Available online at www.ejal.eu
http://dx.doi.org/10.32601/ejal.543773
Eurasian Journal of Applied Linguistics, 5(1), 1–22

EJAL
Eurasian Journal of Applied Linguistics

Effect of Input Mode on EFL Free-Recall Listening Performance: A Mixed-Method Study

Yali Shi [a] *

[a] *School of International Studies, 866, Yuhangtang Road, Hangzhou, 310058, PRC*

Received 24 April 2018 | Received in revised form 23 November 2018 | Accepted 10 December 2018

APA Citation:
Shi, Y. (2019). Effect of input mode on EFL free-recall listening performance: A mixed-method study. *Eurasian Journal of Applied Linguistics, 5*(1), 1-22. **Doi: 10.32601/ejal.543773**

Abstract

This study conducted a mixed-method study of the influence of audio and video input mode on free-recall listening performance. It first explored quantitatively whether input mode significantly influenced 34 sophomores' performance in general and across two genres (passage and long dialogue) and three ranks of idea units (the discourse topic, main point, and supporting detail). Then it investigated qualitatively how four of the participants interacted with the audio and video input. T-test results showed the video mode significantly facilitated listening performance in general and for long dialogue in particular, as well as the recall of supporting details for long dialogue. Qualitative findings revealed that the participants' interaction with the video varied with their language proficiency and the visual-verbal relationship in the input, suggesting the threshold of language proficiency for the visuals to take effect and the intervening effect of visual input and task features.

Keywords: input mode; audio-video input; free-recall listening

1. Introduction

In large-scale English proficiency tests nowadays, the input in the listening section is predominantly audio-mediated for both practicality and validity concerns. However, typical of TLU (Target Language Use) domain, video is not only popularly and effectively used in listening instruction (e.g., Cross, 2009; Feak & Salehzadeh, 2001; Mohamadkhani, Farokhi, & Farokhi, 2013) but also frequently present in the listening activity in real life. Therefore, it is necessary to incorporate visuals into the input to achieve the authenticity of the listening test task. Yet, on the other hand, there was still controversy concerning the construct (i.e. what is actually measured) of the listening test with visuals, which necessitates further study into the influence of visuals

* Corresponding author.
E-mail address: ylshi@zju.edu.cn

on listening performance in terms of whether the influence is present and how the influence takes place.

2. Literature review

The past three decades have witnessed an abundant body of research on the influence of visuals on listening performance, which can be divided into three focuses.

2.1. Effect of visuals on listening test performance

As to whether visuals exert facilitative or inhibitory effects, there have been mixed results. Most studies showed that participants performed better in listening tests with visuals (Brett, 1997; Lesnov, 2017; Sueyoshi & Hardison, 2005; Wagner, 2010a, 2013), whereas several found the effect of visuals either negative (Ginther, 2002; Lenz, 2014; Pusey & Suvorov 2009) or absent (Lesnov, 2018). Some studies went further by introducing a variety of moderating variables such as genre (Batty, 2015; Ginther, 2002; Suvorov, 2009), the access to the test questions (Wagner, 2013), working memory (Pusey & Lenz, 2014), and proficiency level (Batty, 2015; Ginther, 2002; Pardo-Ballester, 2016; Sueyoshi & Hardison, 2005). Yet the findings lack consistency given the difference in specific variables and contexts.

To begin with, with regard to the interaction with proficiency level, Sueyoshi and Hardison (2005) found that when listening to academic lectures, advanced learners of ESL performed the best in visual condition with face image, while low-intermediate learners performed the best in visual condition with both gesture and face image. Similarly, in the case of Spanish passage listening, Pardo-Ballester (2016) concluded that the performance on inference items were better with video format for higher-proficiency learners, and the opposite was true for lower-proficiency ones.

Secondly, with respect to the interaction with genre, Suvorov (2009) identified a negative effect of video input on listening performance in lectures, and no effect of video input on performance in dialogues.

In addition, concerning the interaction with genre as well as proficiency level, Ginther (2002) found that both mini-talks and academic discussions with content visuals facilitated performance, while mini-talks with context visuals did not; and dialogues either with visuals or not had no effect on performance. Meanwhile, the effect of visuals did not differ between high- and low-proficiency groups. However, in Batty (2015)'s study, no significant interaction was identified between video/audio format on the one hand, and either three genres (monologue, conversation, and lecture) or four proficiency levels on the other hand.

Finally, no significant interaction effect was found between some other variables (e.g., the access to the test questions, working memory) and listening input mode on listening performance (e.g., Pusey & Lenz, 2014; Wagner, 2013).

To sum up, despite the absence of a consensus on the influence of visuals on listening performance in previous studies, one thing is certain that the influence depends on

specific contexts as well as other intervening factors (e.g., language proficiency of examinees, type of listening material, features of visual input, and type of listening task). Especially, the simple and intuitive division of visuals into either content or context ones failed to provide a finer-grained picture of the issue, which is one of the reasons behind the inconsistency in previous findings.

2.2. Examinees' interaction with visuals

Studies on how examinees interact with visuals in listening comprehension are mainly conducted in two ways—qualitatively and quantitatively. Qualitatively, think-aloud, observations and interviews were mainly used. For example, Ockey (2007) found that in computer-based listening tests examinees hardly had any engagement with still images but did interact with a video stimulus in different ways and to different extent. Whereas this study analyzed the data somewhat intuitively, others did it in accordance with a specified framework.

For instance, Seo (2002) examined how audio-only and audiovisual formats influenced participants' use of strategies. It was found that more cognitive strategies (e.g., inferencing, elaborating, evaluating information) were elicited by audio-visual format, while audio-only format elicited more metacognitive strategies (e.g., identifying problems, self evaluating).

