Exploring the Feasibility of Video-Mediated Listening Test in a Nation-wide Proficiency English Examination in China

Ying Xiao*
Fudan University, Shanghai, China

Zheng-liu Liang
Fudan University, Shanghai, China

Qiang Li
National Medical Examination Center, Beijing, China

Ruo-jun Jia
National Medical Examination Center, Beijing, China

Abstract

Thanks to the increased accessibility of multimedia technology, video texts have been widely used in teaching English in Chinese foreign language classrooms. This trend provides great impetus for changes in the means of assessing listening in both achievement and proficiency tests, and suggests the potential possibility of supplanting the audio-only with a video-mediated listening test. This paper aims to explore such a possibility in the Foreign Language Admission Test for Medical Doctoral Students (FATMD) administered by the National Medical Examination Center in China. A total of 148 MD and PhD students who had just passed the equivalent of a language admissions test voluntarily enrolled. Participants were divided into an experiment group for a video-mediated listening test and a control group for an audio-only test. The results indicated no significant differences on average scores between the video-mediated and audio-only group. However, they did suggest the potential of supplanting the audio-only listening test with a video-mediated equivalent in FATMD without upsetting its well-established credibility and reliability, in light of the increasing use of computer-assisted English proficiency testing in China.

Keywords: English proficiency test, listening test, audio-mediated, video-mediated, EFL, ESP

Introduction

In the English-as-a-foreign-language classroom, mediums of teaching are evolving, which has much to do with rapid technological changes. In China, due to the increasing prevalence of multimedia technology, video has been playing an increasing role in language teaching and learning at different levels. The wider use of video texts has been reflected in learners' positive attitudes and in pedagogical effects in classrooms across the country, which echoes the findings from previous research on the extensive use of video texts as the medium of instruction (Baltova, 1994; Dunkel, 1991; Parry and Meredith, 1984; Progosh, 1996; Sueyoshi & Hardison, 2005; Wagner, 2002). This medium change in the classroom has also exerted an impact on the medium of delivering the listening comprehension sections on an achievement test. Consequently, video texts have come to take the place of audio listening tests, as illustrated by previous research (Baltova, 1994; Gruba, 1993; Parry & Meredith, 1984; Progosh, 1996; Sueyoshi & Hardison, 2005; Wagner, 2010, 2013).

* *Tel.: (+86) 21-65648523. E-mail: yxiao471247@fudan.edu.cn. Address: College of Foreign Languages and Literatures, Fudan University, 220 Handan Rd., Shanghai, 200433, People's Republic of China*

Inevitably, the question is, what is the difference between a video-mediated listening comprehension test compared to an audio-only test? The present study seeks to investigate whether test-takers would demonstrate different proficiencies on the two listening tests, audio-only and video-mediated, on the nation-wide Foreign Language Admission Test for Medical Doctoral Students (FATMD) administered by National Medical Examination Center (NMEC) in China, and explore the feasibility of supplanting the former with the latter.

Audio- and Video-Mediated Listening Tests

A review of literature on the issue revealed contradicting evidence. There has been ample evidence that listening comprehension is facilitated by video-mediated texts—e.g., the increased comprehension provided through visual information when paired with an explanation for students with little prior knowledge of a topic (Mayer, 1997); significantly better performance in a video-mediated version of a Spanish listening test compared to an audio-only test because the test-takers had more interest and greater motivation to pay attention (Parry & Meredith, 1984); enhanced general comprehension through informative visual cues as well as the affective and attention advantages (Baltova, 1994); examinees' reported preference for video-mediated listening comprehension tests (Baltova, 1994; Progosh, 1996; Sueyoshi & Hardison, 2005); and the greater ease of taking video-mediated listening tests compared to audio-only counterparts (Baltova, 1994; Brett, 1997; Hernandez, 2004; Parry & Meredith, 1984; Shin, 1998; Sueyoshi & Hardison, 2005; Wagner, 2010, 2013).

On the other hand, much evidence has also been reported on the insignificant difference and even contradictory or conflicting results in video-mediated listening tests compared to audio-only ones (Baltova, 1994; Brett, 1997; Coniam, 2001; Cubilo & Winke, 2013; Gruba, 1993, 1997; Kellerman, 1990, 1992; Ockey, 2007; Progosh, 1996; Shin, 1998; Sueyoshi & Hardison, 2005; Suvorov, 2013). Gruba (1993) compared the performances of university students who took an audio-only test and who took a context-only video version of the same test and found no difference in their test performance. A case study done by Coniam (2001) revealed a similar finding, there were no significant differences between the audio and video testing-taking groups. Coniam also noted that the video test-taking group gained no advantage from the video mode and some test takers even reported being distracted by the visual images.

Despite the divergent evidence on this issue and the fact that video text is much less used than audio-only in assessing second language (L2) listening ability, more research is needed to gauge the value of videos in listening assessments. As previously reported, more authentic input for L2 listeners, along with the assistance of such non-verbal elements as gestures, facial expressions and other kinesic behaviors could contribute to listening comprehension (Baltova, 1994; Gruba, 1997; Kellerman, 1990; Progosh, 1996; Wagner, 2010). Gruba (1997) has provided us with four reasons for using video texts in listening assessment. First, the use of video in listening assessment is theory-driven because according to models of language comprehension, real-life communication in most cases involves both visual and verbal elements. Listeners cannot help receiving extra-linguistic signals from speakers as well as from their surroundings; consequently, the verbal and visual signals form a complementary *co-text* for listening comprehension (Rost, 2002). This gets test developers to look beyond the audio tape as a mode of presentation for authenticity, and face validity can be improved in the video-mediated listening tests. Next, using video texts in listening assessment is pedagogy-related, and the widespread use of visual aids for a variety of pedagogical reasons calls for the integration of improved video-mediated assessment practices. Third, video is used in listening assessment because "there are features of the process, or setting, of how the language is being used which cannot be separated from its meaning" (Gruba, 1997, p. 339), as indicated by some *language specific* medical settings where the use of video media can help viewers understand some proper medical practices. Additionally, the use of video in listening assessment results from the mode of delivery, i.e., the increasing use of videos in distance learning programs, correspondence courses, and video conferencing raises the possibility of video-mediated listening tests.

