The Role of Content-Rich Videos in the L2 Academic Listening Assessment Construct: Dissertation 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, AZ 86004, USA.

Contact: roman.lesnov@nau.edu

Abstract

This study compared the difficulty of two modes of delivering a second language academic listening test, namely the audio-only mode and the video-based mode. In the video-based version of the test, content-rich videos were used, defined as videos containing relevant content-related visual aids, such as graphs and pictures, for 60% of the video length. Visual aids in the videos overlapped with the speaker's non-verbal cues. In accord with the theory, the video-based version was expected to generate higher scores and more favorable opinions of test-takers regarding test difficulty, authenticity, motivation, and whether or not videos should be used in high-stakes listening tests. Overall, 143 English as a second or a foreign language learners participated in the study, 16 of which were students in the Program in Intensive English (PIE) at Northern Arizona University. The results showed that the video-based version was generally easier than the audio-only version if test questions were cued by videos. In contrast, the test was harder with the videos than without if test questions were not cued by videos, especially for lower-level students. These findings are discussed in terms of their implications for the PIE.

The Role of Content-Rich Videos in the L2 Academic Listening Assessment Construct: Dissertation Study

Background

The field of L2 assessment has been actively debating the issue of including visual processing as part of a listening construct. Many studies showed that the inclusion of visual information could make an L2 listening construct more authentic (e.g., Ockey, 2007; Wagner, 2008, 2010a, 2010b). This view is in line with contemporary conceptualizations of the listening skill, which largely acknowledge the role of visual information for successful listening (e.g., Field, 2008; Flowerdew and Miller, 2005). In addition to recognizing the importance of non-verbal cues (e.g., eye expression), scholars nowadays have started gravitating towards considering visuals that are content-related (e.g., visual aids in a lecture) as part of listening (e.g., Rost, 2016). Despite this, L2 listening tests developers continue to favor "pure" definitions of listening, which view visuals as a source of construct-irrelevant variance (e.g., Buck, 2001; Kang, Gutierrez Arvizu, Chaipuapae, & Lesnov, 2016; Lado, 1964). Taking into account the widespread availability of video technology, which is thought to be most capable of reflecting the visual reality of L2 contexts and no longer technologically problematic, the unwillingness to build video-based listening assessments remains a mystery (Gruba, 2014; Li, 2013).

For *academic* L2 listening assessments, videos would be expected to include content-related visuals along with non-verbal cues and situational visuals (Lynch, 2011). Evidence validating such a video-based construct could come from several sources, two of which are (1) test-takers' performance on a test and (2) stakeholders' perceptions of the test (Bachman and Palmer, 2010; Chapelle, Enright, & Jamieson, 2008; Gruba, 2014). This study aimed to obtain

such evidence by comparing (1) L2 students' performance on an audio-only versus a video-based version of an academic listening test, the latter containing content-rich videos, and (2) L2 students' and teachers' perceptions about helpfulness of content-rich videos for comprehension and the use of such videos in listening tests.

Research Questions

The study was governed by the following two major research questions.

1. Do content-rich videos affect L2 academic listening comprehension difficulty? An affirmative answer for this question was expected, signaling a difference in the nature of the L2 academic listening construct and supporting the inclusion of content-related visuals into the assessment construct.
2. Do test-takers' perceptions lend support for using content-rich videos in the L2 academic listening assessment construct? L2 learners were expected to favor the use of content-rich videos in listening tests, further advancing the argument for including content-rich visuals into the construct of L2 academic listening assessment (e.g., Gruba, 2014).

Methods

Methodology is what made this study innovative. First, the study employed content-rich videos, or videos that displayed predominantly content-related visuals in the form of graphs and images overlapping with non-verbal cues from the lecturer and with a small amount of text. Graphs and images in the video were semantically congruous with the respective audio chunks, with pictures fulfilling an illustrating function and graphs serving as illustrators and/or organizers. Second, this study controlled for the degree to which content-rich videos were helpful for answering individual test items, coined as item video-dependence. Although previous studies have attempted to investigate the effects of content videos (e.g., Suvorov, 2015), they

overlooked the extent of videos' saturation with content-related visuals and item video-dependence.

