Using Videos in L2 Listening Achievement Tests: Rasch Analysis on Difficulty Effects

Roman Lesnov

Northern Arizona University

Abstract

Even though modern video technology has been used in a variety of educational contexts, second language (L2) listening testing remains one of the areas that has made little use of video support. This study investigated how audio-only and video-enhanced delivery formats of listening passages compared in terms of difficulty for English as a second language (ESL) students. Also, interactions between difficulty, video type (context vs. content), and individual items were investigated. The study utilized listening achievement test data from 44 high-intermediate ESL students who were enrolled in an American intensive English program in Fall 2015. The Rasch procedure was used to analyze the data in Facets. No effect of video on the overall test difficulty was found. However, the findings suggested that the more content clues were in videos, the less difficult listening comprehension was. Furthermore, the testlet enhanced with a video without content clues was significantly harder for participants than its audio-only analog. Finally, individual items were both advantaged and disadvantaged by the presence of content-related videos. The findings are discussed in terms of their practical significance for ESL teachers and their theoretical implications for L2 listening assessment.

Keywords: audio, content, context, listening test, video

Using Videos in L2 Listening Achievement Tests: Rasch Analysis on Difficulty Effects

The place of visuals in the listening construct has been a topic of debate among researchers. Thus far, videos have been included in L2 listening tests with the purpose of investigating their effect on comprehension, which would serve as evidence in the argument for or against the inclusion of videos in tests, and, on a larger scale, visuals in the listening construct. To uncover this effect, researchers compared L2 test takers' performance on audio-only versus video-enhanced listening passages with mixed results. Some studies found positive effects of videos on L2 listening comprehension (e.g., Londe, 2009; Wagner, 2010b), while others yielded a negative (e.g., Suvorov, 2009; Wagner, 2010a) or no effect (e.g., Coniam, 2001; Cubilo & Winke, 2001). This discrepancy may be partly attributable to the failure on the researchers' part to control for video type (i.e., context versus content videos), with only one study (i.e., Suvorov, 2014) having taken video type into account.

Surprisingly few studies (i.e., Batty, 2015; Wagner, 2010b) have attempted to investigate how individual items functioned under the video condition. This is the area where item response theory (IRT) can offer better insights than classical test theory (CTT) because it compares the difficulty of individual items and the performance of individual test takers without relying on overall raw scores. Except for Batty (2015), there was no research that would employ IRT to compare the effect of videos on overall L2 listening test performance as well as the comprehension of each item. The present work sought to fill this gap by employing Multi-Faceted Rasch Analysis (MFRA) to investigate the interactions between format (i.e., video-

enhanced versus audio-only), video type (i.e., context versus content), and performance on individual listening comprehension items.

Research questions

Three research questions (RQ) were developed to accomplish the above-stated goal.

1. How does listening achievement test difficulty depend on delivery format (audio-only vs. video-enhanced)?
2. Is there interaction between testlet difficulty, delivery format, and video type?
3. Is there interaction between delivery format and individual item difficulties? If so, does it relate to item type (main idea, detail, inference)?

Methods

Participants

The data were obtained from 44 international students who studied in an intensive English program (IEP) at an American university in Fall 2015. The program had six levels of proficiency, Level 1 being the lowest and Level 6 the highest. All the 44 students were enrolled in Level 5, which consisted of four classes (5A, 5B, 5C, and 5D). In terms of overall language proficiency, Level 5 students were approximately in the range of 57-69 on the Test of English as a Foreign Language (TOEFL iBT) scale.

Instrument

A listening test (LT) – Level 5 LT – was developed to monitor students' progress in their Listening and Speaking courses. To create Level 5 LT, four video passages were found on the Internet. Features of the passages are listed in Table 1 including the main topic as well as the amount of content visual information and context non-verbal cues in each passage. As indicated in the table, one of the video passages was mostly content ("Ethics," 63%) while the others were

more of the context type (i.e., containing less than 50% of content clues). It should be mentioned that content information in Ethics as well as in MTask testlets could be used by test-takers to answer comprehension questions as it might contain or allude to the right answer via text, picture, or scheme.