In recent years, we have seen the rise of a preference for the use of videos in classroom instruction at colleges and universities in China (Zhou & Yang, 2004), which could potentially affect the delivery of nation-wide English proficiency tests in China. However, there is a dearth of literature on comparisons of video-

mediated and audio-only listening tests in terms of the acceptability, reliability, and validity of such a large-scale English proficiency test in China.

The current research intends to explore the feasibility of supplanting the audio-only listening test with video-mediated listening on the FATMD administered by the NMEC in China. Established in 1997, the FATMD is offered in English, Japanese, and Russian to those who apply for PhD and MD positions at medical colleges, with the English test-takers accounting for 99.9% of approximately 15,000 applicants. In such a large-scale high-stakes language test, it is vital that the FATMD scores reflect test-takers' language proficiency, particularly their communicative competence, the ability to communicate in a medical setting (National Medical Examination Center, 2009). Bachman and Palmer (1996) argued:

> If the scores from a language test are used to make inferences about individuals' language ability, and possibly to make various types of decisions, it is necessary to demonstrate how performance on that language test is related to language use in specific situations other than the language test itself. (p. 10)

For FATMD, to ensure that the test provides an accurate indication of communicative language ability, it is necessary to include the ability to use the language in specific situations, including the understanding of non-verbal or subtle visual information, which is a part of communicative competence (Bachman, 1990). Without such consideration, the validity of the listening test tasks could be compromised. It must be noted that audio media is commonly used for language listening tests, which might be explained as a legacy of limited technology. However, high-stakes decisions require a high-quality test, which demands considerable time, effort, and resources, and the available resources are crucial as they provide one basis of determining the success of a test's development (Buck, 2001). Even though digital videos are readily accessible to language listening test developers, the question remains how they can be used to improve the quality of such a large-scale and high-stakes test.

Research Questions

Guided by previously reported studies and based on the findings from our previous study on FATMD candidates' visual perceptions and subjective acceptance of the video-mediated listening test (Sun, Xiao, Liang, Li, & Jia, 2015), this study focuses on the feasibility of supplanting audio-only with video-mediated listening tests in the FATMD. To be specific, we intend to address two major research questions:

1. Is there any significant performance difference between the test-takers on a video-mediated listening test and their counterparts on the audio-only one?
2. What factors can affect the performance of the video group in comparison with the audio group?

Methods

Test Design

Based on the test developing guidelines for the FATMD, many video clips from medicine/health-related websites were downloaded and cut into video segments, from which 15 short conversations, 1 long conversation, and 2 monologue passages were chosen to form a simulated FATMD listening comprehension test. As in the case of the audio-only version of this listening test, one multiple choice question was set for each video-mediated short conversation; five for a long video-mediated conversation; and five for each one of the video-mediated monologue passages. Based on the video and their transcripts, all the test questions were developed and discussed before being decided on by all the researchers. The clips were edited into one complete test video following the layout of the FATMD audio test by an experienced teacher of English with 18 years of concurrent working experience for NMEC. The directions (see Table 1) and questions were audio-recorded in advance by two

experienced native speakers, who work for Shanghai Foreign Language Education Press and were later edited into the two sets of listening tests.

Table 1
Comparison of the Directions for the Video-mediated and Audio-only Test

For video-mediated	For audio-only
Section A *Directions*: In this section, you will watch 15 short video clips. In each video, there is a short conversation between two speakers. At the end of each conversation, you will hear a question about what has been said. For each question, you can read four possible answer choices marked A, B, C, and D on the screen *before* and *after* the video has been played. Listen to the question carefully, and choose the best answer by ticking the corresponding letter on the answer sheet.	*Section A* *Directions*: In this section, you will hear 15 short conversations between two speakers. At the end of each conversation, you will hear a question about what has been said. The question will be read only once. After you hear the question, read the four possible answer choices marked A, B, C, and D. Choose the best answer by ticking the corresponding letter.
- Here is an example. Read quickly the four suggested answer choices marked A, B, C and D on the screen. Now watch the video and listen carefully. - Read the choices again and listen to the following question: *What can be the problem with a barbeque dinner according to the man?* - Tick the corresponding letter on the answer sheet before you do the next. - Now let's begin.	(An example was not included because students are quite familiar with this testing style in China)
Section B *Directions:* In this section, you will watch three video clips. In one, there is a long conversation, and in the rest, there are two passages. After each one, you will hear five questions. For each question, you can read four possible answer choices marked A, B, C, and D on the screen *before* and *after* the video has been played. Listen to the question carefully, and choose the best answer by ticking the corresponding letter on the answer sheet.	*Section B* *Directions*: In this section, you will hear one long conversation and two passages. After each one, you will hear five questions. After each question, read the four possible answer choices marked A, B, C, and D. Choose the best answer by ticking the corresponding letter.

