academic listening assessment construct?

Methods

Participants

Participants in the study were 29 ESL and EFL learners from multiple locations, recruited in the summer of 2017. Sixteen of them (mainly 18-to-35-year-old Chinese) were students in the Program in Intensive English at Northern Arizona University.

Instruments

There were three instruments in the study – the academic listening comprehension (ALC) test, the anchor test, and the questionnaire. Each of them is described below.

ALC test. The listening test contained four passages. To turn the four passages into testlets, each passage was followed by six 4-option multiple-choice questions assessing students' ability to infer main ideas ($k = 1$), to identify supporting details ($k = 3$), and to make inferences based on the listening ($k = 2$). Each question was dichotomously scored (i.e., 0 or 1), setting the overall possible score to 24 points.

The development of each of the four testlets consisted of four main steps. First, four authentic video passages were found on YouTube. Second, four new videos were recorded to reflect the content and visual patterns of the original YouTube videos. Third, six comprehension items for each passage were developed, forming the four testlets. Fourth, the items were trialed as part of the prototyping process. Table 1 summarizes the characteristics of the listening passages in terms of length, speech rate, lexical complexity, and content-related visuals.

Table 1

ALC Test: Characteristics of the Listening Passages

Title	Major details of a lecture	Length		Speech rate		Lexical Complexity		Content-related visuals		
		Word count	Input length	Words per minute	Syllables per second	Oxford 3000	Academic Word List	Pictures % (#)	Graphs % (#)	Total % (#)
Homeostasis	A lecture explaining the concept of homeostasis in human bodies. (1) The control of weight by our bodies. (2) The mechanism of homeostasis. (3) The importance of controlling body temperature.	734	03:58	185	4.29	91%	6%	20.6% (3)	40.0% (5)	60.6% (8)
Food Tax	A lecture about the effect of taxes on human behavior. (1) A tobacco tax precedent: Tax rates across the US. (2) Positive effects of tobacco taxes on health (3) Can food taxes affect eating behavior?	747	04:08	180	4.02	92%	3%	20.9% (4)	39.7% (4)	60.6% (8)
Compassion	A lecture about mechanisms that make people feel compassion towards others. (1) Compassion as a function of similarity. (2) Experiment providing evidence for (1).	779	03:57	197	4.23	91%	5%	17.1% (5)	42.5% (5)	59.6% (10)
Exoplanets	A lecture about detecting the motion of exoplanets. (1) The definition of a barycenter. (2) The light Doppler effect. (3) The radial velocity method.	863	04:16	202	4.22	91%	6%	18.6% (3)	40.7% (8)	59.3% (11)

Anchor test. The anchor test was developed to control for test-takers' proficiency. It consisted of two testlets (two YouTube lecture excerpts). Their length and complexity characteristics are summarized in Table 2.

Table 2

Features of the Anchor Listening Passages

Title	Major details of a lecture	Length		Speech rate		Lexical Complexity	
		Word count	Input length	Words per minute	Syllables per second	Oxford 3000	Academic Word List
Cyber-security	A lecture about the lack of trust online. (1) The essence of the problem. (2) Three types of cyber-attacks	693	04:15	163	3.71	90%	8%
Language	A lecture about how children learn their first language. (1) Basic facts about children's L1 acquisition. (2) Puzzles associated with children's L1 acquisition.	532	03:47	141	3.25	92%	5%

Reflecting the ALC test item structure, the anchor testlets contained three detail and three inference questions, the latter including one main idea question. Appendices A and B contain the revised items, answer key, scripts, and table of specification for the ALC and anchor tests respectively.

Test-takers' questionnaire. A questionnaire was developed to elicit test-takers' perceptions about the effects of videos on listening comprehension. A portion of this questionnaire was administered after each of the four testlets in the ALC test. The questionnaire had two versions – version 1, which came after the testlets in the audio-only version of the test, and version 2, which came after video-based versions of the testlets. Version 1 had questions about the effect of videos on listening difficulty, motivation, and authenticity, and whether or not videos should be used in academic tests. Version 2 also sought perceptions about viewing

behavior and helpfulness of content-rich videos for answering comprehension questions. Both versions ended with four items eliciting learners' demographic information, including first language, school affiliation, age, and gender. The questionnaire design is reflected in Table 3 below (also see Appendix C). The table lists the seven above-mentioned content areas, for each of which the following is provided: item type and scale, and number of items for each version.

Table 3

Test-takers' Questionnaire Design

#	Content area (construct)	Item type and scale	Number of items	
			Version 1	Version 2
1	Viewing behavior	5-point Likert equivalent	-	1
2	Video effects on listening difficulty	7-point semantic differential	1	1
3	Video effects on motivation	7-point semantic differential	1	1
4	Video effects on authenticity	7-point semantic differential	1	1
5	Video helpfulness for answering questions	4-point Likert	-	1
6	Use of videos in academic tests	4-point Likert	3	3
7	Demographic information	multiple-choice	2	2
		open-ended	2	2
	Overall		10	12

Procedures

The three test-takers' assessments, namely the ALC test, the anchor test, and the questionnaire, were combined in one academic listening assessment battery, which operated on an online testing platform run by Survey Gizmo. The battery started with the academic listening test, with section 1 of the test-takers' questionnaire appearing after each testlet, continued with the anchor test, and concluded with sections 2 and 3 of the test-takers' questionnaire.

The administration of the test-takers' assessment battery took place online, at each participants' convenience and preferred location. After listening to the instructions and electronically signing the informed consent, the test-taker was randomly assigned to either the audio-only or the video-based group. The testing software automatically ran directions, listening passages, videos, and the questionnaire as well as controlled the allocation of listening time. The

system did not allow for video replays and automatically sent the test-taker to comprehension questions upon the completion of each lecture. Test-takers were not allowed to pre-view comprehension questions. While listening, test-takers were free to take notes if needed. There were no time constraints on answering test or questionnaire items.

Results

Psychometric Properties of Tests

ALC test. ***Item fit statistics.*** The first step was to analyze the items' fit statistics. Table 4 below shows items' infit and outfit mean square values along with their *z*-scores. It can be seen that three items did not fit into the assessment construct (items 10, 22, and 23). Their infit and outfit values indicated statistically significant deviations from the overall item difficulty pattern.

Separation reliabilities. Item separation reliability index of 0.75 was approaching the desired value of 0.80. The strata value was 2.67, indicating that the test could consistently distinguish between more than two difficulty levels of items. The reliability associated with separating test-takers' abilities was 0.76 (> 0.70). The strata value was 2.68, indicating that the test could consistently distinguish between at least two ability levels of test-takers.

Item facility and discrimination. Three items displayed unusually high item facility indices (i.e., item 1 with 0.90; item 5 with 0.83; and item 10 with 0.97). This indicates that items 1, 5, and 10 were excessively easy for test-takers. Regarding item discrimination, items 5, 10, and 22 had values lower than 0.19. As evidenced by Table 4, point-biserial correlations of five items did not approach the cut-off value of 0.25 (items 4, 5, 14, 22, and 23). Because items 22 and 23 also had relatively large outfit values, they were candidates for major revisions.

Distractor analysis. According to Table 4, 10 out of 62 distractors were implausible, including distractors for items 1, 2, 3, 5, 6, 10, 12, and 13.