Similarly, Gruba (2004), in order to probe into how participants process visuals, developed a seven-category framework based on the constructivist perspective of comprehension. Using this framework, he analyzed the immediate retrospective verbal reports of L2 Japanese tertiary level students as they were viewing three authentic Japanese news broadcasts. Judging from the result that visual elements worked in a number of ways that go beyond merely 'supporting' verbal elements, he argued for considering visuals as integral resources to comprehension.

In the further and more elaborate follow-up study, Gruba (2006), based on the framework of media literacy for "playing the text", conducted a descriptive case study of 22 participants' thinking processes in watching three digitized news clips. As a result, nine different forms of interactions were summarized, which were respectively—making sense of setting and context, no longer attending directly and consciously to every single element, developing macrostructure, exploring ideas without commitment, playing without advance preparation, playing as joyfulness, recovering from comprehension failure, stalling decision making, and establishing signposts and boundary lines.

Quantitatively, there are two approaches. One focused on the viewing behavior of participants, and the other elicited views of participants on the effect of visuals by means of questionnaires. As regards the first approach, Wagner (2007, 2010b) investigated examinees' behavior in terms of the amount of time they make eye contact with the video monitor as well as the rate they actually view the video texts. The result showed that examinees oriented to the video monitor 69% of the time in video format, and they tended to view the video at a higher rate when the dialogue rather than the

lecture was given; besides, there was a moderate negative correlation between viewing rate and test performance. Suvorov (2015), employing more precise measurement such as eye-tracking technology, explored the relationship between examinees' viewing behavior in terms of fixation rate, dwell rate, and the total dwell time on the one hand and their performance in academic listening with two types of videos (context and content) on the other hand. Despite no significant relationship between the three eye-tracking measures and the test scores, fixation rates and total dwell time values were found to be significantly different between two types of videos.

As to the second approach, some studies (e.g., Pardo-Ballester, 2016; Sueyoshi & Hardison, 2005) identified participants' general preference of visual format over audio one in listening tasks, and it was found that the main merit of multimedia format lay in its efficiency and ongoing feedback to listening tasks (Brett, 1997). Nevertheless, when relating participants' preference for certain visual input to their performance, Suvorov (2009) observed they did not perform better with the visual input they preferred.

Taken together, notwithstanding either the much specified and elaborate qualitative analysis or the use of advanced technology (i.e. eye-tracking) in the quantitative analysis of examinee-visuals interaction in previous studies, most of them treated listening comprehension as a whole without considering the possibility that the interaction may vary by different levels of comprehension (e.g., grasping main idea, explicit details).

2.3. The construct of video-mediated listening assessment

Accompanying the mixed finding about the effect of visuals on listening comprehension, contention exists in terms of what is actually measured in listening assessment with visuals. While Shin (1998) found both concurrent and construct validity evidence for the videotape-formatted listening test as a measure of academic lecture listening ability, other researchers take the use of visuals into consideration in exploring the construct of video-mediated listening assessment.

For example, in the construct validation of a video-based listening assessment, Wagner (2002) provided quantitative evidence for a two-factor model of listening ability: ability to comprehend explicitly stated information and implicit information in aural texts, and some sort of method effect related to item type. By means of verbal report methods, Wagner (2008) found that the examinees attended to and utilized the nonverbal information in different manners and thus recommended not only integrating the ability to utilize the nonverbal information into the construct of listening ability but also classifying the ability of using visual information as a kind of pragmatic knowledge. Holding similar ideas, Cubilo and Winke (2013) recommended incorporating knowledge of how to interpret nonverbal cues into the construct of the listening-writing integrated task. Likewise, Lesnov (2017) argued for the consideration of understanding context video as part of the academic listening comprehension construct.

However, holding a more reserved point, other researchers (e.g., Lesnov, 2018; Li, 2013) called for more empirical research and theoretical thinking to further clarify the construct of video-mediated listening assessment. In this sense, more studies in the field would help to shed light on it.

2.4. Research questions

The above review has revealed the lack of conclusive evidence on the effect of visuals on listening performance, the predominant use of multiple-choice items, and the failure to probe into the specific feature of visuals and examinees' interaction with the visuals on different levels of comprehension. In view of this, the present study is conducted.

Specifically, it provides a mixed-method investigation of the influence of input mode (audio vs. video) on free-recall listening performance in non-academic TLU domain. First, it explores quantitatively the effect of two modes on performance in general and across two genres (passage and long dialogue) and three ranks of idea units (the discourse topic, main point, and supporting detail). Then, it investigates qualitatively how participants interact with two modes of input through retrospective verbal reports and follow-up interviews. To be specific, the study examines the effect of input mode by addressing the following three research questions:

1. RQ1: Does input mode significantly influence free-recall listening performance in general and across two genres respectively?
2. RQ2: Does the influence vary across three different ranks of idea units in two genres?
3. RQ3: How do participants interact with visual and audio input in listening?

3. Method

3.1. Participants

The participants consisted of an intact College English teaching class in a university in mainland China, totaling 34 sophomores majoring in humanities. Based on College English Test Band 4 (CET-4)[1†] scores, two higher-proficiency participants (two females) and two lower-proficiency ones (one female and one male) were randomly chosen for the qualitative investigation. They are respectively Participant 1 (female, higher-proficiency), Participant 2 (female, higher-proficiency), Participant 3 (female, lower-proficiency), and Participant 4 (male, lower-proficiency).

3.2. Instrument

[†] †College English test is a large-scale English language test which is administered by the National College English Testing Committee on behalf of the Ministry of Education of People's Republic of China, aiming to examine the English proficiency of undergraduate students in terms of whether they reach the required English levels specified in the National College English Teaching Syllabuses. It consists of three tests: Band 4 (CET-4), Band 6 (CET-6), and the CET–Spoken English Test (CET-SET) (Zheng & Cheng, 2008).