After the choices were laid out and clips were pieced together into a complete test video, two sets of listening tests were developed. To better reflect face-to-face communication and real-life language tasks, questions and choices were not written on the test paper of the video-mediated group based on Wagner's (2013) findings that "having access to the MC questions while the text is played does not lead to increased test performance" (p. 191). In the video-mediated test, choices were projected on a screen before and after the test video had been played. Participants were supposed to mark the best answer on the answer sheet with question numbers, each followed by four letters (A, B, C, and D) representing four possible answer choices. In the audio-only test, while listening to the audio recording converted from the test video, the control group worked with a

test paper presenting directions, question numbers, and four word choices for each question. Following the designed workflow (see Table 2), the tests were administered.

Table 2
Workflow of the Video-mediated and Audio-only Test

Section A		
For the video-mediated version		
A PPT slide showing four multiple choices on a big screen, lasting 3 seconds	A short video-mediated face-to-face conversation between a man and a woman[1]	An audio-only question for the four possible answer choices on the screen, followed by a pause of 7 seconds, during which the subjects are supposed to read the choices marked A, B, C, and D and choose the best one by ticking the corresponding letter on the answer sheet
For the audio-only version		
A short audio-only conversation between a man and a woman, with the four suggested answer choices to read on the controls' test paper[2]	An audio-only question, followed by a pause of 10 seconds, during which the controls were supposed to read the four possible answer choices marked A, B, C, and D and choose the best one by ticking the corresponding letter on the test paper	
Section B		
For the video-mediated version		
Five PPT slides for each video, each showing four answer choices on a big screen, lasting 15 seconds (3x5)	1) A long face-to-face video-mediated conversation[3] 2) A video-mediated passage delivered by a woman (above the waist), with a few meaningful still pictures[4] 3) A video-mediated passage delivered by a man behind the meaningful settings[5]	Five audio questions to be asked one by one, each followed by a pause of 7 seconds, during which the subjects were supposed to read on the screen four possible answer choices marked A, B, C, and D and choose the best one by ticking the corresponding letter on the answer sheet
For the audio-only version		
1)A long audio-only conversation 2) An audio-only passage delivered by a woman 3) An audio-only passage delivered by a man	Following each audio text, five audio-only questions to be asked one by one, each followed by a pause of 10 seconds, during which the controls were supposed to read the four possible answer choices marked A, B, C, and D and choose the best one by ticking the corresponding letter on the test paper	

Notes.
[1] Fifteen short video-mediated conversations on diabetes, hair, stroke, hypertension, multi-vitamins, cancer, online medical information, AIDS, antibiotics, diet, chemicals, online teenagers, dynamic communication, patient and physician and weight loss, respectively
[2] The same 15 conversations but made audio-only
[3] On a medical myth: alcoholic drink and sleep
[4] On chronic stress

[5] On diet and Alzheimer's disease

Pre-Test

Before the implementation of the pre-test, five teachers, each with more than five years of teaching medical English and developing achievement tests, were randomly chosen for a video-mediated pre-test. After the test, they were encouraged to provide insightful and useful feedback on how to choose testable video-mediated texts and how to design the corresponding questions, especially on how to match the current difficulty level with that of the well-recognized FATMD. They suggested choosing clips with higher resolution, controlling the background noise, making the choices of questions shorter, and including more questions covering non-verbal clues. Based on their feedback, the video clips were finalized, replacing some low-resolution clips and questions were improved with shorter choices without changing the difficulty level in relation to the FATMD.

Inclusion and Exclusion Standards

A total of 148 students voluntarily signed up for the current study after the study notice was handed out among the target 400+ MD and PhD students at the Shanghai Medical Center of Fudan University in Shanghai. Those who were of ethnic minority origin and English native speakers were excluded, a total of 3 students, because according to the school policy, they were exempted from the school's English entrance examination. Consequently, the video-mediated group consisted of 73 participants, and the audio-only group, 72 participants. The students were full-time doctoral freshmen who had just passed the school's English entrance examination, almost similar in age to the potential test-takers of the FATMD. Currently, they were taking an advanced course on medical English as a required course. After being selected for the study, they were divided, based on the last odd/even figure of their student ID numbers, into subject and control groups.

Testing Implementation

Two weeks before the study, handouts regarding its schedule were distributed during the break of medical English classes for doctoral students, followed by a brief explanation on participation, liability, and reward. After that, the handouts were collected from those who were willing to participate in the listening test.

The listening test was scheduled to be given when the participants were available after their English classes. When the classes were over, they were asked to come into one of the classrooms to be divided into the subject and control groups according to the last odd/even figure of their ID numbers; the former were guided into another classroom for the video-mediated listening test, and the latter stayed for the audio-only one. At the same time, simple answer sheets were handed out to the subjects, while multiple-choice test papers were provided to the control group before the start of the tests. For both groups, the listening test lasted half an hour.

A total number of 145 submissions, 73 answer sheets and 72 test papers, were collected, with an effective response rate of 100%.

Data Management

A data bank was established using Epidata3.1 based on the input of the answers from both the video-mediated and audio-only tests by two statisticians. The data went through a consistency check, and exhibiting no difference, they were exported to DTA file format. The data were re-checked using SAS9.2.

Data Analysis

A data analysis was performed using SPSS17.0. The frequency distributions were established based on the number of correct answers from the lowest to the highest scores, and from the score percentage to the accumulative percentage of the two groups.