Table 4

ALC test: Analyses of Items' Psychometric Properties

Psychometric property	Expected value or range	Testlet / Items																							
		Homeostasis (1-6)							Food Tax (7-12)										Compassion (13-18)						Exoplanets (19-24)					
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
		D	D	I	I	D	M	I	D	D	I	D	M	I	D	D	D	I	M	D	I	D	D	I	M
Rasch analysis**																									
Difficulty logit		-1.81	0.14	0.83	0.73	-1.14	0.66	0.14	-0.04	1.17	**-3.06**	0.49	-0.23	-0.65	1.00	0.32	0.83	0.83	-0.88	-0.43	0.32	0.49	-0.04	0.49	-0.23
(b) Infit MS	0.50-1.50[1]	0.87	1.16	0.97	1.23	1.25	0.93	1.13	0.85	1.07	0.85	0.88	0.76	0.77	1.19	0.92	0.89	1.11	0.68	0.85	1.12	0.91	1.22	1.39	0.72
Infit *Z*	-2.00-2.00[2]	-0.10	0.90	-0.10	1.30	0.80	-0.40	0.70	-0.80	0.40	0.10	-0.70	-1.20	-0.90	1.10	-0.40	-0.60	0.70	-1.20	-0.60	0.70	-0.40	1.10	**2.20**	-1.50
Outfit MS	0.50-1.50[1]	0.58	1.20	0.94	1.41	1.18	0.90	1.11	0.74	0.99	**0.24**	0.80	0.62	0.67	1.42	0.84	0.96	1.26	0.46	0.76	1.10	0.86	**1.80**	**1.67**	0.59
Outfit *Z*	-2.00-2.00[2]	-0.30	0.80	-0.10	1.70	0.40	-0.40	0.50	-0.90	0.00	-0.20	-0.90	-1.20	-0.70	1.70	-0.60	-0.10	1.10	-1.20	-0.60	0.50	-0.60	**2.30**	**2.60**	-1.40
(c) Item separat.	> 0.80	reliability = 0.75; strata = 2.67; chi-square = 60.1, p < 0.01																							
(d) Person separat.	> 0.70	reliability = 0.76; strata = 2.68; chi-square = 99.2, p < 0.01																							
Classical analysis**																									
(e) Item facility	0.30-0.80[3]	**0.90**	0.62	0.48	0.48	**0.83**	0.52	0.62	0.66	0.41	**0.97**	0.55	0.69	0.76	0.45	0.59	0.48	0.48	0.79	0.72	0.59	0.55	0.66	0.55	0.69
(f) Item discrimin.	> 0.19[4]	0.29	0.43	0.57	0.43	**0.14**	0.57	0.43	0.71	0.43	**0.14**	0.86	0.71	0.71	0.29	0.71	0.71	0.43	0.86	0.57	0.29	0.71	**0.14**	0.29	0.71
(g) PB correlation	> 0.25[3]	0.42	0.27	0.47	**0.17**	**0.12**	0.50	0.30	0.56	0.38	0.37	0.54	0.64	0.58	**0.19**	0.49	0.49	0.29	0.69	0.53	0.31	0.49	**0.18**	**0.07**	0.67
Distractor analysis*																									
"A"	% students	**0.00**	0.14	*0.48*	*0.48*	**0.00**	0.03	0.17	0.07	0.07	**0.00**	*0.55*	0.10	*0.76*	0.24	0.14	0.07	0.28	*0.79*	0.10	*0.59*	0.07	*0.66*	0.03	0.07
"B"	selecting	*0.90*	0.24	**0.00**	0.10	*0.83*	**0.00**	0.14	*0.66*	0.03	**0.00**	0.14	0.21	0.21	*0.45*	*0.59*	0.24	0.10	0.03	0.10	0.24	*0.55*	0.03	0.24	*0.69*
"C"	each	0.07	*0.62*	0.07	0.10	0.17	0.45	0.07	0.21	0.48	*0.97*	0.07	**0.00**	**0.00**	0.07	0.14	*0.48*	0.14	0.07	0.07	0.14	0.21	0.10	*0.55*	0.14
"D"	option	0.03	**0.00**	0.45	0.31	**0.00**	*0.52*	*0.62*	0.07	*0.41*	0.03	0.24	*0.69*	0.03	0.24	0.14	0.21	*0.48*	0.10	*0.72*	0.03	0.17	0.21	0.17	0.10

** problematic parameters are **in bold**; * keys for items are in *italics;* MS = Mean Sqaure; *Z* = z-value for MS; PB = point biserial; D = detail item; I = inference item; M = main idea item;
[1]see Linacre (2012); [2]see McNamara (1996); [3]see Fulcher (2010); [4]see Ebel & Frisbie (1986);

Descriptive statistics. Table 5 contains descriptive statistics (i.e., sample size, mean, standard deviation, and confidence interval) for the ALC test by participants' location and item type for each of the delivery modes. We can see from the table that Facebook-based test-takers performed better (M = 17.78) than USA-based (M = 13.31) and Russia-based (M = 15.75) participants on average. Performances on global and local items were very similar in general and within each of the delivery modes. None of the confidence intervals overlap, indicating that differences in means did not reach statistical significance.

Table 5

Descriptive Statistics for the ALC Test

Delivery mode		Participants' location			Item type		Total (out of k=24)
		USA	Russia	Facebook	Local (out of k=12)	Global (out of k=12)	
Audio-only	*n*	10	1	2	13	13	13
	M	12.50	-	19.00	7.08	7.00	14.08
	SD	4.84	-	1.41	2.78	2.80	5.17
	CI	[9.68; 15.32]	-	[12.70; 25.30]	[5.65; 8.51]	[5.55; 8.45]	[10.95; 17.20]
Video-based	*n*	6	3	7	16	16	16
	M	14.67	14.33	17.43	7.69	8.13	15.81
	SD	1.21	6.03	4.72	2.27	2.34	4.05
	CI	[11.03; 18.30]	[9.19; 19.48]	[14.06; 20.79]	[6.40; 8.98]	[6.82; 9.43]	[13.65; 17.97]
Total	*n*	16	4	9	29	29	29
	M	13.31	15.75	17.78	7.41	7.62	15.04
	SD	3.96	5.68	4.18	2.49	2.60	4.59
	CI	[11.20; 15.42]	[6.71; 24.79]	[14.57; 21.00]	[6.47; 8.36]	[6.64; 8.60]	[13.29; 16.78]

Note: n = number of test-takers; k = number of items; CI = confidence interval

Anchor test. ***Item fit statistics.*** Table 6 below shows item infit and outfit mean square values for the anchor test, along with their z-values. The table follows the same format as for the ALC test. One item (item 2) had a high outfit value, indicating a possible misfit.

Separation reliabilities. Item separation reliability index was 0.82, falling in the preferred range. The strata value of 3.19 indicated that the items were at three distinct difficulty levels. The

reliability associated with separating test-takers' abilities was 0.57. The strata value of 1.86 indicated that the test could distinguish between more than one ability level of test-takers.

Item facility and discrimination. One item had unexpectedly high item facility index (i.e., item 6 with 0.97), indicating that the item was too easy for test-takers. Item 6 also had a low item discrimination index. Point-biserial correlations of two items did not approach the cut-off value of 0.25 (items 2 and 5).

Table 6

Anchor Test: Analyses of Items' Psychometric Properties

Psychometric property	Expected value or range	Testlet Items											
		Cybersecurity (1-6)						Language (7-12)					
		1	2	3	4	5	6	7	8	9	10	11	12
		I	D	D	D	I	M	D	D	I	D	M	I
Rasch analysis**													
Difficulty logit		-0.12	-0.95	-0.12	1.10	0.75	-3.11	-0.12	1.67	0.41	0.58	-0.51	0.41
(b) Infit MS	0.50-1.50[1]	0.78	1.14	0.81	1.15	1.26	0.95	1.06	1.15	0.83	0.92	1.00	0.97
Infit *Z*	-2.00-2.00[2]	-1.30	0.50	-1.10	0.80	1.40	0.20	0.30	0.70	-1.00	-0.40	0.00	-0.10
Outfit MS	0.50-1.50[1]	0.64	**2.14**	0.68	1.21	1.17	0.39	0.95	1.25	0.73	0.85	1.04	0.90
Outfit *Z*	-2.00-2.00[2]	-1.20	1.80	-1.00	0.80	0.70	0.00	0.00	0.80	-1.10	-0.50	0.20	-0.30
(c) Item separat.	> 0.80	reliability = 0.82; strata = 3.19; chi-square = 60.1, p < 0.01											
(d) Person separat.	> 0.70	reliability = 0.57; strata = 1.86; chi-square = 37.9, p < 0.01											
Classical analysis**													
Item facility	0.30-0.80[3]	0.66	0.79	0.66	0.41	0.48	**0.97**	0.66	0.31	0.55	0.52	0.72	0.55
Item discrimination	> 0.19[4]	0.63	0.25	0.75	0.50	0.38	**0.13**	0.63	0.25	0.75	0.63	0.25	0.38
PB correlation	> 0.25[3]	0.62	**0.17**	0.59	0.32	**0.24**	0.27	0.37	0.31	0.59	0.50	0.37	0.47
Distractor analysis*													
"A"	% students	0.14	*0.79*	**0.00**	0.34	0.03	*0.97*	0.14	0.45	0.24	0.07	0.10	0.24
"B"	selecting	0.10	0.10	*0.66*	*0.41*	0.14	**0.00**	0.14	0.10	0.10	0.17	*0.72*	0.03
"C"	each option	*0.66*	0.03	0.14	0.17	0.34	**0.00**	*0.66*	***0.31***	*0.55*	0.24	0.03	*0.55*
"D"		0.10	0.07	0.21	0.07	*0.48*	0.03	0.07	0.14	0.10	*0.52*	0.14	0.17