To minimize the potential effects of construct-irrelevant factors, the difficulty of listening material, topic, speed, and the quality of video were taken into consideration in the selection. Thus, two videos were finally selected as below in Table 1.

Table 1. Listening material information

Category	Theme	Genre	Duration	Speed
Audio/Video 1	Facebook addiction	Passage	84s	182w/m
Audio/Video 2	Shopping for healthy food	Long dialogue	120s	258w/m

3.3. Research design and procedure

All participants were first asked to provide basic information of gender and CET 4 scores. Given the variation of individual examinees in terms of language ability, cognitive style, and the ability to utilize the non-verbal information (Wagner, 2008), the present study used the within-subjects design in which all participants were exposed to two modes of input, in order to avoid score variances caused by individual differences (Suvorov, 2009). Meanwhile, there could be the influence of listening material features (e.g., text type, topic, characteristics and relationship of interlocutors) on listening performance (Dunkel, Henning, & Chaudron, 1993; Rubin, 1994). Given this, it is tough, if not totally impossible to select fully comparable listening material for each of the video and audio mode. Such being the case, for the purpose of minimizing as much as possible the variances caused by different listening material in the experiment, the same materials were used for both modes. Nevertheless, the issue of memory effect would possibly come out of the repetitive use of the same material. To solve this problem, participants were given the listening task in the audio mode first in view of the transient nature of aural input, and then the same task in the video mode, with a one-day interval between the two. To lessen the memory effect, immediately after the completion of the listening task in the audio mode, an intervening procedure was introduced in the format of a new one in the same mode. This practice turned out effective to certain extent. For one thing, the participants claimed listening to the material assigned for the first time, when they were assigned the same material in the video mode the next day. For another, even after the researcher told them that they had listened to the same material as the day before, most of them said they barely had any memory of it, whereas several said they did have some impression, but did not think it helped greatly their free-recall performance. In all, the aforementioned measures were to isolate the independent variable of either audio or video mode in listening tasks. Specifically, the research design was organized this way:

- Participants first listened to the audio version of passage and long dialogue respectively, and immediately after this, finished a free-recall within 5 minutes for each, in which they were asked to write down whatever they got out of listening.
- The next day participants watched the video versions of the same material and again finished free-recall within 5 minutes for each in which they were asked to write down whatever they got out of viewing and listening.

- Four of the participants agreed to give retrospective verbal reporting of the listening process and also to have a follow-up interview.

3.4. Analytic scoring of free-recalls and quantitative analysis

For each participant, there were two free-recall protocols for both passage and long dialogue listening in two modes. Yet, four of the participants either lacked one or two recall protocols or left blanks in the protocol, and thus their data were considered invalid. As a result, there were 120 free-recall protocols in total (30 × 4).

To score the free recalls, the two listening scripts in this study were first analyzed in terms of different levels of idea units—the discourse topic, main point, and supporting detail, as shown in Tables 2 and 3.

Table 2. A hierarchy of idea units in the passage

Level of Idea	Discourse topic	Mark	Main Points		Mark	Level of Idea	Supporting Details	Mark
1	Facebook addiction	2 points						
2			2-1	Changing your profile picture too often	1	2-1-1	Worrying about your image	0.5
			2-2	Changing your status update too often	1	2-2-1	Unnecessary	0.5
			2-3	Access Facebook via mobile phone	1	2-3-1	Thinking about it	0.5
			2-4	Adorn your page with too many apps	1	2-4-1	Getting rid of them	0.5
						2-4-2	Refusing new ones	0.5
			2-5	Being taken over by Facebook in vocabulary	1	2-5-1	Spending time with friends in real life	0.5

Table 3. A hierarchy of idea units in the long dialogue

Level of Idea	Discourse topic	Mark	Level of Idea	Main Points	Mark	Level of Idea	Supporting Details	Mark
1	Shopping for healthy food	2						
2			2-1	Have an idea of a menu	1.5	2-1-1	Take a look at what your plate looks like	1
						2-1-1-1	Components of a plate	0.5
						2-1-2	Have recipes and menus on the websites	1
			2-2	Focus on the produce cell	1.5	2-2-1	Necessary to have lots of vegetables and fruits	1
						2-2-2	Try different things in different ways	1
						2-2-2-1	Be creative	0.5

Specifically, first, the discourse topic refers to the gist or theme of the text, i.e., what is mainly discussed. For example, the passage centers on Facebook addiction while the dialogue focuses on how to shop for healthy food. Then, the main point refers to the aspects from which the topic is addressed, as exemplified by the five aspects to illustrate the phenomenon of Facebook addiction and the two key points for attention in food shopping. Third, the supporting detail refers to specific reasons or evidence to specify the main point. For instance, for each of the five aspects of the Facebook addiction, there is one or two backing evidence or further specification; and for each of the two notes in food shopping, there are three further details provided. Together, all these idea units constitute the meaning or content of the whole text.

The total mark was 10 for both materials. For each idea unit, the differential score value was assigned according to their levels, as illustrated in the two tables above. For example, in the passage, the discourse topic was assigned 2 points, five main points were respectively assigned 1 point, and six supporting details were respectively assigned 0.5 point. In the long dialogue, the discourse topic was assigned 2 points, and two main points were respectively assigned 1.5 points; for six supporting details, four of them with more content were respectively assigned 1 point, and two others with less content were respectively assigned 0.5 point. As a result, both a holistic score and analytic scores for each of the three ranks of idea units were generated.