Unless otherwise specified, a *t*-test was used for the comparison of the total scores on average between the two groups with $P< 0.05$ considered to be significantly different and $P > 0.05$ to have no significant difference. Non-inferiority testing was approached by their difference value based on the minimum fractional differential (non-inferiority margin) of 2 points, which is formally practiced in the school's test scoring; therefore, the difference value by less than 2 points ($P< 0.05$) meant no significant difference in listening comprehension between the video-mediated and audio-only group, and the difference value by no less than 2 points ($P> 0.05$) meant the opposite.

A comparison was made of each individual average score between the two groups via *t*-test with $P< 0.05$ and $P> 0.05$ used to indicate the different results, respectively. The correct and incorrect answers to each of 30 listening questions were addressed using Chi-square test to observe the individual difference between the test and control group with $P< 0.05$ indicating the different scoring accuracy of each question, respectively.

Results

Frequency Distribution

The 30 multiple-choice listening comprehension questions exhibited typical scores of a normal distribution as a whole (n = 145; mean: 15.94; SD: 3.264), for both the video-mediated group (n = 73; highest: 24; lowest: 9; mean: 15.49; SD: 3.396), and the audio-only (n = 72; highest: 24; lowest: 11; mean: 16.39; SD: 3.084) (see Table 3 and Figure 1), based on a *t*-test applied to the various comparisons of mean.

Table 3
Score Frequency Summary

Scores	Audio-only Group			Video-mediated Group			Total Score		
	No	%	Total %	No	%	Total %	No	%	Total %
9	0	0.0	0.0	2	2.7	2.7	2	1.4	1.4
10	0	0.0	0.0	2	2.7	5.5	2	1.4	2.8
11	2	2.8	2.8	4	5.5	11.0	6	4.1	6.9
12	4	5.6	8.3	6	8.2	19.2	10	6.9	13.8
13	8	11.1	19.4	10	13.7	32.9	18	12.4	26.2
14	9	12.5	31.9	7	9.6	42.5	16	11.0	37.2
15	6	8.3	40.3	7	9.6	52.1	13	9.0	46.2
16	12	16.7	56.9	5	6.8	58.9	17	11.7	57.9
17	7	9.7	66.7	12	16.4	75.3	19	13.1	71.0
18	5	6.9	73.6	4	5.5	80.8	9	6.2	77.2
19	7	9.7	83.3	5	6.8	87.7	12	8.3	85.5
20	5	6.9	90.3	4	5.5	93.2	9	6.2	91.7
21	3	4.2	94.4	0	0.0	93.2	3	2.1	93.8
22	1	1.4	95.8	3	4.1	97.3	4	2.8	96.6
23	1	1.4	97.2	1	1.4	98.6	2	1.4	97.9
24	2	2.8	100.0	1	1.4	100.0	3	2.1	100.0
Total	72	100.0		73	100.0		145	100.0	

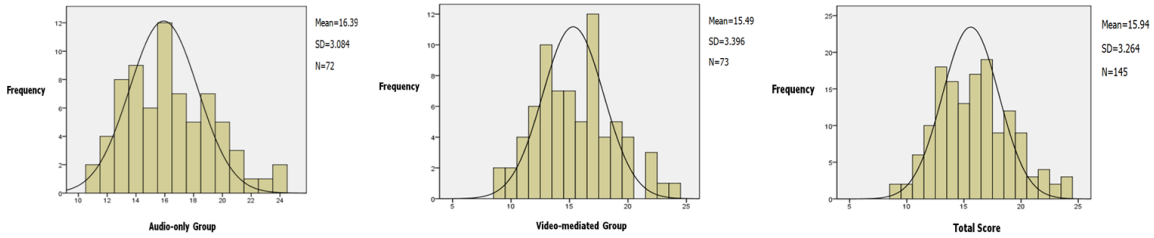

Figure 1. Frequency Distribution in Video-mediated Group, in Audio-only Group and as a Whole

Population Mean

The 30 listening comprehension questions were hierarchically scaled from easy to medium to difficult as required by the FATMD. As indicated by the population variance test, F-value and significance were 0.885 and 0.348 in the total number of 30 questions; 0.788 and 0.376 for the 8 easy questions; 0.660 and 0.418 from the 13 medium ones; and 0.046 and 0.831 for the 9 difficult ones (see Table 4).

Table 4
Population Mean

	No	Mean	SD	T-value	Df	Sig	95%CI
30 Qs in Total							
Audio-only	72	16.39	3.084	1.662	143	0.099	-0.170-1.961
Video-mediated	73	15.49	3.396				
8 Easy Questions							
Audio-only	72	5.04	1.272	1.507	143	0.134	-0.098-0.730
Video-mediated	73	4.73	1.250				
13 Medium Questions							
Audio-only	72	7.71	2.024	0.632	143	0.528	-0.487-0.944
Video-mediated	73	7.48	2.322				
9 Dif. Questions							
Audio-only	72	3.64	1.367	1.546	143	0.124	-0.098-0.800
Video-mediated	73	3.29	1.369				

Non-Inferiority Test

The non-inferiority test was performed with reference to the smallest fractional differential (non-inferiority margin), a score of 2, which was adopted on the English test at Fudan University—e.g., to decide between the grades B and B-. The test results indicated that the difference value was less than a score of 2, a small probability event.

As indicated in Table 5, no significant difference was observed between the video-mediated and audio-mediated group in terms of average score (non-inferiority test adjusted $P > 0.05$). The difference value of 0.90 between the experiment and control group was much less than a score of 2.