** problematic parameters are **in bold**; * keys for items are in *italics;* MS = Mean Sqaure; *Z* = z-value for MS; PB = point biserial; D = detail item; I = inference item; M = main idea item; [1]see Linacre (2012); [2]see McNamara (1996); [3]see Fulcher (2010); [4]see Ebel & Frisbie (1986).

Distractor analysis. As detailed in Table 6, three distractor malfunctioned. They were distractors for item 3 and item 6. In addition, the key for item 8 may not have functioned properly because it only attracted 31% of test-takers. Item 4 may have had a similar issue.

Descriptive statistics. The descriptive statistics for the anchor test is provided in Table 7 below. They are summarized by delivery mode, participants' location, and overall. The table shows that, on average, USA-based test-takers' listening proficiency was slightly lower

compared to Russia- and Facebook-based test-takers. However, this difference was not significant, as indicated by the overlapping confidence intervals. The means of the audio-only group (M = 7.23) and the video-based group (M = 7.31) were very similar, supporting the assumption that the two groups were equivalent in terms of listening proficiency. The mean for the whole body of test-takers was 7.28 out of 12. This value was used to determine a cut-off point in operationalizing listening proficiency. Test-takers scoring higher than 7.00 on the anchor test were assigned to the higher-proficiency group. Test-takers with a total anchor score of 7.00 and below were assigned to the lower-proficiency group.

Table 7

Descriptive Statistics for the Anchor Test

Delivery mode		Participants' location			Total (out of k=12)
		USA	Russia	Facebook	
Audio-only	*n*	10	1	2	13
	M	6.50	-	9.00	7.23
	SD	2.72	-	1.41	2.80
	CI	[5.01; 7.99]	-	[5.67; 12.33]	[5.54; 8.92]
Video-based	*n*	6	3	7	16
	M	6.67	7.33	7.86	7.31
	SD	1.51	1.53	2.41	1.92
	CI	[4.74; 8.59]	[4.61; 10.05]	[6.08; 9.64]	[6.29; 8.34]
Total	*n*	16	4	9	29
	M	6.56	8.25	8.11	7.28
	SD	2.28	2.22	2.21	2.31
	CI	[5.35; 7.78]	[4.72; 11.78]	[6.42; 9.81]	[6.40; 8.16]

Note: n = number of test-takers; k = number of items; CI = confidence interval

Test Performance

The effects of mode were investigated based on the mode measurement report generated by the Rasch analysis. The logit difficulty values for the audio-only and video-based groups were not far apart, M = -0.60 (SE=0.14) and M = -0.85 (SE=0.12) respectively. As indicated by the

chi-square statistics, chi-square (df = 1) of 0.03, separation index of 0.95, and $p = 0.17 > 0.05$, the two delivery modes were not significantly different in terms of their difficulty for test takers.

The relationship between the effect of mode and proficiency was investigated by running a Rasch bias/interaction analysis. Table 8 shows the following information for each of the proficiency levels: listening difficulty when a content-rich video is absent (audio-only target measure) or present (video-based target measure), target contrast (difference between the target measures), and the significance of the contrast. We can see that the presence of content-rich videos did not affect listening comprehension for either lower-level or higher-level test-takers.

Table 8

Bias/Interaction Analysis: Listening Difficulty, Delivery Mode, and Proficiency

Proficiency	Target measure (S.E.)		Target contrast	Joint S.E.	*t*	Welch *d.f.*	*p*
	Audio-only	Video-based					
Lower (n = 13)	0.00 (0.17)	0.00 (0.18)	0.00	0.30	0.00	272	0.99
Higher (n = 16)	0.00 (0.24)	0.00 (0.16)	0.00	0.23	0.00	381	0.99

Note: S.E. = Standard Error

Regarding effects of mode at the item level, there was one significant interaction between individual item difficulty and delivery mode. Item 2 was found to be significantly easier in the video-based mode (-0.80) than in the audio-only mode (1.13). Although the effect of delivery mode was noticeably high for two more items (items 4 and 21), it did not reach statistical significance. The output for the item-mode interaction analyses is given in Appendix D.

Questionnaire Responses

Descriptive statistics for the questionnaire data are summarized in Table 9. It shows subsample sizes, means, and standard deviations for perceptions of listening difficulty, motivation, authenticity, and use of videos in listening tests. Regarding listening difficulty, no effects of delivery mode and proficiency were found. Test-takers generally perceived the lectures

to be quite difficult (5.32 out of 7.00), regardless of mode and proficiency. Regarding motivation, no effects of delivery mode and proficiency were found. Test-takers generally perceived the lectures to be moderately motivating (4.45 out of 7.00), regardless of mode and proficiency. Regarding listening authenticity, no effects of delivery mode and proficiency were found. Test-takers generally perceived the lectures to be authentic (5.30 out of 7.00), regardless of mode and proficiency.

Table 9

Descriptive Statistics for RQs 2.1-2.5

Dependent variable (RQ)		Deliver mode / Proficiency						Total
		Audio-only mode			Video-based mode			
		Lower	Higher	Total	Lower	Higher	Total	
Difficulty (RQ 2.1)	*n*	5	8	13	8	8	16	29
	M (out of 7.00)	5.00	6.09	5.67	5.00	5.09	5.05	5.32
	SD	0.40	0.63	0.77	1.44	1.18	1.27	1.10
	CI	[4.03; 5.97]	[5.33; 6.88]	[5.21; 6.13]	[4.24; 5.77]	[4.33; 5.86]	[4.40; 5.72]	[4.91; 5.75]
Motivation (RQ 2.2)	*n*	5	8	13	8	8	16	29
	M (out of 7.00)	4.50	4.53	4.51	3.93	4.87	4.41	4.45
	SD	0.25	1.18	0.91	1.04	0.93	1.07	0.99
	CI	[3.60; 5.40]	[3.82; 5.24]	[3.95; 5.09]	[3.23; 4.65]	[4.17; 5.58]	[3.91; 4.91]	[4.08; 4.83]
Authenticity (RQ 2.3)	*n*	5	8	13	8	8	16	29
	M (out of 7.00)	5.75	5.28	5.46	5.00	5.34	5.17	5.30
	SD	1.03	0.91	0.95	1.00	1.26	1.30	1.17
	CI	[4.66; 6.84]	[4.42; 6.14]	[4.82; 6.21]	[4.13; 5.86]	[4.48; 6.20]	[4.56; 5.78]	[4.87; 5.74]
Use of videos in tests (RQ 2.5)	*n*	5	8	13	8	8	16	29
	M (out of 4.00)	2.60	3.00	2.85	3.25	3.50	3.38	3.14
	SD	0.72	0.47	0.59	0.66	0.50	0.58	0.63
	CI	[2.06; 3.14]	[2.57; 3.42]	[2.49; 3.20]	[2.82; 3.67]	[3.07; 3.92]	[3.06; 3.69]	[2.90; 3.38]

Regarding the use of videos in listening tests, an effect of mode approached statistical significance, F (1, 25) = 6.78, p = 0.015, at the adjusted p-value of 0.05/4 = 0.0125. The partial eta squared effect size index was 0.21, indicating a large effect (Cohen, 1988). This result shows that test-takers in the video-based group had more favorable opinions about including videos in listening tests (M = 3.14; SD = 0.63) than test-takers in the audio-only group (M = 2.85; SD = 0.59).