Finally, the data was administered the paired-samples t-test to test whether performance significantly differed between two modes of input (i.e. audio vs. video) in general and across two genres (i.e. passage and long dialogue) and three different ranks of idea units.

3.5. Qualitative analysis of free-recalls and coding of retrospective verbal reports

Firstly, in accordance with the level of idea units presented in Tables 2 and 3, free-recall protocols of both passage and long dialogue from the four randomly-chosen participants were analyzed and compared in terms of which idea unit was recalled and which level it belonged to.

Table 4. Integration of verbal and visual information (Schriver, 1997)

Category	Function
Redundant	words and pictures convey identical content
Supplementary	words and pictures provide different content, with one mode presenting the main idea and the other mode supplementing it
Complementary	words and pictures provide different content, with both modes being necessary to understand the main idea
Juxtapositional	words and pictures provide different content, with both modes presenting the ideas that clash
Stage-setting	words and pictures present different content, with one mode providing the content and another mode giving the main idea

Secondly, based on the five kinds of visual-textual integration (Schriver, 1997) in Table 4 above, two videos were analyzed in terms of the manner of integration between

concurrent visual and audio input, and the result of coding was presented in Tables 5 and 6.

Table 5. A representation of audio-visual image in passage video (Analysis of 11 frames)

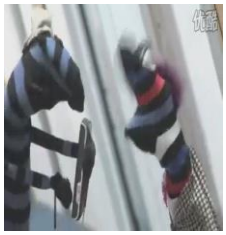
Level 1 – redundant				
				
Level 2-1 Stage-setting	Level 2-2 Stage-setting	Level 2-3 Redundant	Level 2-4 Stage-setting	Level 2-5 Supplementary
				
Level 2-1-1 Supplementary	Level 2-2-1 Supplementary	Level 2-3-1 Supplementary	Level 2-4-1 & 2-4-2 Redundant	Level 2-5-1 Supplementary
				

Table 6. A representation of audio-visual image in long dialogue video (Analysis of 9 frames)

Turn 1
Level 1 Stage-setting


Turn 2

Level 2-1 /

Level 2-1-1 /

Level 2-1-1-1 Stage setting

Turn 3

Level 2-1-2 Supplementary

Turn 4

Level 2-2 Supplementary

Level 2-2-1 /

Level 2-2-2 Supplementary

Turn 5

Level 2-2-2-1 /

Note: "/" refers to that there is no relation between the verbal and visual information.

Thirdly, to code how the participants processed the visual and verbal input in the listening process, Paivio (1986)'s Dual Coding Theory was drawn on. The theory, through assuming the co-existence of verbal and non-verbal cognitive processing subsystems, proposes three types of concurrent processing of two types of input, as presented in Table 7.

Table 7. Processing type of visual and verbal information

Processing type	Content
a. representational	the direct activation of verbal or non-verbal representations
b. referential	the activation of the verbal system by the nonverbal system or vice versa
c. associative	the activation of representations within the same verbal or nonverbal system

Finally, to supplement the retrospective verbal reports, a follow-up interview was given in terms of how participants processed audio and video inputs and why they got or missed certain idea units.

4. Results

4.1. Research question 1: Does input mode significantly influence listening free-recall performance in general and across two genres respectively?

Table 8 presents the descriptive statistics for the free-recall performance in the audio and video modes in general and for both passage and long dialogue.

Table 8. Descriptive statistics of listening performance in general and across two genres

Overall	*Mean*	*Max*	*Min*	*SD*
Audio	2.43	5.00	.00	1.27
Video	2.91	6.50	.00	1.46
Passage	*Mean*	*Max*	*Min*	*SD*
Audio 1	1.80	4.00	.50	.77
Video 1	2.20	5.00	.50	1.03
Long Dialogue	*Mean*	*Max*	*Min*	*SD*
Audio 2	3.05	5.00	.00	1.37
Video 2	3.62	6.50	.00	1.49

Note: 1—passage; 2—long dialogue

As is shown in Table 8, overall as well as for both passage and long dialogue respectively, performance in the video mode was higher than that in the audio one, suggesting that participants recalled better in the video mode; and the range was also wider in the video mode. Meanwhile, performance in the video mode spread out more widely than in the audio one.

Paired-samples t-test results suggested that performance in the video mode (M = 2.91) was significantly higher than that in the audio mode (M = 2.43) in general (t (59) = -2.890, p = .005, r = .35). The same was true for long dialogue (t (29) = -2.142, p = .041, r =.37). Both significant effects reached the medium size. However, there was no significant difference between performances in two modes for passage (t (29) = -1.922, p = .065). Put differently, the video mode facilitated free-recall performance in general and for long dialogue in particular.

4.2. Research Question 2: Does the influence vary across three different ranks of idea units in two genres?

Table 9 lists the descriptive statistics for the free-recall performance in the two modes on three ranks in both genres.

Table 9. Descriptive statistics of listening performance on three ranks

Performance	Audio 1		Video 1		Audio 2		Video 2	
on	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Rank 1	1.12	.41	1.37	.67	1.12	.70	1.08	.78
Rank 2	.27	.58	.45	.65	.77	.64	.97	.82
Rank 3	.42	.23	.38	.22	1.17	.75	1.57	.92

Note: 1—passage; 2—long dialogue; Rank 1—discourse topic; Rank 2—main point; Rank 3—supporting detail

As Table 9 presents, in passage listening, mean scores in the recall of discourse topic and main points were both higher in the video mode, while the opposite was true for the recall of supporting details. In long dialogue listening, the mean scores in the recall of both main points and supporting details were both higher in the video mode, whereas the opposite was true for the recall of discourse topic. As to the dispersion of scores, for both genres and modes, performance on rank 2 spread out more widely than on the other two ranks.