Table 1

Features of Level 5 LT Video Passages

Passage	Main topic	Non-verbal cues	Content clues
Passage **Ethics:** "Business Ethics"	An introductory lecture (monologue) on business ethics delivered by a professor of economics.	Upper half of the body: face, gestures	63% of the time: organized text
Passage **Fraud:** "A Fraud Triangle"	An interview with an expert about the causes of fraud in a workplace and the ways to deal with it. Mostly monologic.	Upper half of the body: face, gestures	0% of the time: No content clues
Passage **Zombies:** "Economics and Zombies"	A presentation (monologue) on the topic of the relationship between economics and ethics.	Upper half of the body: face, gestures	0% of the time: No content clues
Passage **MTask:** "Multitasking and Switchtasking"	A presentation (monologue) about the ineffectiveness of switchtasking and managing one's time properly.	Entire body: face, gestures	29% of the time: pictures, text, schemes

Five to six multiple-choice questions of different types (i.e., main idea, detail, and inference questions) were developed for each passage. Eventually, the test contained 4 listening testlets. Each testlet contained either audio-only or video-enhanced passages; audio-only passages being a product of excluding video channel from the original video passages. The sequences of passages and formats as well as other features of the four versions of Level 5 LT are presented in Table 2.

Table 2

Features of Level 5 LT Versions

Passage	Length	Speed	Number of Hearings	Items answerable from a video input	Answer time, per item	Order and format of testlets in versions			
Ethics	04:31	medium	1	5 out of 5	25 sec	1A	4A	3V	2V
Fraud	04:52	medium to fast	1	0 out of 6	25 sec	2V	3V	4A	1A
Zombies	04:13	medium	1	0 out of 5	25 sec	3A	2A	1V	4V
MTask	03:05	fast	1	2 out of 6	25 sec	4V	1V	2A	3A

Note: Ethics = "Business Ethics"; Fraud = "Fraud Triangle"; Zombies = "Economics and Zombies"; MTask = "Multitasking and Switchtasking"; A = audio-only; V = video-enhanced;

Data analysis

Order effects. To compare raw scores among the four versions of the LT, a one-way ANOVA was run. No significant differences between mean scores was found, F (3, 40) = 1.18; p = 0.33, which showed the equivalence of the four versions.

Effects of Videos. The effects of videos on listening difficulty in general as well as in relation to video type and individual items were analyzed with the MFRA procedure in Facets (Version 3.71.4). Four facets were entered into the Rasch model – examinee, format, video type, and item. Table 3 shows the following information for each of the facets: range of values, labels, and other specifications.

Table 3

Rasch Model Specifications

Facet	Range of values	Labels	Other specifications
Test taker	1-44	Individual test takers' IDs	Positive facet Group-anchored at 0*
Format	1-2	1 = audio-only 2 = video-enhanced	Non-centered ("floating" facet)
Video type	00, 29, and 63	00 = 0% of content clues 29 = 29% of content clues 63 = 63% of content clues	Dummy facet
Items	1-22	Individual items' numbers	

* Note: group-anchoring was used to eliminate the disconnectedness of the data

Results

Before answering the research questions, the infit statistics, the item-variable map, and separation reliabilities for items and test takers were checked. They confirmed that the instrument was well suited for the sample and the data met the assumption of unidimentionality.

Overall Format Effects

Format measurement report (see Table 4) was used to determine test takers' comprehension across the two formats. As can be seen in Table 4, the logit difficulty values for audio and video were not far apart, M = -1.00 (SE=0.12) and M = -1.10 (SE=0.12) respectively. As indicated by chi-square statistics, separation = 0.00, chi-square (df = 1) = 0.03, p = 0.57 > 0.05, formats were not significantly different in terms of their difficulty for test takers.

Table 4

Format Measurement Report

	Difficulty logit	Model error	Infit MS	Infit Z	Correlation
Audio	-1.00	0.12	0.95	-0.90	0.50
Video	-1.10	0.12	1.06	1.00	0.48
M	-1.05	0.12	1.00	0.10	0.49
SD	0.05	0.00	0.05	1.00	0.01

Note: Reliability = 0.00; Separation Index = 0.00; Fixed Chi-square = 0.03 (df = 1, p = 0.57)

Interaction between Format and Video Type

Pairwise comparisons of the difficulties of each video type under each of the formats revealed significant differences. According to Table 5, the presence of videos with no content clues resulted in the increased difficulty of the listening message; t (407) = -2.72, p = 0.007 < 0.05. In contrast, the presence of videos containing 29% content clues did not make a difference for difficulty; t (226) = 0.10, p = 0.922 > 0.05. Finally, a significantly large positive t-value, t (178) = 3.56, p = 0.001 < 0.05, indicated that videos with 63% of content clues made the listening message significantly easier for test takers.