Discussion

The results of this study generated little evidence to support the inclusion of content-rich videos in L2 academic listening tests in terms of both test-takers' test performance and perceptions. However, they point to the possibility of more promising findings for the future dissertation study. Although not to the point of statistical significance, the audio-only mode in this study tended to be noticeably harder than the video-based mode at the test level. Similarly, effects for some individual items were close to approaching statistical significance.

The present pilot study may have had limited statistical power due to the small sample size. Having a considerably larger sample size may enable the future dissertation study to detect significant effects of content-rich videos foreshadowed by this pilot study. Therefore, drawing theoretical implications based solely on the results of this study may be a premature act. Instead, limitations of this study should be addressed in the future dissertation project, aiming to bring more evidence about the role of content-rich visuals in the construct.

Relevance to PIE and Second Language Learning

This study's inconclusive findings have not provided strong support for including content-rich video-based visuals in English as a second language listening assessment instruments used in the PIE. However, they have not challenged the recommendation to use such

visuals either, since the lack of evidence does not constitute counter-evidence. Moreover, the use of content-rich videos in listening tests would be a better representation of authentic university listening contexts in the US (Lynch, 2011). Awaiting more conclusive evidence from the forthcoming dissertation study, it is suggested that listening sections of the PIE placement and exit tests include audio-only testlets for the time being and that lower-stakes classroom assessments continue utilizing videos in their designs.

Although not to the point of statistical significance, test-takers in this pilot study seem to have found the video-based version of the test slightly easier than the audio-only version (5.05 vs 5.67 out of 7.00 respectively; interpretation: the lower, the easier). They found the video-based version to be somewhat motivating (4.41 out of 7.00) and authentic (5.17 out of 7.00). They also were in favor of using content-rich videos in second language listening tests (3.38 out of 4.00). These perceptions support the common practice of utilizing visual information to facilitate learners' listening comprehension (Chapelle & Jamieson, 2008). They also corroborate that it is worth using visuals in second language listening classrooms for sustaining students' motivation and improving listening authenticity.

References

Batty, A. (2015). A comparison of video- and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning. *Language Testing*, *32*, 3-20.

Chapelle, C. & Jamieson, J. (2008). *Tips for teaching with CALL: Practical approaches to computer-assisted language learning.* White Plains, NY: Pearson Education.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Coniam, D. (2001). The use of audio and video comprehension as an assessment instrument in the certification of English language teachers: A case study. *System, 29,* 1-14

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

Ebel, R., & Frisbie, D. (1986). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice-Hall.

ETS Guidelines for Fair Test and Communication. (2015). Princeton, NJ: Educational Testing Service.

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

Fulcher, G. (2010). *Practical language testing.* London: Hodder Education.

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

Kang, T., Gutierrez Arvizu, M. N., Chaipuapae, P., & Lesnov, R. (2016). Reviews of academic English listening tests for non-native speakers. *International Journal of Listening.* Published online on June 27.

Lesnov, R. (2017). Using videos in ESL listening achievement tests: Effects on difficulty. *Eurasian Journal of Applied Linguistics, 3,* 67-91.

Lievrouw, L. A. (2011). *Alternative and activist new media*. Cambridge, UK: Polity.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7,* 328.

Linacre, J. M. (2012). Many-Facet Rasch measurement: Facets tutorial 1/2012. Retrieved from http://www.winsteps.com/a/ftutorial2.pdf

Linacre, J. M. (2017) Facets computer program for many-facet Rasch measurement, version 3.80.0. Beaverton, Oregon: Winsteps.com

Lynch, T. (2011). Academic listening in the 21st century: Reviewing a decade of research. *Journal of English for Academic Purposes, 10,* 79-88.

McNamara, T. (1996). *Measuring second language performance.* Harlow, Essex, UK: Addison Wesley Longman Ltd.

Ockey, G. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing, 24,* 517–537.

Rost, M. (2016). *Teaching and researching: Listening* (3rd ed.). New York, NY: Routledge.

Royce, T. D. (2007). Multimodal communicative competence in second language contexts. In T. D. Royce & W. Bowcher (Eds.), *New directions in the analysis of multimodal discourse* (pp. 361–90). New York, NY: Erlbaum.

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning, 55,* 661-699.

SurveyGizmo (2017). Professional survey solution. Retrieved from https://www.surveygizmo.com/

Suvorov, R. (2009). Context visuals in L2 listening tests: The effects of photographs and video vs. audio-only format. In C. Chapelle, H. 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, 21,* 1-21.

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

Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly, 5,* 218–243.

Wagner, E. (2010a). Test-takers' interaction with an L2 video listening test. *System, 38,* 280-291.

Wagner, E. (2010b). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing, 27,* 493-513.

Appendix A

Revised Listening Test (scripts, items, table of specification)

Testlet 1. Homeostasis (Questions 1-6)

I want to talk about some concepts in physiology that are really important for this course in biomedical engineering. I want you to try to imagine a table that has characteristics of an average person – an adult male, 30 years old, average height, average weight, average surface area, ah average temperature, just a lot of average characteristics of an average person. And let's just take a look at one of these, let's look at weight. So weight is something that is actually a very carefully controlled parameter for a person. Ahm we take in a lot of food, we take in a lot of drink ah but we don't really gain a lot of weight, our weight stays pretty stable. And if you try to lose weight - you're too young to try to lose weight too much, but as you get older your metabolism changes, you realize how hard it is to lose weight, and we know it's hard because we spend so much energy talking about it. Now ah weight is pretty carefully controlled and your body does it on its own, you don't have to think about it. Now ah also, temperature. Temperature is something that is within a narrow range, stays pretty constant. You go from inside to outside, you go into a hot room, your temperature doesn't change that much, it stays within this range of 36.5 to 37.5 degrees. And it's so stable, it's so important that it's stable that when it changes just a little bit, we know that something is wrong. You measure your temperature, it goes up and down. And if it's a little bit up, we know something's wrong – you have a fever. We know it because it's so stable.

So, you could go through a lot of these parameters and think about them in the same way that these things are really very highly controlled. And this process of control to maintain a constant environment within our bodies, whether it's mass or chemical composition, or temperature, is called homeostasis. And your body has very elaborate mechanisms for maintaining this state of homeostasis. Ah in spite of the fact that we take in a lot of chemicals and ah in different ways, and we have to do that to stay alive, but we have mechanisms to control the process very well. Now homeostasis is enabled by both complex and simple control mechanisms. And we can describe them in ways that are actually probably pretty similar to control mechanisms mechanisms that you're already familiar with. So, let's take for example the

thermostat in your dorm. Maybe this is a bad example, maybe you don’t have control over your thermostat or maybe your thermostat doesn’t work very well. But just imagine a perfect thermostat. No matter what the temperature it is outside, it maintains the constant temperature inside your room. Now this perfect thermostat ah works through a control mechanism that’s called negative feedback. And so it works like this. You have a thermostat that’s measuring the temperature and it’s sending signals to a heater somewhere. And when the temperature level drops below a certain level, then it sends a signal to turn on, the heater turns on, and it’s just heating, it’s just heating until it receives the second signal. So when does it receive the second signal? When the temperature goes above the certain level, then the second signal is sent, and it turn off. So the heater’s on, it’s just heating, heating, heating and it gets the signal to turn off. It says 'oh we've gone too high', and it shuts down. So our bodies have these same mechanisms like that, they mainly use this principle of negative feedback to control the parameters that are important for life within certain ranges.