Table 5
Non-inferiority Test of Population Mean

	No	Mean	SD	Group Difference	D-value	df	Sig	95%CI
Audio-only Lowest score: 11 Highest score: 24	72	16.39	3.084	0.9	0.896	143	0.042	-0.1961-0.170

Video-mediated			
Lowest score: 9	73	15.49	3.369
Highest score: 24			

Single Question Data

As indicated by most of the answers in Section II (see Table 6), the number of incorrect answers outnumbered correct ones. And only in question 26, the number of correct answers was significantly lower in the video-mediated than in the audio-only group ($P<0.05$).

Table 6
Single Questions

Q #	Audio-only Correct		Wrong		Video-mediated Correct		Wrong		Total Correct		Wrong		P-value[9]
	No	%	No	%	No	%	No	%	No	%	No	%	
Section I: Question 1-15													
1 (e)[1]	70	97.2	2	2.8	71	97.3	2	2.7	141	97.2	4	2.8	0.989
2 (m)[2]	55	76.4	17	23.6	54	74.0	19	26.0	109	75.2	36	24.8	0.736
3 (e)	48	66.7	24	33.3	50	68.5	23	31.5	98	67.6	47	32.4	0.814
4 (e)	64	88.9	8	11.1	57	78.1	16	21.9	121	83.4	24	16.6	0.080
5 (m)	55	76.4	17	23.6	58	79.5	15	20.5	113	77.9	32	22.1	0.657
6 (d)[3]	46	63.9	26	36.1	38	52.1	35	47.9	84	57.9	61	42.1	0.149
7 (m)	45	62.5	27	37.5	34	46.6	39	53.4	79	54.5	66	45.5	0.050
8 (m)	47	65.3	25	34.7	49	67.1	24	32.9	96	66.2	49	33.8	0.814
9 (e)	63	87.5	9	12.5	56	76.7	17	23.3	119	82.1	26	17.9	0.090
10 (m)	50	69.4	22	30.6	52	71.2	21	28.8	102	70.3	43	29.7	0.814
11 (e)	55	76.4	17	23.6	59	80.8	14	19.2	114	78.6	31	21.4	0.515
12 (d)	*31*	*43.1*	*41*	*56.9*	*26*	*35.6*	*47*	*64.4*	*57*	*39.3*	*88*	*60.7*	0.359
13 (d)	*17*	*23.6*	*55*	*76.4*	*26*	*35.6*	*47*	*64.4*	*43*	*29.7*	*102*	*70.3*	0.114
14 (d)	*10*	*13.9*	*62*	*86.1*	*15*	*20.5*	*58*	*79.5*	*25*	*17.2*	*120*	*82.8*	0.289
15 (d)	39	54.2	33	45.8	*32*	*43.8*	*41*	*56.2*	*71*	*49.0*	*74*	*51.0*	0.213
Section II: Question 16-30													
16 (m)	44	61.1	28	38.9	39	53.4	34	46.6	83	57.2	62	42.8	0.350
17 (d)	51	70.8	21	29.2	48	65.8	25	34.2	99	68.3	46	31.7	0.511
18 (m)	55	76.4	17	23.6	57	78.1	16	21.9	112	77.2	33	22.8	0.808
19 (m)	*33*	*45.8*	*39*	*54.2*	42	57.5	31	42.7	75	51.7	70	48.3	0.159
20 (d)	*20*	*27.8*	*52*	*72.2*	*15*	*20.5*	*58*	*79.5*	*35*	*24.1*	*110*	*75.9*	0.309
21(e)	*6*	*8.3*	*66*	*91.7*	*6*	*8.2*	*67*	*91.8*	*12*	*8.3*	*133*	*91.7*	0.980
22 (m)	*32*	*44.4*	*40*	*55.6*	*22*	*30.1*	*51*	*69.9*	*54*	*37.2*	*91*	*62.8*	0.075
23 (e)	*16*	*22.2*	*56*	*77.8*	*19*	*26.0*	*54*	*74.0*	*35*	*24.1*	*110*	*75.9*	0.592
24 (d)	*25*	*34.7*	*47*	*65.3*	*22*	*30.1*	*51*	*69.9*	*47*	*32.4*	*98*	*67.6*	0.555
25 (m)	59	81.9	13	18.1	52	71.2	21	28.8	111	76.6	34	23.4	0.128
26 (e)	41	56.9	31	43.1	*27*	*37.0*	*46*	*63.0*	*68*	*46.9*	*77*	*53.1*	0.016
27 (d)	*23*	*31.9*	*49*	*68.1*	*18*	*24.7*	*55*	*75.3*	*41*	*28.3*	*104*	*71.7*	0.330
28 (m)	*35*	*48.6*	*37*	*51.4*	*32*	*43.8*	*41*	*56.2*	*67*	*46.2*	*78*	*53.8*	0.564
29 (m)	*31*	*43.1*	*41*	*56.9*	40	54.8	33	45.2	*71*	*49.0*	*74*	*51.0*	0.157
30 (m)	*14*	*19.4*	*58*	*80.6*	*15*	*20.5*	*58*	*79.5*	*29*	*20.0*	*116*	*80.0*	0.868

Notes.
[1] e: stands for easy
[2] m: stands for medium
[3] d: stands for difficult

Section Data

According to the results of the homogeneity of variance test, F-value and significance were 0.073 and 0.788, respectively, for all the questions in Section I (see Table 7); 0.821and 0.366, respectively, for the easy questions; 0.966 and 0.327, respectively, for the medium questions; and 0.709 and 0.401, respectively, for the difficult questions.

In all the questions in Section II (see Table 8), F-value and significance were 0.023 and 0.881, respectively; for the easy questions, 0.082 and 0.775, respectively; for the medium questions, 2.665 and 0.105, respectively; and for the difficult questions, 0.709 and 0.401, respectively.