Table 5

Bias Analysis for Format vs. Video Type: Pairwise Comparisons

Video Type	Target measure (S.E.)		Target contrast	Joint S.E.	*t*	Welch *d.f.*	*p*
	Audio	Video					
0% content clues	-0.35 (0.18)	0.33 (0.17)	-0.68	0.25	-2.72	407	0.007*
29% content clues	0.02 (0. 24)	-0.01 (0.21)	0.03	0.32	0.10	226	0.922
63% content clues	0.49 (0.21)	-0.76 (0.28)	1.25	0.35	3.56	178	0.001*

Note: S.E. = Standard Error; * = significant at the 0.05 alpha level

Interaction between Format and Individual Items

The bias analysis for format against every individual item revealed significant interactions for 5 items, as illustrated in Table 6 below. Three of the five items (i.e. items 1, 3, and 20) had positive contrasts between their difficulties under audio vs. video formats, indicating that they were significantly easier when administered in video-enhanced testlets. Specifically, items 1 and 3 belonged to the "Ethics" testlet whereas item 20 was in "Mtask." Items 1 was a main idea item while items 3 and 20 were details. In contrast, items 7 (inference) and 17 (main idea) were negatively biased by the presence of video, meaning that they were significantly more difficult when testlets "Fraud" (item 7) and "Mtask" (item 17) were accompanied by videos.

Table 6

Significant Format-Item Interactions

Item	Type	Target measure (S.E.)		Target contrast	Joint S.E.	*t*	Welch *d.f.*	*p*
		Audio	Video					
Item # 1	MI	1.50 (0.45)	-0.01 (0.56)	1.52	0.72	2.11	34	0.042
Item # 3	DET	0.90 (0.45)	-1.62 (0.78)	2.52	0.90	2.80	30	0.009
Item # 20	DET	0.48 (0.54)	-1.76 (0.77)	2.25	0.94	2.40	38	0.022
Item # 7	INF	-0.43 (0.58)	1.60 (0.45)	-2.03	0.73	-2.76	34	0.009
Item # 17	MI	-1.72 (0.78)	0.39 (0.46)	-2.10	0.91	-2.31	31	0.028

Note: S.E. = Standard Error; MI = Main Idea; DET = Detail; INF = Inference

Relevance for PIE and Second Language Studies

The findings that content-related visuals can reduce L2 listening test difficulty may trigger the discussion about their place in the listening construct. At the core of such a discussion lays the fact that some authentic listening settings do include content-related visuals in one form or another (e.g., visual aids during lectures). In light of this, it would be unfair to deprive L2 listeners of information they would normally encounter in a corresponding listening situation. Another unfair consequence of not including content-related visuals may be the increase of listening message difficulty. To avoid these biases, it seems reasonable to follow Suvorov's (2014) suggestion to define a listening construct for each specific TLU domain of interest.

Including context visuals in the construct is more questionable. The finding that context visuals negatively influence the difficulty of lecture-like listening passages may serve as an argument against their inclusion. On the other hand, context visuals are a part of authentic academic lectures, which makes them desirable accompaniments to the listening construct in order to better represent TLU domains. To solve this dilemma, the following question needs to be considered: Should listening be inclusive of processing semantically unrelated visuals even if they may be distracting or adding an additional cognitive load for listeners? The answer is yes, if it reflects what occurs in real-life listening contexts. In other words, even though context visuals may be detrimental for L2 listeners' comprehension during authentic lectures, it is the reality that L2 listeners should face and be prepared for.

A refined definition of listening to academic lectures, encouraged by the results of this study, may bear several implications in regard to the use of content and context visuals in L2 tests. One suggestion is to accompany passages in L2 listening tests with videos that would best

mimic real-life situation. Such videos would ideally contain the combination of content and context visual information. Even though creating and adding such videos to listening tests may conflict with practicality considerations, it would be a move to the right direction for both high- and low-stakes L2 listening tests, as it will enable test developers to assess academic listening in a more realistic way. Moreover, it will help to avoid unfair increase (in case of the absence of content-enriched visuals) or decrease (in case of the absence of context-oriented visuals) of a listening message.

Another suggestion would be for L2 teachers to refrain from using purely context videos in instructional classroom activities for beginner-level students. Since context-oriented videos have a potential to be distracting, their presence may lead to an unnecessary cognitive overload of lower proficiency students, which, in turn, may thwart the primary purpose of developing the listening skill. Similarly, the use of context videos in formative and summative classroom assessments for students at the beginning stages of their L2 listening ability development should be taken carefully. Lower-level students may be especially susceptible to being distracted from listening due to their unstable L2 ability. Consequently, test scores may dramatically suffer from the presence of semantically unrelated videos and lead to misinformed instructional or diagnostic decisions.

References

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

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.

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

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

Suvorov, R. (2014). 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. (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.