So why is temperature, for example, so important to keep at 37 degrees? Well it’s because that's the temperature at which many of the molecules in our bodies operate most efficiently. So enzymes are the best example of this. Enzymes are molecules that catalyze chemical reactions and our bodies are basically networks of chemical reactions, and enzymes operate most effectively at 37 degrees Celsius. So when we're off from that temperature then enzymes don't work properly any more, and then the chemical reactions don’t run as well as they should. And there are other examples as well, but that's why it's important.

1. According to the speaker, which statement about weight is true?
   (A) It is easy for older people to lose weight.
   (B) Our body carefully controls its weight.
   (C) Eating food makes our weight unstable.
   (D) We do not normally talk about weight.

2. The normal body temperature range is _____ degrees Celsius.
   (A) 36.5-37.0
   (B) 36.0-37.5
   (C) 36.5-37.5
   (D) 36.6-37.6

3. We can infer that thermostats are _____.
   (A) quite familiar to students
   (B) not relevant to the lecture
   (C) not helpful for understanding homeostasis
   (D) in a perfect condition in college dorm rooms

4. Our body will most likely send the second control signal when _____.
   (A) we have a fever
   (B) we are cold
   (C) our temperature is normal
   (D) our temperature is negative

5. Temperature control is important because it _____.
   (A) reduces body's energy use
   (B) helps molecules work effectively
   (C) increases the number of enzymes
   (D) stops negative feedback signals

6. This lecture is mainly about _____.
   (A) how our body keeps its weight constant
   (B) which body parameters are most typical
   (C) why body temperature is important
   (D) how our body controls its environment

Testlet 2. Food Tax (Questions 7-12)

So today let's return to that idea of unhealthy foods that we've been talking about and think about how it interacts with taxes. Now, the most radical change of all when it comes to proposed ahm policies and food politics has to do with the idea of taxes. Taxing foods and will it actually be viable to put a tax on certain foods to help improve public health? And the rationale for doing something like this with taxes has to do with what we've been talking about in class. Those un ah healthy foods just simply cost more to make and to to provide than unhealthy foods do. As a result, those unhealthy foods are more affordable for the poor. We could you use a tax policy to to discourage that affordability of unhealthy foods and we could take that money and use it as a subsidy for the foods we want, fresh produces, fruits and vegetables. And this is a topic that we've been thinking about for years. There is a precedent for this in the arena of tobacco. Now you know there're different taxes on packs of cigarettes that vary state by state by state around the country. And there's a huge difference between the biggest taxes of about two dollars and fifty cents a pack in New Jersey and Rhode Island versus the smallest tax of ah seven cents a pack in South Carolina. And the research in this areas has shown for years that taxes are the single most effective way to curb smoking. Other things do matter but taxes are the most effective. Those are current data that I just presented. But I also have data that are about a year older. If you compare the four states with the highest tax and the four with the lowest. So that's Montana, Michigan, New Jersey, and Rhode Island ah more than two dollars a pack, and Mississippi, Missouri, and the Carolinas, less than 20 cents a pack. You can see that difference is huge.

Of course, you can probably guess what I'm going to tell you next, which is the rate of smoking in the state with higher versus lower tobacco taxes. There's not a perfect relationship because in Michigan we can see quite a high level of smoking despite having one of the highest taxes in the country. But in general, ahm we can see that states with higher taxes have remarkably lower rates of smoking; the states with low cigarette taxes do have many more smokers. So taxes do matter, they do affect behavior. And we wonder if there could be something equivalent in the area of food. To show you just how much of a difference ah these taxes can make, let's look at California. In California, there is a heavy tax on cigarettes, with the money specifically earmarked to go to anti-tobacco programs, and that doesn't happen in every

state. This ahm started in 1988 with a twenty-five cents per pack increase in taxes on cigarettes. And it generated about ninety million dollars a year, all going to these anti-tobacco ahm campaigns. And you might have seen those Truth Campaign ads that painted tobacco executives ahm really negatively. By 1999, this resulted in a twenty-seven percent decrease in smoking and nineteen percent decrease in deaths due to lung cancer, about 10 percent better than the rest of the country. Now that's a powerful finding: a nineteen percent reduction in deaths, just from a tax. Could you imagine trying to do that through education? You wouldn’t be able to do it. It would cost way too much, and nobody would come up with that kinda money. Or you can just write a law that changes tax.

Now those are staggering findings, this these changes in behavior just from a tax. And it didn't come from small steps. It didn't come from advice like 'go get a dog and walk it.' That came from changing the law and placing a tax on the thing we want to discourage. And if a tax is done in this way, it potentially has many beneficial effects. So these different suggestions for food taxes have come up in countries, in England, in Ireland, in Australia. And it probably will happen at some point. So, the question I leave you with today is what role should government play in this whole process? And, is it taking a constructive role right now? That's for you to think about.

7. We can infer that taxing fast food will _____.
   (A) weaken public health
   (B) raise people's objections
   (C) make people wealthier
   (D) increase fresh food sales

8. Cigarette tax rates are _____ across states in the US.
   (A) relatively similar
   (B) largely different
   (C) mostly high
   (D) mostly low

9. In California, smoking-related deaths _____.
   (A) increased by 27%
   (B) increased by 19%
   (C) decreased by 27%
   (D) decreased by 19%

10. The purpose of discussing tobacco in this lecture was to show _____.
   (A) the effects of smoking on health
   (B) the benefits of non-smoking
   (C) an effective tax policy
   (D) a good way to pay income taxes

11. Based on the listening, which statement is NOT true?
   (A) Educating about tobacco is better than taxing it.
   (B) Tobacco taxes may fund anti-tobacco programs.
   (C) Some countries have considered a food tax.
   (D) Adding a new tax requires changing the law.

12. This lecture is mainly about _____.
   (A) tax rates and educational achievement
   (B) tobacco tax rates across the US
   (C) tobacco tax and smoking in California
   (D) tax rates and human behavior

Testlet 3. Compassion (Questions 13-18)

Compassion is a really interesting thing to study because the world is full of more people who need help than we can possibly help. Right, if we try to feel compassion for everyone, it will be impossible and overwhelming. And so the question is: Out of all the people in the world who need help, how do we decide who it is most beneficial to help, ah who is most worthy of compassion? And what I wanna suggest to you is that one way that we go about deciding whether or not to help someone or whether or not to show compassion to them is based on a simple analysis: Do we see ourselves in them? And so I wanna suggest that one way compassion works is based on that simple metric, and that metric is similarity. The idea is: The more similar someone is to me, the more likely I am to feel compassion for them, even if they're suffering the same tragedy as another individual. And what this suggests is that distress is really in the eye of the beholder. How much compassion I feel for someone isn't a function of what's befallen them, it's a function of their links to me. Now if I said to you, on a battle field an American soldier comes upon a wounded member of Taliban and a wounded American soldier, and they feel more compassion towards the wounded American soldier, that might not be surprising to you. Those groups were in conflict for a long time. But what I wanna suggest is that this bias is so deeply embedded in the mind that we can see it even with the subtlest of cues.

And so the cues I really wanna look at, stripping it down to bare bones, is simple motor synchrony, right, moving in time together. If you move your body in time together, it's a marker that right now, in this moment, two individuals are one. Their purposes are joined, and their goals are joined. And those are the individuals who long-term are most likely going to help me. So, how do we do this? We bring individuals into a lab. We sit them down at a table, and they put on earphones. They think they're in the music perception study. And their goal is simple: Tap your hands to the tones you hear. The only difference is: Sometimes they tap their hands in unison, and sometimes the tones are random, so they tap in a completely asynchronous way. They don't talk, they don't do anything else. What happens next is that you see the partner who you were tapping with, engaging in another study that you're observing, in which they are being cheated by another subject and being stuck with this onerous, tedious task. And then simply what we do is we ask them if they wanna help that person or not. We don't ask them as experimenters because that might add some extra pressure. Ah the end of the experiment, the computer simply

says to them: There's more work to be done; if for some reason you'd like to help somebody else, please find one of the experimenters and let them know.