Paired-samples t-test results suggested that performance on rank 3 for long dialogue in the video mode (M = 1.57) was significantly higher than that in the audio mode (M = 1.17) (t (29) = -2.887, p = .007) and the effect reached almost the large size (r = .47). However, there was no significant difference between performances in the two modes on rank 1 for either passage (t (29) = -1.634, p = .113) or long dialogue (t (29) = .195, p = .847), on rank 2 for either passage (t (29) = -1.249, p = .222) or long dialogue (t (29) = -1.099, p = .281), and on rank 3 for passage (t (29) = .626, p = .536). Put another way, the video mode only facilitated the recall of supporting details for long dialogue listening. It did not facilitate the recall of either the discourse topic or the main point across genres.

4.3. Research Question 3: How do participants interact with visual and audio input in listening? Ancillary analyses

4.3.1. Passage Processing Analysis

Table 10 demonstrates in detail which level of idea units the four participants recalled, as well as how they processed visual and audio input during passage listening.

Table 10. Passage recall protocol analysis and coding for input processing

		Level 1	Level 2					Level 3				
Participant	Mode	L1	L 2-1	L2-2	L 2-3	L 2-4	L 2-5	L 2-1-1	L 2-2-1	L 2-4-1	L 2-4-2	L 2-5-1
1	Audio	√	√		√							√
	Video	√ a	√		√ b		√ b					√
2	Audio	√										√
	Video	√ b	√ b									√
3	Audio											√
	Video	√a										√
4	Audio	√										√
	Video	√a										√

(Note: √—the presence of the idea unit on that level in the recall protocol; "a"—representational processing; "b"—referential processing; higher-level participants—1 & 2; lower-level participants—3 & 4)

As Table 10 shows, for Participant 1, there were three differences between the two modes. First, concerning Level 1, there was only "Facebook" in the audio mode while in the video one the expression "Facebook addiction" was complete. Second, as to Level 2-3, the idea unit was mentioned incompletely by the expression "access the Facebook" in the audio mode but completely in the video one. Third, for Level 2-5, the idea unit was missing in the audio mode.

Her retrospective verbal report shed some light on the differences. For Level 1, she said that she saw the words "what is Facebook addiction" at the beginning of the video and thus judged it must be concerned with the main idea. As to Level 2-3, she said that she saw the puppet was doing something with a cell phone in his hand and thus integrated this with the meaning she grasped by audio input. With respect to Level 2-5, she said that she judged from the words "I'm going to do some Facebooking" and the facial expression of the man and got the idea of "using too much Facebook in language". Concerning all the missing idea units in both modes, she explained in the interview that she just missed them when listening, and when viewing she was totally distracted and confused by so many pieces of visual information that she could not absorb them at all.

For Participant 2, there were two differences. First, for Level 1, there was an incomplete idea unit as "Facebook" in the audio mode while in the video one the discourse topic was incorrectly summarized as "the function of Facebook". Second, concerning Level 2-1, the idea unit was missing in the audio mode and was incomplete in the video one.

According to her verbal report, concerning Level 1, she said that she totally neglected the words at the beginning of the video and was not sure of what was talked about the Facebook. When listening, she made a guess by some key words and assumed it was about the popularity of Facebook; when viewing, she made a guess by some images and assumed it was about the influence of Facebook. When it came to Level 2-1, she said that when viewing she noticed something was constantly changing on the home page but missed what changed. For all the missing idea units in both modes, she explained

in the interview that she just missed them when listening, and when viewing she just got so blank-minded that she could not relate what she saw with what she heard.

For Participant 3, concerning Level 1, the idea unit was missing in the audio mode and incomplete in the video one. To explain this, she said in her verbal report that when listening, she understood it as the influence of modern technology on human life because she heard such words as "Facebook" and "cellphone" mentioned for several times. When viewing, she said that the words at the beginning on the screen gave her some clue. As to all the missing idea units in both modes, she honestly disclosed in the interview that she listened word by word and remembered general meaning of some sentences, but could only deliver it in the form of frequently-spoken words. When viewing, she just paid attention to such elements as subtitles, facial expression and movements.

For Participant 4, for Level 1, there was only "Facebook" in the audio mode while in the video one the expression was complete. For that, he gave similar explanations to that of Participant 3. As regards all the missing idea units in both modes, he said in the interview that he did not catch the words in detail but formed a general impression based on background knowledge and association. The only difference between the two modes was that he made associations by the key word heard in the audio mode but by the impressive image seen in the video mode.

As a whole, in passage listening, the interaction between the participants and the listening input seemed to vary with the listening proficiency of participants and the visual-verbal relationship. Specifically, on the one hand, while higher-proficiency participants were able to make use of visual input to access verbal information, lower-proficiency ones often failed to do so, being unable to undertake concurrent processing of two kinds of input. On the other hand, visual input tended to facilitate comprehension especially when they conveyed the same information as audio input or supplemented the main idea expressed by it.

4.3.2. Dialogue processing analysis

Table 11 demonstrates in detail which level of idea units the four participants recalled and how they processed visual as well as verbal input during dialogue listening.

Table 11. Dialogue recall protocol analysis and coding for input processing

		Level l	Level 2		Level 3					
Participant	Mode	Level 1	Level 2-1	Level 2-2	Level 2-1-1	Level 2-1-2	Level 2-1-1-1	Level 2-2-1	Level 2-2-2	Level 2-2-2-1
1	Audio	√	√		√		√			
	Video	√a	√				√			
2	Audio	√							√	
	Video	√	√b		√b		√b		√	
3	Audio	√			√				√	
	Video	√			√		√b		√	
4	Audio		√					√		
	Video	√a							√c	

(Note: √—the presence of the idea unit on that level in the recall protocol; "a"—representational processing; "b"—referential processing; "c"—associative processing; higher-level participants—1 & 2; lower-level participants—3 & 4)

As Table 11 shows, for Participant 1, there were two differences across two modes. First, concerning Level 1, the discourse topic was summarized incompletely as "healthy food" in the audio mode but completely as "take control of shopping for healthy food" in the video one. Second, for Level 2-1-1, the idea unit was missing in the video mode.

