Korean Journal of English Language and Linguistics
Vol. 19, No. 4, Winter 2019, 688–712
10.15738/kjell.19.4.201912.688

Using Virtual Reality to Test Academic Listening Proficiency

Areum Lee
(Ewha Womans University)

Lee, Areum. 2019. Using virtual reality to test academic listening proficiency. *Korean Journal of English Language and Linguistics* 19–4, 688–712. This study examines the validity of a virtual reality (VR) listening test in an English as a foreign language (EFL) context. While a VR listening test has the possibility of being a more valid listening test than other listening test formats, its effectiveness has not yet been tested. This study compared test–takers' performance from the VR test against that of an audio test and a video test. A total of 54 Korean EFL university students participated. Due to the small number of participants, the test scores of the three different test modes were analyzed using a nonparametric Friedman test and a Wilcoxon signed–rank test. Differences in test–takers' performance were not statistically significant between the VR and the video groups, but were statistically significant between these two groups and the audio group. Specifically, the VR group demonstrated a better ability to search for detailed information than the audio group. In general, participants in the VR listening test responded positively to the test, situating VR as useful for listening assessments.

Keywords: English listening test, head mounted display (HMD), L2 listening assessment, semi–immersive, virtual reality listening test

1. Introduction

Listening is a complex and dynamic process in a listener's mind (Brindley 1998, Lynch 1998). Although numerous researchers have attempted to identify the components of listening (Anderson and Lynch 1998, Morley 1991, Peterson 1991, Underwood 1997), the construct of listening has not yet been defined or universally accepted (Janusik 2004). However, many researchers agree that listening ability involves not only verbal components but also non–verbal components (Baltova 1994, Gruba 1997, Harris 2008, Progosh 1996, Wagner 2007). Emphasis on non–verbal components situates the proper handling of visual information as one component of listening ability (Baltova 1994, Burgoon 1994, Ockey 2007, Richards 1983, Rost 2011,

Wagner 2008, 2010, Weir 2005). This recognition of the importance of visual information has led to the argument that a listening test should present visual inputs.

Until now, listening tests primarily used only audio inputs. However, audio-only listening tests are often criticized (Brindley 1998, Buck 2001, Lynch 1998, White 1998) because a valid listening test must present test-takers with a task that reflects the natural settings of real-life. Listening researchers claim that audio-only listening tests do not provide authentic and communicative target language-use situations (Brindley 1998, Buck 2001, Lynch 1998, Wagner 2010, White 1998), and advise that visual inputs be included in listening assessments to increase test validity (Baltova 1994, Batty 2015, Chung 1994, Coniam 2001, Gruba 1993, Kim 2004, Shin 1998, Suvorov 2009, Wagner 2007, 2008).

Accordingly, efforts to create more valid listening tests were often made to employ still images (Chung 1994, Ginther 2002, Ockey 2007). However, researchers have long argued that the still images used in listening assessments hardly represent authentic language-use situations. The speakers in the static images do not show any social interactions, and test-takers have the lack of the variety of clues that normally help listeners to infer meanings, utilizing visual clues.

Video was then used in listening assessments to offer test-takers with more authentic visual inputs than a static image alone. Video listening tests are often assumed to increase the authenticity (Baltova 1994, Brett 1997, Coniam 2001, Gruba 1993, Shin 1998, Sueyoshi and Hardison 2005) since video enables test-takers to observe the speakers. In that way, video listening tests may allow for more valid inferences to be made from the results of those assessments (Wagner 2010).

While videos can show the dynamics of a particular interaction, video listening tests can still make listeners feel like passive observers-a positionality at odds with that the listener occupies in real-life settings (Brett 1997, Gruba, 1994). The use of a monitor-a testing channel that is not immersive-leaves test-takers with a sense of remoteness, and this type of listening differs from most real-life listening situations. Sometimes listeners do not watch videos even if videos are given because test-takers are too busy focusing on written test items (Alderson et al. 1995, Brett 1997, Gruba 1994, Wagner 2007). Thus, language-testing developers must rethink how listening performance might best be assessed (Field 2004, 2008, Hughes 2003, Ockey 2007).

With the advancement of technology, virtual reality (VR) may facilitate better listening assessments by intensifying a test-taker's sense of presence, immersion, and realism (Chung 2016, Dalgarno and Lee 2010, Dolgunsöz et al. 2018, Witmer and

Singer 1998). As VR allows test-takers to immerse themselves into realistic listening contexts, test-takers can experience a VR listening test as if they are taking part in actual communication. Accordingly, participants in VR listening tests are expected to be more cognitively engaged in listening and better identify the situations their tests present. Moreover, VR can provide test-takers a 360° view in any direction, and test-takers can decide where to devote their attentional resources. By their own viewing choice, visual information can be efficiently used. Taking this potential to task, this study tests the validity of a VR listening test over and against the validities of audio and video listening tests. In addition, it explores VR listening test-takers' perceptions of the VR listening test with a questionnaire.