And what we've found, I have to admit to you, was rather astounding to me. The simple act of tapping your hands in time makes people feel more similar. Now they couldn't tell us why they were more similar, they would create stories about how they were similar. They didn't even talk to the other person, and yet they still felt similar. And what that similarity did is it gave the long-term mechanisms of the mind greater power to increase the compassion that we were gonna feel. And so the amount of compassion they felt was also influenced by whether or not they tapped in time with that person – if they did, they felt more compassion. But remember, in each case the person is victimized in the same way and cheated in exactly the same way. But how much compassion we feel for them is really a function of how similar we feel to them. Moreover, if you look at the decisions to help, there's a really large difference, right? 17 out of 35 people decided to help the person with whom they tapped their hands in time. Only 6 out of 34 decided to do that in cases where there was less similarity. And if you look at the time they spent helping, it's even more dramatic, right? If I feel similar to you, I helped you for much longer than I did if I felt that you and I were not similar.

13. According to the speaker, we will probably feel more compassion for a person who _____.
    (A) has a lot in common with us
    (B) got in serious trouble
    (C) is in conflict with someone
    (D) is similar to a famous celebrity

14. In the experiment, what happened after the tone tapping?
    (A) The tones were changed.
    (B) One participant was cheated.
    (C) Participants were seated.
    (D) Experimenters helped participants.

15. If people tapped in time with a partner, they _____ their partners.
    (A) felt less similar to
    (B) more often helped
    (C) felt less compassion for
    (D) more often looked at

16. Which statement is NOT true?
    (A) Moving together is a sign of having one goal.
    (B) Participants were cheated in the same way.
    (C) Participants knew why they felt similar.
    (D) Talking was not allowed in the experiment.

17. Two partners would probably feel less similar if _____.
    (A) one of them was not cheated
    (B) both of them were cheated
    (C) their tasks were not tedious
    (D) they heard tones at different times

18. The passage is mainly about _____ compassion.
    (A) what makes people feel
    (B) how to do research on
    (C) how to have people appreciate
    (D) why it is important to study

Testlet 4. Exoplanets (Question 19-24)

This lecture focuses on one of the main methods for detecting exoplanets - the radial velocity method. As we'll discuss, the radial velocity method uses the motion, or the wobble, of a star to indicate the presence of a planet. As I alluded to when we talked about planetary motions, planets don't exactly orbit the Sun. We probably learned that the Sun's at the center and the planets orbit around the Sun. Well, that's not exactly true. Planets don't orbit the Sun. They orbit the barycenter, which is kind of a balance point. It's a balance point in mass between all the planets and the Sun. And that's hard to explain, when we consider all eight of the planets in our solar system. So let's just consider the biggest planet, Jupiter, and let's see how that goes with the Sun. So the Sun and the Jupiter play kind of cosmic balancing act. It's as if they're on a seesaw, if you will, and they have to balance each other. So if you put the Sun and Jupiter on a seesaw, Jupiter will be mhhs much farther away. It's 1,000 times less massive than the Sun. And the Sun will actually sit very close to the center, but not perfectly at the center. That balance point of the seesaw is what is called the barycenter. These two are balancing each other So as Jupiter goes around in its orbit, the Sun also has to balance out Jupiter's mass and go round in its orbit. Turns out the barycenter of the Sun with respect to Jupiter is actually outside the surface of the Sun. And therefore, as Jupiter is going around in its orbit, the Sun, too, is going around in its orbit. So we can actually see, if you were looking at the solar system from above you'd actually see as Jupiter is going around, the Sun too is orbiting. It's making a much smaller orbit, but it too is making an orbit.

So this wobble, or this effect of a star having to orbit its own barycenter, is a telltale sign of planets around that star. But how can we detect them? There's some tricks that we can do for seeing the star's motion as it comes towards us and away from us. One of those tricks is the Doppler effect. The Doppler effect is an effect that most of you probably know because you've encountered it with sound. In fact, if you're walking down the street or you've heard a police car or an ambulance come towards you or going away from you, ah you hear, as that car comes towards you, the sound waves are compressed, and the pitch gets higher. Kind of goes -- beeeep. And as the car goes away from you, the sound waves are elongated, and the pitch goes down. Ah you hear kind of ahh baaooo. And of course, the engine or the siren of the police vehicle hasn't changed its pitch at all. It's just your perception. The waves have actually been compressed as they come to your ear. So many of us have heard that with sound. But the same principle applies

to light. In fact, as an object comes towards you, the waves are compressed. The wavelength gets smaller, gets bluer. And as an object goes away from you, the waves are elongated, or get redder, as they get to longer wavelengths. And the faster an object moves, either towards you or away from you, the larger that shift is. So this is the light version of a Doppler effect.

But what can we use to study that? We know that now, if we can measure this light Doppler shift, if we can measure a star as it wobbles towards us, it should get a little bit bluer. And as it goes away from you, it should get a little bit redder. And in fact, that motion towards us and away from us is actually what's called radial velocity. That's why this technique is called radial velocity method. And we define radial velocity, positive radial velocity, as the motion away from us. So as the light gets a little bit redder, we call that positive radial velocity. As it gets bluer when it comes towards us, we call that negative radial velocity. So if we see that star go towards us, then away from us, then towards us, then away from us, we'll be detecting that star wobbling. And that's, again, the telltale sign that star has a planet in orbit. So what we can do is monitor these stars, take spectra, or distribution of colors coming from stars, and actually watch as these colors themselves wobble back and forth. We can actually observe the spectral features doing that, and the degree of the spectral shift tells us about the speed of that star’s wobble. So the very first detection of an extrasolar planet around a star like our Sun was done in 1991 using this radial velocity method. It was done around the star 51 PEG. And so we call the exoplanet 51 PEG B, for the first exoplanet around that system.

19. A barycenter is a/an _____.
    (A) planet's core or midpoint
    (B) orbit or path of planets
    (C) planet detection method
    (D) balance point of planets

20. We can infer that a planet with less mass _____.
    (A) sits far from its barycenter
    (B) has a smaller orbit
    (C) has its barycenter inside
    (D) completes its orbit faster

21. According to the speaker, which statement is NOT true?
    (A) The Sun goes around in its orbit.
    (B) Car sirens change their pitch.
    (C) Planets do not orbit the Sun.
    (D) The Doppler Effect applies to light.

22. If an object comes toward us, it has _____.
    (A) shorter waves
    (B) lower pitch
    (C) redder colors
    (D) positive radial velocity

23. A radial velocity method is probably NOT _____.
    (A) effective
    (B) reliable
    (C) new
    (D) popular

24. This passage is mainly about detecting the _____.
    (A) barycenter of a planet
    (B) motion of a planet
    (C) planets' sound waves
    (D) orbits of Jupiter and the Sun

Academic Listening Test Answer Key

Testlet 1. Homeostasis	Testlet 2. Food Tax	Testlet 3. Compassion	Testlet 4. Exoplanets
1. B	7. D	13. A	19. D
2. C	8. B	14. B	20. A
3. A	9. D	15. B	21. B
4. A	10. C	16. C	22. A
5. B	11. A	17. D	23. C
6. D	12. D	18. A	24. B

Academic Listening Test – Table of Specification

Listening Testlets	**Sub-constructs**			**# items**	**%**
	Main Ideas	Details	Inferences		
Testlet 1. Homeostasis	1	3	2	6	25%
a) 03:58 b) 1 speaker c) Physical science d) moderately fast e) video-based version: 20.6% pictures, 40.0% graphs	6	1, 2, 5	3, 4		
Testlet 2. Food Tax	1	3	2	6	25%
a) 04:08 b) 1 speaker c) Social science d) moderately fast e) video-based version: 20.9% pictures, 39.7% graphs	12	8, 9, 11	7, 10		
Testlet 3. Compassion	1	3	2	6	25%
a) 03:57 b) 1 speaker c) Social science d) moderately fast e) video-based version: 17.1% pictures, 42.5% graphs	18	14, 15, 16	13, 17		
Testlet 4. Exoplanets	1	3	2	6	25%
a) 04:16 b) 1 speaker c) Physical Science d) moderately fast e) video-based version: 18.6% pictures, 40.7% graphs	24	19, 21, 22	20, 23		
Items per sub-construct	4	12	8	24	100%
Points per item	1	1	1		
Points per sub-construct	4	12	8	Raw Pts: 24	

Appendix B

Anchor Listening Test (scripts, items, table of specifications)

Anchor testlet 1. Cybersecurity (Questions 1-6).