In her retrospective verbal report, she said that for Level 1 she grasped the term "healthy food" stressed by the first speaker at the beginning when listening, and when viewing she noticed the caption introducing the topic at the bottom of the screen, which gave her a hint. As to Level 2-1-1, she said that she clearly heard the sentence "take a look at what the plate looks like"; but when viewing, she just focused on the image, without attending to what was said. When it came to all the missing idea units in both modes, she explained in the interview that she either missed it or failed to recall it when listening, and when viewing she got so attracted by the dishes of food on the desk that she failed to process other input.

For Participant 2, concerning Levels 2-1, 2-1-1, 2-1-1-1, the idea units were absent in the audio mode but present in the video one. Her verbal report shed some light on the difference. First, concerning Levels 2-1 and 2-1-1, she got some clue from the interaction (e.g., gestures, eye fixations) between the two speakers despite the lack of relationship between the verbal and the visual input. Concerning Level 2-1-1-1, she said that the video clearly illustrated the meaning of the words which helped her quickly grasped the audio information. Third, concerning the missing idea units of Levels 2-2, 2-1-2, 2-2-1 and 2-2-2-1, she gave a similar reason to that for passage listening.

For Participant 3, the idea unit on Level 2-1-1-1 was absent in the audio mode but present in the video one. To explain this, she said in her verbal report that she understood what she heard as soon as she saw the video which illustrated the content conveyed by the audio information. Concerning all the missing idea units, she gave a similar reason to that for passage listening.

For Participant 4, there were four differences. First, concerning Level 1, the discourse topic was summarized incompletely as "healthy eating and shopping make easy" in the video mode but missing in the audio one. Second, the idea unit for Level 2-2-2 was

present in the video mode but absent in the audio one. Third, as to Levels 2-1 and 2-2-1, the idea units were present in Chinese in the audio mode but absent in the video one.

According to his verbal report, for Level 1, he noticed the caption which introduced the topic at the bottom of the screen, but in the audio mode he missed the beginning because he couldn't concentrate quickly. And he got the idea unit on Level 2-2-2 through the association of the image when viewing. As to Levels 2-1 and 2-2-1, he got both idea units through key word association in the audio mode, but in the video mode he just had no impression of them. Finally, with respect to all the missing idea units in both modes, he offered similar reasons to those for passage listening.

As a whole, similarly, in long dialogue listening, visual input tended to facilitate comprehension for both proficiency levels of participants especially when they illustrated or specified the main idea expressed by the verbal information. Meanwhile, higher-proficiency participants tended to make use of visual input to access verbal information more frequently, whereas lower-proficiency ones did not exhibit such tendency. Instead, they tended to exclusively rely on visual input to cope with failure in listening comprehension.

5. Discussion

5.1. Effect of input mode on listening performance

The study revealed a significantly facilitative effect of video input on listening performance in general, which echoes the findings in previous studies (e.g., Brett, 1997; Lesnov, 2017; Sueyoshi & Hardison, 2005; Wagner, 2010a, 2013). When this overall significant effect was further probed into, it was found that the video mode significantly facilitated the general performance for long dialogue instead of passage listening. This contradicts the previous finding that the video input had no significant effect on performance for dialogue listening (e.g., Batty, 2015; Ginther, 2002; Suvorov, 2009). The contradiction would probably be related with different situational contexts (daily vs. academic context) or the authenticity of the video (real-life vs. acted). Moreover, when this significant effect of video input on performance in long dialogue listening was more deeply looked into, it turned out that the video mode only significantly facilitated the recall of supporting details rather than either the discourse topic or the main point. This adds the indirect evidence to the finding that in listening tests administered in the traditional audio mode, participants generally performed worst on detail items than the main idea and inference ones (Shang, 2005).

To explain the absence of the significant video effect on the recall of either the discourse topic or the main point, it is necessary to draw on the research concerning the nature of listening comprehension. Specifically, Rost (1990) argues that listening is essentially an inferential process in that it includes constructing propositional meaning through supplying relational links and assigning underlying links in the discourse. Similarly, Field (2008) states that listening involves not only extracting raw information from the speech, but also dealing with it by means of selecting what

information is relevant, monitoring the information to ensure it is consistent with what has gone before, integrating the information into a representation of what has been said so far, and building an information structure of macro- and micro-points. Hence, the problem of the participants can be diagnosed as the failure to handle idea units in the sense of forming a mental model of hierarchical ideas based on the links among them. Simply put, they were unclear if not totally ignorant of the relationship among the separate ideas they got, and so they only recalled separate minor details which were not logically inter-related in their mind, as was indicated in their recall protocols in which the idea units recalled were just piled up rather than organized logically.

The above suggests the threshold of language proficiency in video-mediated listening: visual input would take effect substantially only when listeners are able to deal with audio input within a global network. This lends support to the finding that examinees made use of the nonverbal information in different manners (Wagner, 2008) and that video format only facilitated listening performance on inference items for higher-proficiency learners (Pardo-Ballester, 2016). Yet it contrasts with the absence of interactions of participants' proficiency levels with visuals found in previous studies (e.g., Batty, 2015; Ginther, 2002).