Table 7
Data of Section I

	No	Mean	SD	T-value	Df	Sig	95%CI
All Qs							
Audio-only	72	9.65	2.022	1.128	143	0.261	-0.285-1.043
Video-mediated	73	9.27	2.023				
Easy Qs							
Audio-only	72	4.17	0.822	1.045	143	0.298	-0.136-0.442
Video-mediated	73	4.01	0.935				
Medium Qs							
Audio-only	72	3.50	1.222	0.591	143	0.556	-0.273-0.506
Video-mediated	73	3.38	1.150				
Dif. Qs							
Audio-only	72	199	1.204	0.582	143	0.561	-0.262-0.481
Video-mediated	73	1.88	1.053				

Table 8
Data of Section II

	No	Mean	SD	T-value	Df	Sig	95%CI
All Qs							
Audio-only	72	6.74	1.986	1.524	143	0.130	-0.154-1.187
Video-mediated	73	6.22	2.097				
Easy Qs							

Audio-only	72	0.88	0.786	1.303	143	0.195	-0.084-0.410
Video-mediated	73	0.71	0.716				
Medium Qs							
Audio-only	72	4.21	1.352	0.439	143	0.661	-0.394-0.619
Video-mediated	73	4.10	1.709				
Dif. Qs							
Audio-only	72	1.99	1.204	0.582	143	0.561	-0.262-0.481
Video-mediated	73	1.88	1.053				

Data of Score Groups

As indicated by Table 9, the total score of 30 was accordingly divided into four ranges of correct scoring percentages for the Chi square test: below 40%, 40%-60%, 60%-80%, and above 80%. It showed that the distribution of correct scoring was not consistent between the video-mediated and audio-only group, as indicated in the 1st and 4th range (P=0.05).

The results of the *t*-test showed that the average score was significantly higher in the audio-only than in the video-mediated group (P=0.048<0.05) (see Table 10).

Table 9
Data of Score Groups

Score groups	Audio-only		Video-mediated		In total		P-value[1]
	No	%	No	%	No	%	
Lowest-11	2	2.8	8	11.0	10	6.9	
12-18	51	70.8	51	69.9	102	70.3	
19-23	17	23.6	13	17.8	30	20.7	0.050
24-highest	2	2.8	1	1.4	3	2.1	
Total	72	100.0	73	100.0	145	100.0	

Note.
[1] p-values calculated via Chi-Square test

Table 10
Score Group Comparisons

Score Groups	No	Mean	SD	T-value	Df	Sig	95%CI
Lowest-11							
Audio-only	2	11.00	0.000	2.393	7.000	0.048	0.009-1.491
Video-mediated	8	10.25	0.886				
12-18							
Audio-only	51	15.08	1.798	0.423	100	0.674	-0.580-0.893
Video-mediated	51	14.92	1.948				
19-23							

Audio-only	17	20.06	1.197	-0.517	28	0.609	-1.234-0.736
Video-mediated	13	20.31	1.437				
24-highest							
Audio-only	2	24.00	0.000	N/A	1	0.000	0.000-0.000
Video-mediated	1	24.00					

Discussion

The results of the non-inferiority test on the mean showed that results of the video-mediated and audio group showed no statistically significant difference (Table 5; P<0.042), which was in line with previous research findings: the Educational Testing Service (ETS) reporting no main effect for the presence or absence of visuals on TOEFL (Ginther, 2002); Coniam (2001) found no statistically significant difference between the video-mediated and audio-only groups in his study involving 104 Hong Kong English language teachers in a post-graduate diploma program; and Gruba (1993) found no difference in the scores of the two versions of a test on listening about air traffic safety for two intact classes of ESL students studying at a US university. However, it was noted that the average score was 0.90 lower in the video-mediated than in the audio-only group (see Table 5), which could be ascribed to the random sampling and the lack of training for the new mode of testing on the part of the video-mediated group.

Based on the data for the individual questions, statistically significant difference was observed between the video-mediated and audio-only group on Q26 (P<0.016), which had been subjectively defined as easy with reference to the question difficulty scale for the FATMD. Such a difference could be explained by the possibility that the visual components, which were interactional (i.e., conversational) and transactional (e.g., a lecture) (Buck, 2001), did nothing but hinder listening comprehension (Gruba, 1999), or distracted attention away from the aural input (MacWilliam, 1986). Another explanation could be that the four written choices to be read on the screen, or the spoken question which followed the video listening were hard to understand on the part of the video-mediated group, regardless of what they had watched. The video-mediated group had no chance to read the four written choices while watching and listening, while the audio-only group listened with their eyes busy reading the same choices on paper from which they might pick up some hints of well-targeted information for their predictions and decisions; thus, recalling the question-related information was more demanding for the former than the latter, even though both were allowed the same amount of time for a question.

Such explanations could be applied to the results indicated by Table 10, where significant differences were observed between the audio-only and video-mediated group, as indicated by 2 vs. 8 persons and 2 vs. 1 person in the score ranges of lowest-11 and 24-highest, respectively (P=0.048<0.05). The difference suggested that the video-mediated text could be more demanding than its counterpart in terms of the listening comprehension test because it deprived the subjects of the chance to bridge what they heard and what they could read while watching and listening.