Testlet 1. Cybersecurity

The reality is when you are online there is no way to be sure that the person you think you are communicating with or the website you're … are going to is really that person or that website. There is no 100% certainty with the basic architecture of the network. And so … we've had to think about how do you manage this problem. The problem is maybe best encapsulated by a New York … New Yorker magazine cartoon – I think it goes back fifteen to twenty years – it's back in the days of big clunky ah PCs on the desk. And there's a drawing of a PC and there's two dogs talking to each other. And one dog says to the other: "On the Internet nobody knows you're a dog." And in many ways that sums up the problem. So with this lack of trust and with the ability of people to mascaraed as others and use it as a way to gain entry to our own networks, what we've seen again and again is the capability that people have if they're bad actors to corrupt information, to steal information, to deny access or introduce latency or delay in the transmission of information, to destroy and overwhelm networks and of course to steal all kinds of inf information for financial gain.

If I would group … these types of consequences, I would say in the main they fall into three main categories. The one that's maybe the most long-standing set of security challenges and the one that we still read about the most and probably the one that touches us personally the most is the use of the network to steal financial information for the purposes of committing fraud – identity information, credit card information, access to bank accounts. Ah as you've read there've been literally millions of dollars stolen in this way. In the last couple of years, for example, there was one organized criminal effort to gain access to ATMs. What they did was they hacked into a couple of firms overseas, they were managing debit cards and ATM withdrawal cards, and they had the withdrawal limits on those cards removed. Then, on a single day, ahhp individuals working as part of this conspiracy were sent out to ATM machines all over the world to withdraw all the money from the machines. Because the withdrawal limits were gone, they could take every cash bit of cash that was in those machines. And on a single day

before it was shut down tens of millions of dollars were stolen. So, that's a classic example of the fact that because the Internet is now where the money is, it's like [??]. To paraphrase [??], you don't have to rob banks any more by going in with a gun – you just rob it through the ATM or the credit card.

A second area of things that we have seen are denial of service attacks. Ah these aren't maybe the most sophisticated attacks, they don't ultimately destroy ah systems or networks, they don't kill people, but they interfere with the ability to get access to your … perhaps your bank or some other facility that you need to communicate with. And they create an enormous burden and dragging expense for enterprises.

But the third and most consequential from a national security standpoint, the third type of category of attacks we worry about are attacks that actually could be corruptive or destructive. Imagine what would happen if ah malevolent actors penetrated into banks and were able able over a period of time, in a very subtle way, to change bank records. If you didn't have a back-up for transactions, you might have a crisis of confidence in banks something like what we saw in 2008 when we had our financial crisis. You could have destruction of critical infrastructure but unlike in Sony which destroyed business enterprises, tools and and and information technology architecture, you could actually have attacks on critical infrastructure that deals with transportation – the train that I came up with, the airplane I'm flying, maybe power. And that could actually cause loss of life as well as significant economical property damage.

1. The cartoon about two dogs was discussed to illustrate the _____.

(A) solution for the lack of trust online

(B) disadvantages of early computers

(C) problem of trusting online resources

(C) types of online communication

2. According to the speaker, which cyber-crime will touch people personally the most?

(A) Stealing credit card information.

(B) Robbing a bank's ATM machine.

(C) Destroying a government office.

(D) Denying access to a bank website.

3. To take all the money from ATMs, the criminals _____.

(A) shut down power in the banks

(B) removed limits from credit cards

(C) robbed banks with a gun

(D) broke open the ATM machines

4. For national security, the most serious category of cyber-crimes is _____.

(A) stealing financial information

(B) corruptive or destructive attacks

(C) denial-of-service attacks

(D) robbing banks with a gun

5. A cyber-attack of *the third type* would most likely target a _____.

(A) family-owned business

(B) person's Facebook account

(C) government official's email

(D) country's energy system

6. This lecture is mainly about the _____.

(A) security problems of online systems

(B) secure access to bank computers

(C) problems of insecure ATM machines

(D) lack of trust among modern people

Anchor testlet 2. Language (Questions 7-12)

Hello there and welcome back to the introduction to English linguistics. In this video I'd like to talk about language acquisition - how do children learn a first language. And to start out with, let me give you a few basic facts about language learning. First of all, there is no genetic predisposition for learning any one particular language. A baby born to English-speaking parents will of course learn English but the same baby, if it grows up around people talking in Finnish or in Mandarin or in Sinhalese or in Welsh, will acquire any of those languages with the same speed and ease. All human languages are equally easy to acquire as a first language and not only that - children can acquire two or more first languages with ease. Yeah. Ahh having two or more first languages – that's called bilingualism or multilingualism and it has been shown that there are strong cognitive advantages to being bilingual. Bilinguals they have two language systems in their mind and in order to use one, they have to inhibit the other, so they have to concentrate on one thing and defocus another thing. And you can imagine that this helps in a whole lot of other cognitive tasks – you concentrate on one thing and selectively ignore the other thing.

Right. More facts about language acquisition. Ahhm I said that the process seems to be effortless - very easy and very rapid so that all essential parts of language – the grammatical structures, pronunciations, all of that, is in place by age five to six, so there kids talk pretty much like adults. Now of course they don't talk completely like adults – they don't have the same capabilities that adults have. Think of telling a good joke or understanding irony. There kids catch up over the years, but in terms of grammatical rules, pronunciations, knowledge of different words – the basics really are in place by age five to six. All this happens without formal instruction. You don't have to tell kids: this is right, this is wrong, this is what the rules are. They figure that out by themselves and, interestingly, the outcome is almost always the same. Everybody learns how to talk and ah even though there may be some people that talk really really well, that are super eloquent, that know how to talk in public, ahhm … well this is a skill that you have to learn as an adult. Yeah ahh everybody learns instinctively how to talk well enough to hold a conversation.

Right. Now there are certain puzzles associated with language acquisition. For one thing, kids say things that they've never heard before. How do they do that? Kids get things right without being corrected. How is that? How do they figure that out? And then they master grammar by age 5 but they don’t master things that are equally complex or comparable to

language like mathematics, differential equations. Mmm they have trouble doing that at age 15 and yet at age five they chatter away, yeah, they have trouble tying their shoelaces but they use relative clauses – that seems to be remarkable. Now linguists try to explain these puzzles with theories of language acquisition.

7. According to the speaker, which statement is *true*?

(A) Children learn some languages faster than others.

(B) Learning two languages may be difficult for children.

(C) Children learn any human language equally easily.

(D) Some children are slower at learning languages.