Participants

143 adult English learners from two USA-based and four foreign English schools took online instruments. Sixteen of the 143 recruited participants were students in the Program in Intensive English (PIE) at Northern Arizona University (NAU). They were native speakers of Chinese (n = 15) and Korean (n = 1). On average, they were about 22 years old. Thirteen of them were females and three males.

Instruments

To answer the first research question, two instruments were developed – the academic listening comprehension test (ALC; 4 recorded lectures, 24 multiple-choice items) and the anchor listening test (2 authentic lectures, 12 multiple-choice items). Four lectures for the ALC test were video-recorded using the scripts and visual patterns from authentic university lecture excerpts found on YouTube. The video-based version of the ALC test used content-rich videos. Each content-rich video displayed pictures, graphs, and the lecturer for about 20%, 40%, and 100% of the overall video's length respectively. Cronbach's alphas of 0.61 (audio-only group) and 0.75 (video-based group) indicated adequate reliability. Test performance of the audio-only group and the video-based group was compared, using the combination of Rasch analysis and classical test theory analyses, such ANOVA. The anchor test (Cronbach's alpha = 0.65) was used to estimate test-takers' listening proficiency.

Item video-dependence had two values for each item, namely video-dependent or video-independent. Twelve English learners and teachers took a video-dependence questionnaire and the muted version of the video-based ALC test. The results on these two measures informed the

researcher's decision to group the ALC test items into either video-dependent (k = 14) or video-independent (k = 10).

To answer the second research question, test-takers' opinions were investigated. Test-takers' perceptions about listening difficulty, motivation, and authenticity, as well as the use of videos in tests were obtained using a questionnaire. Data from the questionnaire were compared between the audio-only and the video-based groups using ANOVAs.

Administration

The three test-takers' instruments, namely the ALC test, the anchor test, and the test-takers' questionnaire, were combined in one assessment battery, which operated on an online testing platform run by Survey Gizmo. The system randomly assigned each test-taker to either the audio-only or the video-based version of the battery. The administration of the battery took place online (approximately a 45-minute time commitment), at each participants' convenience and preferred location. The ALC test, the anchor test, and the test-takers' questionnaire are found in Appendices A, B, and C respectively.

Analysis

To answer research question 1, Rasch interactions were run to see if academic listening comprehension difficulty was affected by delivery mode, listening proficiency, and item video-dependence. Delivery mode was a categorical independent variable with two values, namely audio-only and video-based. Proficiency level was determined using the anchor test. It had two dichotomous values, lower (person's logit ≤ 0.00) and higher (person's logit > 0.00). Item video-dependence was a yes-or-no property of each individual item. Accordingly, it had two values, namely video-dependent and video-independent, reflecting the video-dependence grouping decisions for each item (see Instruments).

Note that, to connect disjoint data subsets for Rasch analysis, 24 out of the 143 test-takers had to take the test the second time in the opposite mode after a 3-week break. Therefore, Rasch analysis was based on 167 responses.

To answer research question 2, four ANOVAs were run to see if delivery mode and listening proficiency had an effect on test-takers' perceptions about difficulty of, motivation towards, and authenticity of the ALC test, as well as about whether content-rich videos should be used in high-stakes L2 academic listening tests. Each perception area was measured by a number of questions on a 1-6 scale each, as showed in Table 1 below.

Table 1

Operationalizations of Difficulty, Motivation, Authenticity, and Video Use Perceptions

Perception area	Scale	Meaning of the scale	# questions	Overall range
Difficulty	1-6	1-very easy 6-very difficult	4	1-24
Motivation	1-6	1-very boring 6-very interesting	4	1-24
Authenticity	1-6	1-not realistic 6-very realistic	4	1-24
Video use	1-6	Questions similar to "Should videos be used in listening tests?" 1-strongly disagree 6-strongly agree	3	1-18

Results

The video-based mode was easier for test-takers than the audio-only mode. The logit difficulty values (interpretation: the lower, the easier) for the video-based and audio-only groups were M = -0.07 (SE=0.05) and M = 0.07 (SE=0.05) respectively. As indicated by the chi-square statistics, χ^2 (1) = 3.80, separation index of 1.68, and p = .05, this difference approached statistical significance. This pattern is also supported by the total ALC raw scores (interpretation:

the lower, the harder): *M* = 13.10, *SD* = 3.67, CI: [12.03; 14.18] for the video-based mode and *M* = 11.53 out of 24, *SD* = 3.67, CI: [10.69; 12.38] for the audio-only mode.