Since 1997, the FATMD has been audio-only on the listening comprehension test; consequently, test-takers have grown accustomed to the format. Viewed from this standpoint, it should be noted that the current study reflects an important issue of habitual adaptation on the part of the participants. Even though they were given the instructions at length and guided with an illustration prior to the experiment, the video-mediated group could not have physically and psychologically prepared for or adapted to a new format of the listening comprehension test within such a short period, suggesting that unfamiliarity may have undermined their scoring to some extent. Viewed from this point of view, it is argued that once the FATMD is made video-mediated, it is anticipated that future test-takers will develop a familiarity, which will be beneficial to their testing performance, as indicated by Wagner's (2010) study where the video group and audio-only group achieved similar scores on the pre-test, while in the post-test the experimental group scored 6.5% higher than the control group.

Further examination of the significant difference observed among the lower achievers from the audio-only and video-mediated group; our findings showed only 2 candidates in the lowest scoring range from the audio-only group, compared to 8 from the video-mediated group (Table 10; $P=0.048<0.05$). This finding might suggest a relation between the variables of the video-mediated effect and English proficiency level. Lower English proficiency level, which could be an important contributing factor for participants whose results were in the lowest range more so than in others, could be immune to the promising effects of a video-mediated text; in this case, the visuals may have hindered the listening comprehension process.

It is worth noting that in the administration of the test, not all the test-takers were continuously watching the video clips. Some of them just watched a small portion and then focused on taking notes while simply listening to them. However, we cannot say for certain that the video-mediated listening test does not yield the expected results. In real life, the different listening skills are not often used entirely in isolation. This may be interpreted as a normal response to a context where active listening is required, like that of attending a lecture. Ockey (2007) argues for the use of at least some sort of visual stimulus because "most target language-use situations include visual stimuli" (p. 517). In this sense, the video portion is a complement to the audio one, which may set the scene, helping test-takers ease into the test. In further research, more advanced technology is needed to engage the test-takers' with the video, for example, personal computers. More extensive research is needed as the current experimental investigation was conducted in one place and is not representative enough to be generalized to all of China, which is important since the FATMD is a nation-wide proficiency English test.

Implications and Future Research

We can draw the conclusion that no significant difference was observed between the video-mediated and audio-only groups, and that the video-mediated group's performance was affected by two important factors, the presentation format of the test questions and the test-takers' familiarity with the new mode of input. Despite a slightly lower average score for the video-mediated groups, the fact that there was no significant difference between the performance of the video-mediated and the audio-only group suggests that the audio-only FATMD administered by NMEC in China could potentially be switched to a video-mediated version without upsetting its well-established credibility and validity (Yu, 2008). This has been documented in many studies that have reported a tendency towards video-mediated listening tests (Baltova, 1994; Brett, 1997; Hernandez, 2004; Parry & Meredith, 1984; Shin, 1998; Sueyoshi & Hardison, 2005; Wagner, 2013). On the other hand, if the two factors identified above are effectively controlled, it is reasonable to assume that the video-mediated test-takers' performance would improve.

Although audio-only and video-mediated listening tests may produce similar comprehension scores, it seems that the benefits of visual cues to listening comprehension is too important to ignore (Rost, 2002; Rubin, 1990; Sueyoshi & Hardison, 2005), but the research on video-mediated listening tests has turned out conflicting or inconclusive results and test developers seem to avoid video-mediated tests as construct validity is at risk (Bachman, 1990). "Paradoxically, the validity of listening tests that do not consider that most people both hear and see in most communicative situations is just as contentious" (Progosh, 1996, p. 35). Bachman and Palmer (1996), who also lay great stress on the relationship between the language used on the test and the one used in real life situations, argues that "if there is no such relationship, our language tests become mere shadows, sterile procedures that may tell us nothing about the very ability we wish to measure" (p. 350). In the real world, it would be impossible to think of sighted people who can separate listening from seeing, and today even telephone conversations can be visual. Based on the examples of real-world interactions, it is necessary to adapt listening tests to include video texts, if the goal is to have a valid measure of communicative competence in a relevant setting. However, Ockey (2007) warns that the mode of input can affect the construct validity of listening tests because test-takers process verbal information in different ways and perform differently on tests based on different modes of input.

Changing the way that choices are presented can make the video-mediated test more difficult than the audio-only, but at the same time, it may be more likely to draw test-takers' attention to the aural-visual test

materials instead of the hints buried in the four written choices on the test paper. Consequently, video-mediated test-takers are forced to use their integrated listening skills, picking up the key aural- and visual-information, memorizing (or taking notes), piecing together scraps of information they have obtained, and making judgments. These skills are what they need in real-life communicative situations and are what the FATMD intends to measure (National Medical Examination Center, 2009). Integrating these communicative elements into a listening test can have positive backwash effects. If students are encouraged to study for more communicative tasks and skills in the classroom, such positive backwash effects can benefit their language learning. In turn, communication-oriented language learning may help improve their performance on the FATMD. The issue of familiarity with the video-mediated listening test will likely be solved along with an increase in the use of video texts for different purposes.

When the FATMD is made audio-only, the whole listening comprehension test lacks authenticity by nature because it is made up of readings recorded in a studio. If the FATMD is made video-mediated, it could have its authenticity improved because the video-mediated text is well recognized to reflect the way sighted people listen in the real world. The use of visual cues in listening comprehension is inevitable, as indicated by so many researchers who have sought to transform the audio-only listening comprehension test into a video-mediated version in the belief that listening comprehension is, in general, benefitted by seeing the speaker (MacWilliam, 1986).

Reforming such a high-stakes and large-scale test is a daunting challenge. Weir (1990) warned that it is difficult and time-consuming to construct video-mediated communicative tests and it requires increased technical, personnel, and financial resources to administer and demands careful training of examiners for standardization. However, Gruba (1994) reported that the training of university administrators, instructors, and students encountered no significant problems in the development of a 90-min video-mediated test in Japan. Technically, it will not be a big problem for test-developers to adjust themselves to a video-mediated listening test, since it presents no more difficulties than the audio-only in terms of preparation; this also holds true for the 30-min listening section of the FATMD in China.