5.2. Processing of visual input

As regards processing of visual input, it can be influenced by input and task features. To begin with, input features refer in this context to the verbal-visual relationship. There are three kinds of relationship in both the passage (i.e. redundant, supplementary, and stage-setting) and the dialogue (i.e. stage-setting, supplementary, and no relation). A close look at the free-recall performance revealed that participants employed redundant visual input to access verbal information in both genres. However, higher-proficiency participants were found to be more capable of this than lower-proficiency ones. On the one hand, this is similar to the finding that content instead of context visuals in mini-talks and academic discussions improved listening performance (Ginther, 2002). On the other hand, it lent support to the aforementioned threshold of language proficiency. Meanwhile, the use of supplementary and stage-setting visual input depended on the degree of the understanding of simultaneous audio input. If participants failed to catch the latter, the former wouldn't play a role. Furthermore, the no-relation visual input in the dialogue did not hamper participants' recall in general probably because the apparently unrelated visual input provided a transitional point or a cognitive buffer for them to better organize the just-heard information. Yet, for two higher-proficiency participants, much concentration on visual input tended to distract them to some extent, which lends support to the finding that the longer the participants viewed the video, the poorer their listening performance was (Wagner, 2007, 2010b).

Then, task features refer to the way comprehension is measured. Unlike selected task types such as multiple-choice and true-or-false question, free-recall task is constructed in that it requires the construction of the response in a subjective way, and so it is more cognitively demanding (Bachman & Palmer, 2010). Specifically, the recall of

information would bring about heavy memory load and may interfere with attending to some minor details (Shang, 2005). As revealed by retrospective verbal reports, participants mainly conducted referential processing. In other words, they used video input to compensate for the inadequacy in processing audio input or to help them recall some unclear information through schemata-activated association.

Taken together, it suggests that the effect of visuals in listening would vary with visual input and task features: the effect of visuals would be facilitating if it is consistent with audio input; referential processing would be conducted if the task is demanding. The former one is similar to Grimes (1991)'s "belongingness" hypothesis that auditory-visual redundancy would be treated as conveying one message and won't cause attentional capacity exhausted.

5.3. Reflecting on the construct of video-mediated listening assessment

In view of the aforementioned findings, it is possible to provide an attempt to answer the call for a rethinking of what is measured by video-mediated listening assessment (e.g., Lesnov, 2017, 2018; Li, 2013). In accordance with the Interactionalist perspective on construct definition (Chapelle, 2002), there are three essential elements in the construct of listening tests with visuals: trait, context and metacognitive strategies. To begin with, trait includes not only linguistic knowledge but also paralinguistic knowledge and visual literacy. Defined in ACRL Visual Literacy Competency Standards for Higher Education (2011), visual literacy refers to "a set of abilities enabling an individual to effectively find, interpret, evaluate, use, and create images and visual media". However, as it is a broad concept and varies across disciplines, how to define visual-related competence relevant to listening comprehension is a question to be considered. Then, context refers to environmental conditions under which performance is observed. As language use happens in diversified contexts, it is not sufficient to only mark the general TLU domain in the construct. Instead, the most typical features (e.g., location, participants, and channels) of the most representative TLU domains should be incorporated into the construct. For example, even in the same domain of academic listening, seminar and lecture differ in the degree of visual involvement and type of visual input, the processing of which definitely requires different visual skills. Therefore, to specify typical contextual features in the construct definition would be conducive not only to the operationalization of the construct but also to the controlling for visual skills to ensure consistency and fairness. The last factor—metacognitive strategies help examinees apply their linguistic knowledge, paralinguistic knowledge, and visual literacy to specific listening context. Thus, more process-oriented studies, based on visual grammar (Kress & van Leeuwen, 2006) and multi-modal discourse analysis (O'Halloran, 2004), can be carried out to systematically explore how different examinees process different kinds of visual input. As a whole, it suggests that the specification of the construct of video-mediated listening assessment should take into consideration of not only language knowledge required, but also contextual features typical of the represented TLU domain and any visual skills and metacognitive strategies involved in input processing.

6. Conclusions

To conclude, with a view to probing into the influence of input mode on listening performance, and to further clarifying the interaction between listeners and the input with visuals, as well as to shedding light on the construct of video-mediated listening assessment, the present study has conducted a mixed-method investigation of the influence of audio and video mode on free-recall performance of passage and long dialogue listening. Quantitative results showed the video mode significantly facilitated listening performance in general and for long dialogue in particular, and the recall of supporting details for long dialogue. Qualitative findings revealed that the participants' interaction with the video input varied with their language proficiency and the visual-verbal relationship in the input, suggesting the threshold of language proficiency for the visuals to take effect and the intervening effect of visual input and task features.

Given the limitation of the within-subjects design used in this study, it is suggested that the future research could either provide a longer interval between the audio and video presentations or alternate the presentation order (audio-first and video-first), to better control for the memory or order effect. Nevertheless, this study has shed some light on the influence of visuals on listening performance. There are three implications as follows.

Firstly, considering listening as a communicative activity which involves interpretation of the intricate interplay of verbal, paralinguistic and visual information (Suvorov, 2009), it is necessary to incorporate these non-verbal information into the construct for the assessment of real-world listening ability (e.g., Li, 2013; Wagner, 2008). However, the audio-mode listening test is still useful when the ability to understand verbal language purely is the focus. Meanwhile, to further probe into the effect of video on listening performance, attention could be paid to not only the visual-verbal relationship in the video input, the specific listening task assigned, and the language proficiency of students, but also how students make use of the specific video input and whether it facilitates or inhibits listening comprehension.