Video-dependent items' difficulty for both lower and higher proficiency levels was not significantly different for the video-based versus the audio-only mode (see Table 2). Video-independent items were significantly harder for lower-level test-takers in the video-based mode, *t* (787) = -3.25, *p* = .001, but not for higher-level participants.

Table 2

Rasch Interaction: Video-Dependence and Proficiency for Mode

Items	Proficiency	Target measure (S.E.)		Target contrast	Joint S.E.	*t*	*df*	*p*
		Audio-only	Video-based					
Video-dependent	Lower (*n* = 79)	-0.12 (0.09)	-0.30 (0.09)	0.17	0.13	1.31	1103	.190
	Higher (*n* = 88)	-0.07 (0.09)	-0.31 (0.10)	0.24	0.13	1.79	1186	.073
Video-independent	Lower (*n* = 79)	-0.05 (0.11)	0.47 (0.11)	-0.52	0.16	-3.25	787	.001*
	Higher (*n* = 88)	0.13 (0.10)	0.21 (0.11)	-0.08	0.15	-0.55	852	.586

Note: S.E. = Standard Error; *significance at α = .05; *df = Welch degrees of freedom; N* = 167

Table 3 below shows the results of the four ANOVAs on test-takers' perceptions of test difficulty, listening motivation and authenticity, and video use. For each perception area, it also gives means and standard deviations by mode. We can see that test-takers found the video-based mode to be easier than the audio-only mode (interpretation of difficulty means: the lower, the easier). Although perceptions on motivation, authenticity, and video use were not affected by mode, they were high, showing that, on average, test-takers considered the video-based version (along with the audio-only version) motivating and authentic. Test-takers in both the video-based and the audio-only groups seemed to be in favor of including content-rich videos in high-stakes L2 academic listening tests.

Table 3

ANOVAs on Test-Takers' Perceptions

Items	Mean (*SD*)		*F*	*df*	*p*
	Audio-only	Video-based			
Difficulty	16.16 (4.22)	14.59 (4.27)	1.31	(1; 139)	.026*
Motivation	16.33 (4.32)	16.71 (3.46)	0.49	(1; 139)	.485
Authenticity	19.04 (3.25)	18.26 (3.38)	1.65	(1; 139)	.201
Video use	12.56 (2.40)	12.59 (2.48)	0.01	(1; 139)	.910

Note: S.E. = Standard Error; *significance at α = .05; *df = Welch degrees of freedom; N* = 143

Relevance to Second Language Learning and PIE

The results of this study bear two main implications. First, both test-takers' performance on and perceptions of the test showed that content-rich videos facilitate academic listening comprehension on video-dependent items. Test-takers were also supportive of using content-rich videos in standardized academic listening tests. Being in line with theoretical expectations, these findings support the inclusion of content-rich visuals in L2 academic listening assessments. If test developers opt to overlook these findings and keep excluding visual aids that would likely be present in authentic situations, such as PowerPoint slides with graphs and pictures during a lecture, test-takers' listening comprehension may be undermined. To put it differently, visual-free assessment constructs would yield scores that are less predictive of test-takers' behavior in real-life university lectures. It is suggested that the PIE placement and exist assessments include video-based testlets along with audio-only testlets, with the former featuring visual aids reflective of authentic contexts. Similar suggestions can be made for lower-stakes classroom listening assessments, such as achievement tests.

This recommendation should not be challenged by the finding of lower-level test-takers scoring lower on video-independent items in the presence of videos than with audio-only. Due to their limited linguistic capacity, lower-level learners were expected to be overwhelmed by visual information irrelevant for comprehension questions (recall that video-independent items did not rely on visuals). Therefore, video-independent items can still be included in video-based tests along with video-dependent

items. This would be a better reflection of target language use situations. After all, while many assessments that lecturers use would be related to PowerPoint presentations or handouts, some comprehension questions or tasks in authentic university lectures will likely not be covered by visuals.

The second recommendation is for PIE listening teachers to heighten students' awareness of the potential that content-rich visuals offer for listening comprehension. It would be useful for English learners to know that processing visual information is part of the academic listening skill, and thus, can be utilized to facilitate listening comprehension. Pedagogical techniques could be developed for the PIE listening and speaking classes that would train students to process the auditory and the visual channels simultaneously. One approach to this would be to use online listening resources that incorporate "strategies in their instructional designs" and video-enhanced computer-assisted learning activities that "explicitly teach listening strategies" (Chapelle & Jamieson, 2008, p. 145).