An increase in computer-assisted English proficiency testing is inevitable in China, so is the growing application of video-mediated testing, especially for computer-assisted examinations. However, much concern remains about the potential of video texts for distraction, even though they have the potential for enhanced face validity and authenticity (Bejar et al., 2000). Since the published literature has turned out contradictory or inconclusive results on the advantages of video-mediated listening texts, further investigations are merited in the following three areas: the genres of video-mediated texts to be used for listening comprehension tests; their different effects on the performance of test-takers who study English as a foreign language in China; and the relation between English proficiency and the formats used to test listening.

References

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford: Oxford University Press.

Baltova, I. (1994). The impact of video on the comprehension skills of core French students. *Canadian Modern Language Review, 50*(3), 507-31.

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper* (TOEFL Monograph Series Report No. 19). Princeton, NJ: Educational Testing Service.

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

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

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

Dunkel, P. (1991). Computerized testing of nonparticipatory L2 listening comprehension proficiency: An ESL prototype development effort. *The Modern Language Journal, 75*(1), 64-73.

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

Gruba, P. (1993). A comparison study of audio and video in language testing. *JALT journal, 15*(1), 85-88.

Gruba, P. (1994). Design and development of a video-mediated test of communicative proficiency. *JALT Journal, 16*(1), 25-40.

Gruba, P. (1997). The role of video media in listening assessment. *System, 25*(3), 335-345.

Gruba, P. (1999). The role of digital video media in second language listening comprehension. Unpublished PhD thesis, Department of Linguistics and Applied Linguistics, University of Melbourne. Retrieved from http://eprints.unimelb.edu.au/archive/00000244/

Hernandez, S. S. (2004). The effects of video and captioned text and the influence of verbal and spatial abilities on second language listening comprehension in a multimedia learning environment. Unpublished doctoral dissertation, New York University, New York. Retrieved from http://search.proquest.com/pqdt/docview/305166044/abstract/13FE4D5FFBD2C1FDDC/

Kellerman, S. (1990). Lip service: The contribution of the visual modality to speech perception and its relevance to the teaching and testing of foreign language listening comprehension. *Applied Linguistics, 11*(3), 272–280.

Kellerman, S. (1992). 'I see what you mean': The role of kinesic behaviour in listening and implications for foreign and second language learning. *Applied Linguistics, 13*(3), 239–258.

MacWilliam, I. (1986). Video and language comprehension. *ELT Journal, 40*(2), 131-135.

Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions? *Educational Psychologist, 32*(1), 1-19.

National Medical Examination Center. (2009). *The guidelines to national foreign language admission test for medical doctoral students*. Beijing: People's Medical Publishing House.

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

Parry, T. S., & Meredith, R. A. (1984). Videotape vs. audiotape for listening comprehension tests: An experiment. *OMLTA journal, 8*, 47-53.

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

Rost, M. (2002). *Teaching and researching: Listening*. London: Pearson Education Limited.

Rubin, J. (1990). Improving foreign language listening comprehension. In J. E. Alatis (Ed.), *Linguistics, language teaching, and language acquisition: the interdependence of theory, practice, and research* (pp.309-316). Washington, DC: Georgetown University Press.

Shin, D. (1998). Using video-taped lectures for testing academic language. *International Journal of Listening, 12*, 56-79.

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

Sun, Q. X., Xiao, Y., Liang, Z. L., Li, Q., & Jia, R. J. (2015, Spring). Perception and acceptance of videos in the standard test of listening comprehension: An empirical study. *Fudan Forum on Foreign Languages and Literature,* 75-81.

Suvorov, R. (2013). *Interacting with visuals in L2 listening tests: An eye-tracking study*. Doctoral thesis. Iowa State University, Ames, IA.

Wagner, E. (2002). Video listening tests: A pilot study. *Working Papers in TESOL & Applied Linguistics, Teachers College, Columbia University*, *2*(1), Retrieved from http://www.tc.edu/tesolalwebjournal.

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

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.

Weir, C. (1990). *Communicative language testing*. London: Prentice Hall International.

Yu, J. T. (2008). Setting cut-off standard of national foreign language proficiency test for medical doctor with Angoff's method and evaluating the validity. *Chinese Journal of Medical Education, 28*(2), 122-125.

Zhou, G. L., & Yang, S.D. (2004). The effects of visual aid on EFL listening comprehension. *Journal of PLA University of Foreign Language, 27*(3), 58-62.

About the Authors

Ying Xiao, MA, senior lecturer at College of Foreign Languages and Literatures, Fudan University, is interested in applied linguistics, language testing and assessment, English for academic purposes while teaching English as a foreign language, with a special emphasis on its listening strategies

Zhengliu Liang, professor of English at College of Foreign Languages and Literatures, Fudan University, is interested in psycholinguistics and pedagogy

Qiang Li, testing researcher at the National Medical Examination Center under the auspice of the National Health and Family Planning Commission, People's Republic of China, is mainly engaged in Foreign Language Admission Test for Medical Doctoral Students and Chinese Medicine Practitioners Licensing Examination

Ruojun Jia, testing researcher at the National Medical Examination Center under the auspice of the National Health and Family Planning Commission, People's Republic of China, is mainly engaged in Foreign Language Admission Test for Medical Doctoral Students and Chinese Medicine Practitioners Licensing Examination