8. According to the speaker, bilingual children _____.

(A) use two language systems at the same time

(B) may have problems with concentrating

(C) focus on one of the language systems

(D) select a language that they know better

9. Children will most likely _____ by age 6.

(A) need instruction to speak well

(B) be able to tell many good jokes

(C) be able to hold a conversation

(D) know how to speak in public

10. Linguists hope to explain how children can _____.

(A) say what they heard before

(B) solve mathematical problems

(C) understand language theories

(D) speak right without correction

11. The lecture is mainly about _____ children.

(A) formal language instruction for

(B) learning a first language by

(C) the facts about public speaking by

(D) learning a foreign language by

12. The teacher will most likely talk next about _____.

(A) the lives of famous linguists

(B) what languages children should learn

(C) how children learn a first language

(D) how to teach children a first language

Anchor Test Answer Key

Anchor testlet 1. Cybersecurity	Anchor testlet 2. Language
1. C	7. C
2. A	8. C
3. B	9. C
4. B	10. D
5. D	11. B
6. A	12. C

Anchor Test – Table of Specification

Listening Testlets	**Sub-constructs**			**# items**	**%**
	Main Ideas	Details	Inferences		
Testlet 1. Cybersecurity	1	3	2	6	50%
a) 04:15 b) 1 speaker c) Social science d) moderate speed	6	2, 3, 4	1, 5		
Testlet 2. Language	1	3	2	6	50%
a) 03:47 b) 1 speaker c) Social science d) slow to moderate speed	11	7, 8, 10	9, 12		
Items per subconstruct	2	6	4	12	100%
Points per item	1	1	1		
Points per subconstruct	2	6	4	Raw Pts: 12	

Appendix C

Test-takers' Questionnaire

Test-takers' Questionnaire. Version 1

Section 1

1. How interesting was this lecture?

1	2	3	4	5	6	7
very boring						very interesting

2. How difficult was this lecture?

1	2	3	4	5	6	7
very easy						very difficult

3. How realistic was this lecture?

1	2	3	4	5	6	7
not realistic						very realistic

Section 2

4. Academic listening tests should have videos.

( ) Strongly Disagree
( ) Disagree
( ) Agree
( ) Strongly Agree

5. Academic listening tests should be audio-only.

( ) Strongly Disagree
( ) Disagree
( ) Agree
( ) Strongly Agree

6. With videos, academic listening tests are more valid.

( ) Strongly Disagree
( ) Disagree
( ) Agree
( ) Strongly Agree

Section 3

7. What is your first language? _____

8. Which school are you in?

( ) Universidad de Sonora, Mexico
( ) Program in Intensive English, Northern Arizona University, USA
( ) English Language Center, Rochester Institute of Technology, USA
( ) EnglishDom, the Russian Federation
( ) Skyeng, the Russian Federation
( ) Other school
( ) I'm an independent English learner

9. How old are you? _____

10. What is your gender?

( ) Male
( ) Female
( ) Other

Test-takers' Questionnaire. Version 2

Section 1

A. How much of the video did you watch?

( ) I did **not** watch
( ) **Little**
( ) About **half** of the video
( ) **Most** of the video
( ) **All** of the video

1. How interesting was this lecture?

1	2	3	4	5	6	7
very boring						very interesting

2. How difficult was this lecture?

1	2	3	4	5	6	7
very easy						very difficult

3. How realistic was this lecture?

1	2	3	4	5	6	7
not realistic						very realistic

B. Do you agree that you were able to answer some questions because you saw pictures and graphs?

( ) Strongly Disagree

( ) Disagree

( ) Agree

( ) Strongly Agree

Section 2

4. Academic listening tests should have videos.

( ) Strongly Disagree

( ) Disagree

( ) Agree

( ) Strongly Agree

5. Academic listening tests should be audio-only.

( ) Strongly Disagree

( ) Disagree

( ) Agree

( ) Strongly Agree

6. With videos, academic listening tests are more valid.

( ) Strongly Disagree

( ) Disagree

( ) Agree

( ) Strongly Agree

Section 3

7. What is your first language? _____

8. Which school are you in?

( ) Universidad de Sonora, Mexico

( ) Program in Intensive English, Northern Arizona University, USA

( ) English Language Center, Rochester Institute of Technology, USA

( ) EnglishDom, the Russian Federation

( ) Skyeng, the Russian Federation

( ) Other school

( ) I'm an independent English learner

9. How old are you? _____

10. What is your gender?

( ) Male

( ) Female

( ) Other

Test-takers' Questionnaire. Table of Specifications

Version	Content area (Construct)							Total
	Viewing behavior	Video effects on			Video helpfulness for answering questions	Use of videos in academic listening tests	Demographics	
		listening difficulty	motivation	authenticity				
Version 1		1 (#2)	1 (#1)	1 (#3)		3 (#4-6)	4 (#7-10)	10
		10%	10%	10%		30%	40%	100%
Version 2	1 (#A)	1 (#2)	1 (#1)	1 (#3)	1 (#B)	3 (#4-6)	4 (#7-10)	12
	8.3%	8.3%	8.3%	8.3%	8.3%	25%	33.3%	100%

Appendix D

ALC Test: Rasch Item-Mode Interaction Analysis

Audio vs Video 9/4/2017 11:14:54 AM
Table 14.1.2.2 Bias/Interaction Pairwise Report (arranged by N).
Bias/Interaction: 2. Mode, 4. Items

Target Nu	It	Target Measr	S.E.	Obs-Exp Average	N	Mode	Target Measr	S.E.	Obs-Exp Average	N	Mode	Target Contrast	Joint S.E.	t	Welch d.f.	Prob.
1	1	-2.43	1.08	.05	1	audio-only	-1.32	.79	-.04	2	video	-1.11	1.34	-.83	25	.4150
2	**2**	**1.13**	**.64**	**-.19**	**1**	**audio-only**	**-.80**	**.68**	**.16**	**2**	**video**	**1.92**	**.93**	**2.06**	**26**	**.0491**
3	3	1.55	.67	-.13	1	audio-only	.30	.56	.11	2	video	1.25	.87	1.44	25	.1620
4	4	-.07	.64	.17	1	audio-only	1.51	.56	-.14	2	video	-1.58	.85	-1.85	26	.0751
5	5	-2.43	1.08	.13	1	audio-only	-.38	.62	-.11	2	video	-2.05	1.24	-1.65	23	.1128
6	6	1.13	.64	-.09	1	audio-only	.30	.56	.07	2	video	.83	.85	.98	26	.3385
7	7	-.07	.64	.04	1	audio-only	.30	.56	-.03	2	video	-.37	.85	-.43	26	.6673
8	8	.33	.63	-.07	1	audio-only	-.38	.62	.06	2	video	.71	.88	.81	26	.4262
9	9	.33	.63	.16	1	audio-only	1.84	.59	-.13	2	video	-1.51	.86	-1.76	26	.0908
10	10	-2.43	1.08	-.03	1	audio-only	-3.06	1.57	.03	2	video	.62	1.91	.33	26	.7457
11	11	.72	.63	-.05	1	audio-only	.30	.56	.04	2	video	.43	.84	.51	26	.6157
12	12	.33	.63	-.11	1	audio-only	-.80	.68	.09	2	video	1.13	.92	1.22	26	.2338
13	13	-.07	.64	-.10	1	audio-only	-1.32	.79	.08	2	video	1.25	1.02	1.23	26	.2293
14	14	.33	.63	.13	1	audio-only	1.51	.56	-.10	2	video	-1.18	.85	-1.40	26	.1740
15	15	.72	.63	-.08	1	audio-only	-.03	.58	.07	2	video	.75	.86	.88	26	.3882
16	16	.33	.63	.10	1	audio-only	1.20	.55	-.08	2	video	-.87	.84	-1.04	26	.3072
17	17	.33	.63	.10	1	audio-only	1.20	.55	-.08	2	video	-.87	.84	-1.04	26	.3072
18	18	-.50	.67	-.06	1	audio-only	-1.32	.79	.05	2	video	.82	1.03	.80	26	.4333
19	19	-.07	.64	-.07	1	audio-only	-.80	.68	.05	2	video	.72	.93	.78	26	.4449
20	20	1.13	.64	-.16	1	audio-only	-.38	.62	.13	2	video	1.51	.89	1.70	26	.1015
21	21	-.50	.67	.18	1	audio-only	1.20	.55	-.15	2	video	-1.70	.87	-1.96	25	.0608
22	22	.72	.63	-.15	1	audio-only	-.80	.68	.12	2	video	1.52	.92	1.65	26	.1116
23	23	-.07	.64	.11	1	audio-only	.90	.55	-.09	2	video	-.97	.84	-1.16	26	.2581
24	24	-.98	.72	.12	1	audio-only	.30	.56	-.10	2	video	-1.28	.91	-1.40	25	.1734