Secondly, as to listening instruction, considering the pervasiveness of visuals in the modern hi-tech electronic era, training of visual-related skills is not only a need for students to cope with listening comprehension but also a necessity to engage capably and participate fully in a visual culture. In this sense, it is necessary for language teachers to treat visual input not just as a physical medium but as a mode which generates meaning as much as the audio mode, and thus to develop students' visual literacy by teaching them how to process visuals based on visual grammar and multimodal discourse analysis.

Thirdly, with respect to methodology, considering the fact that the mixed-method research has been established as the third paradigm since 1990s, it is high time that more researchers in this field could apply mixed methods to the investigation of the complex construct of listening comprehension with visuals. Jang, Wagner, and Park (2014) appealed for more diversified and integrated use of MMR design rather than the

simple juxtaposition of quantitative and qualitative methods. For example, various MMR designs (e.g., expansion, development and initiation) (Greene, Caracelli, & Graham, 1989) can be used for different research purposes. In all, the above implications point out directions for future research in this area.

References

ALA (2011). *ACRL visual literacy competency standards for higher education*. Retrieved from http://www.ala.org/acrl/standards/visualliteracy

Bachman, L. F. & Palmer, A. S. (2010). *Language assessment in practic*e. Oxford: Oxford University Press.

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*(1), 3-20.

Brett, P. (1997). A comparative study of the effects of the use of multimedia on listening comprehension. *System, 25*(1), 39-53.

Chapelle, C. (2002). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). Beijing, China: Foreign Language Teaching and Research Press.

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

Cross, J. (2009). Effects of listening strategy instruction on news videotext comprehension. *Language Teaching Research, 13*(2), 151-176.

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

Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of an L2 listening comprehension construct: A tentative model for test specification and development. *The Modern Language Journal, 77*(2), 180-191.

Feak, C. B. & Salehzadeh, J. (2001). Challenges and issues in developing an EAP video listening placement assessment: A view from one program. *English for Specific Purposes, 20,* 477-493.

Field, J. (2008). Bricks or mortar: Which parts of the input does a second language listener rely on? *TESOL Quarterly, 42,* 411-432.

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing, 19*(2), 133-167.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11,* 255-274.

Grimes, T. (1991). Mild auditory-visual dissonance in television news may exceed viewer attentional capacity. *Human Communication Research, 18*(2), 268-298.

Gruba, P. (2004). Understanding digitized second language videotext. *Computer Assisted Language Learning, 17*(1), 51-82.

Gruba, P. (2006). Playing the videotext: A media literacy perspective on video-mediated L2 listening. *Language Learning & Technology, 10*(2), 77-92.

Jang, E. E., Wagner, M., & Park, G. (2014). Mixed methods research in language testing and assessment. *Annual Review of Applied Linguistics, 34,* 123-153.

Kress, G. R., & van Leeuwen, T. (2006). *Reading images: The grammar of visual design*. London, UK: Routledge.

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

Lesnov, R. O. (2018). Content-rich versus content-deficient video-based visuals in L2 academic listening tests: pilot study. *International Journal of Computer-Assisted Language Learning and Teaching, 8*(1), 15-30.

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*(2), 61-82.

Mohamadkhani, K., Farokhi, E. N., & Farokhi, H. N. (2013). The effect of using audio files on improving listening comprehension. *International Journal of Learning and Development, 3*(1), 132-137.

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing, 24*(4), 517-537.

O'Halloran, K. (2004). *Multimodal discourse analysis: Systemic functional perspectives*. London, UK: Continuum.

Paivio, A. (1986). *Mental representations: A dual-coding approach*. Oxford, UK: Oxford University Press.

Pardo-Ballester, C. (2016). Using video in web-based listening tests. *Journal of New Approaches in Educational Research, 5*(2), 91-98.

Progosh, D. (1996). Using video for listening assessment: Opinions of test-takers. *TESL Canada Journal, 14*(1), 34-44.

Pusey, K. & Lenz, K. (2014). Investigating the interaction of visual input, working memory, and listening comprehension. *Language Education in Asia, 5*(1), 66-80.

Rost, M. (1990). *Listening in language learning*. London, England: Longman.

Rubin, J. (1994). A review of second language listening comprehension research. *The Modern Language Journal, 78*(2), 199-221.

Schriver, K. A. (1997). *Dynamics in document design: Creating text for readers*. New York, NY: John Wiley & Sons, Inc.

Seo, K. (2002). Research Note: the effect of visuals on listening comprehension: A study of Japanese learners' listening strategies. *International Journal of Listening, 16*(1), 57-81.

Shang, H. F. (2005). An investigation of cognitive operations on L2 listening comprehension performance: An exploratory study. *International Journal of Listening, 19*(1), 51-62.

Shin, D. (1998). Using videotaped lectures for testing academic listening proficiency. *International Journal of Listening, 12*(1), 57-80.

Stokes, S. (2002). Visual literacy in teaching and learning: A literature perspective. *Electronic Journal for the integration of Technology in Education, 1*(1), 10-19.

Sueyoshi, A. & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning, 55*(4), 661-699.

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

Suvorov, R. (2015). The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos. *Language Testing, 32*(4), 1-21.

Wagner, E. (2002). Video listening tests: A pilot study. *Working Papers in TESOL & Applied Linguistics, 2*(1), 1-39.

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning & Technology, 11*(1), 67-86.

Wagner, E. (2008). Video listening tests: what are they measuring? *Language Assessment Quarterly, 5*(3), 218-243.

Wagner, E. (2010a). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing, 27*(4). 493-513.

Wagner, E. (2010b). Test-takers' interaction with an L2 video listening test. *System, 38*(2), 280-291.

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

Zheng, Y., & Cheng, L. (2008). College English Test (CET) in China. *Language Testing, 25*(3), 408-417.