References

Bachman, L. F., & Palmer, A.S. (2010). *Language Assessment in practice.* Oxford: Oxford University Press.

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

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

Chapelle, C., Enright, M., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language.* New York, NY: Routledge.

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

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

Gruba, P. (2014). New media in language assessment. In J. Kunnan (Ed.), *The Companion to language assessment* (pp. 995-1012). London, UK: John Wiley.

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. Retrieved from https://www.tandfonline.com/doi/full/10.1080/10904018.2016.1185210

Lado, R. (1961). *Language testing: The construction and use of foreign language tests.* London, UK: Longman.

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

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

Ockey, G. J. (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.

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. (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

ALC Test: Scripts, Items, Specifications

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) Young people now have serious weight issues.
   (B) We do not have to help our body control its weight.
   (C) Eating food makes our weight quite unstable.
   (D) We do not normally talk much 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) 37.5-38.5

3. We can infer that thermostats are _____.
   (A) quite familiar to students
   (B) weakly related 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 drops fast

5. Temperature control is important because it _____.
   (A) slows harmful chemical reactions
   (B) helps molecules work effectively
   (C) increases the number of enzymes
   (D) manages the body's feedback

6. This lecture is mainly about _____.
   (A) how our body keeps its weight constant
   (B) which body parameters are most important
   (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. We can infer that the teacher used the older data about tax rates to _____.
   (A) show that he is an expert
   (B) present additional evidence
   (C) compare historical data trends
   (D) indicate an ineffective policy

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 anti-tobacco programs
   (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) is our soulmate or close relative
    (B) got in serious trouble or difficulty
    (C) suffers from the war's effects
    (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 *away from* us, it has _____.
    (A) longer waves
    (B) higher pitch
    (C) bluer colors
    (D) negative radial velocity

23. What would be a sign that a planet is orbiting?
    (A) blue colors
    (B) red colors
    (C) both blue and red colors
    (D) no colors and shorter waves

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, 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. Audio-Only Version.

Section 1

1. How interesting was this lecture?

1	2	3	4	5	6
very boring					very interesting

2. How difficult was this lecture?

1	2	3	4	5	6
very easy					very difficult

3. How realistic was this lecture?

1	2	3	4	5	6
not realistic					very realistic

Section 2

4. Academic listening tests should have videos.

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

5. Academic listening tests should be audio-only.

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

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

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

Section 3

7. What is your first language? _____

8. Which school are you in?

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

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

( ) Universidad de Sonora, Mexico

( ) Zaoksky Christian Institute of Arts and Sciences, Russia

( ) White Rabbit, Russia

( ) EnglishDom, the Russian Federation

( ) Skyeng, the Russian Federation

( ) Other

9. How old are you? _____

10. What is your gender?

( ) Male

( ) Female

( ) Other

Test-takers' Questionnaire. Video-Based Version

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
not realistic					very realistic

2. How difficult was this lecture?

1	2	3	4	5	6
not realistic					very realistic

3. How realistic was this lecture?

1	2	3	4	5	6
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	( ) Somewhat Disagree	( ) Somewhat Agree	( ) Agree	() Strongly Agree

Section 2

4. Academic listening tests should have videos.

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

5. Academic listening tests should be audio-only.

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

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

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

Section 3

7. What is your first language? _____

8. Which school are you in?

( ) Program in Intensive English, Northern Arizona University, USA
( ) English Language Center, Rochester Institute of Technology, USA
( ) Universidad de Sonora, Mexico
( ) Zaoksky Christian Institute of Arts and Sciences, Russia
( ) White Rabbit, Russia
( ) EnglishDom, the Russian Federation
( ) Skyeng, the Russian Federation
( ) Other

9. How old are you? _____

10. What is your gender?

( ) Male
( ) Female
( ) Other

Table of Specifications for Test-takers' Questionnaire

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				
Audio-only version		1 (#2)	1 (#1)	1 (#3)		3 (#4-6)	4 (#7-10)	10
		10%	10%	10%		30%	40%	100%
Video-based version	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%