2. Theoretical Background

2.1 The Importance of Visual Inputs in Listening Assessment

Many studies (Hadar 1989, McGurk and MacDonald 1976, McNeil 1985, Richards 1983, Riseborough 1981, Scheflen 1964) have pointed out the importance of extralinguistic information (i.e., visual context, gestures, and facial expressions) in listening. A well-known study by McGurk and MacDonald (1976) reports that speaker lip movements provide a listener with information that helps them better understand what is being said. Along these same lines, hand gestures (McNeil 1985, Riseborough 1981), body posture (Scheflen 1964), and head movements (Hadar 1989) also seem to affect listeners' comprehension. Previous studies suggest that such non-verbal information can be valuable in helping listeners decode meaning in conversations (Kintsch 1998, von Raffler-Engel 1980). Therefore, researchers assert that the ability to process visual inputs needs to be recognized as a part of listening ability (Baltova 1994, Ockey 2007, Rost 2011, Wagner 2008, Weir 2005), and have thus been paying more attention to the role of visual inputs in the construct of listening assessment (Ginther 2002).

To create a more valid listening test, researchers combined still images with audio inputs in listening assessments and examined test-takers' performance (Chung 1994, Ginther 2002, Ockey 2007). Some studies notably look at how the number of still images used in a test affects listening performance. For example, Chung (1994) examines test-takers' listening comprehension across four modes of listening testing:

audio, audio with a single image, audio with several images, and analogue video. He reports that while listening comprehension scores improved significantly with the inclusion of images, displays of multiple still images distracted participants. Although some studies (Chung 1994, Ginther 2002) show that including still images in previously audio-only listening tests improve test-takers' performance, this listening testing format, as noted above, has been criticized because simply including a still image in a listening test fails to reflect the communicative interactions so central to comprehension for test-takers in real-life scenarios.

In order to provide more authentic language-use situations, a growing number of studies then employed videos in listening tests (Coniam 2001, Gruba 1993, Ockey 2007, Shin 1998, Wagner 2007). For example, Ockey (2007) compares the use of a series of still images in a listening test with the use of video in a listening test to observe how tertiary-level ESL test-takers differently engaged the test formats. Ockey (2007) further found that participants in the test with still images demonstrated minimum engagement while participants in the test with video demonstrated more diverse engagement.

As videos closely simulate authentic interactions by portraying a speaker's body language, facial expressions, and gestures (Baltova 1994, Buck 2001, Gruba 1997, Wagner 2007), being able to encounter a speaker in this way enables the listener to make more accurate initial hypotheses about the speakers' roles and the situation (Shin 1998, Wagner 2010). Since situational authenticity is more present in video-based listening tests than in audio-only and still image listening tests, videos may enable listeners to easily identify participants and situations.

The results of the previous studies on video format listening tests, however, revealed inconsistent results. For example, Wagner's (2010) comparison of the success of participants in a video listening test and an audio listening test showed that listeners in the video group scored 6.5% higher than those in the audio group-the difference was statistically significant. Meanwhile, Coniam's (2001) study found no significant performance difference between a video and an audio listening test, and 82% of participants reported that video was not helpful.

A number of studies (Baltova 1994, Shin 1998, Sueyoshi and Hardison 2005) report that test-takers performed better on video listening tests than audio-only listening tests while other studies (Brett 1997, Coniam 2001, Gruba 1994) indicate that the inclusion of videos in formerly audio-only listening tests did not significantly help test-takers' performance. Meanwhile, some researchers (Alderson et al. 1995, Brett

1997, Gruba 1994) argue that the use of video may not be helpful for test-takers' comprehension because listening test-takers are often too busy focusing on their written test items to use the non-verbal information conveyed by the video. And even if the video was helpful, video listening tests cannot fully reflect real-life language-use situations. This aspect prompts the need for a new mode of listening testing to better evaluate test-takers' listening ability. Weaving video and audio in a way that simulates real-life, VR may serve as a more valid mode for listening testing.

2.2 The Potential Advantages of VR in Listening Assessments

Virtual reality is not a new technology. After first emerging in the 1950s, VR became popular with the radical development of technologies running up to and after the new millennium. There are various kinds of VR (Chen and Teh, 2013, Shih and Yang 2008), and one way to categorize VR is through level of immersion (Chen and Teh 2013): non-immersive, semi-immersive, and total-immersive.

Non-immersive VR provides users with a computer-generated environment without a feeling of being immersed in the VR. Non-immersive VR relies on a computer or video game console, display, and input devices like a keyboard and a mouse. Next, semi-immersive VR allows users to experience virtual three-dimensional environments. With semi-immersive VR, users can see what's going on around them, and is generally enjoyed using a head-mounted display (HMD) with goggles. The goggles can provide a 120° field of vision and help to present artificial worlds that simulate the real worlds-far more so than when encountering a simulation on a faraway video screen. Moreover, the head- and eye-tracking capacities of semi-immersive VR make possible natural eye movements. What is crucial to note here is that semi-immersive VR enables users to choose their line of sight, unlike non-immersive VR. Last, in total-immersive VR, users use interactive tools such as motion controllers, allowing for interaction with VR contents. With motion controllers, users can transport their movements and gestures right into the VR. It is possible for users to throw, grab, move objects in VR with intuitive and realistic precision.

Using VR in listening assessments may improve assessments of listening comprehension. VR listening tests may be more valid than video-based listening tests with several reasons. HMD goggles can provide test-takers with the highest quality resolution and the widest field of vision (Jennett et al. 2008). High-resolution graphics intensify immersion, leading to better test-takers' presence, and listeners

may better identify situations and atmospheres (Grabowski and Jankowski 2015, Zhang et al. 2016) than video listening tests. In addition, VR listening test can allow a listener look 360° around themselves and thus to choose visual resources depending on their needs. VR may more powerfully amplify a test-taker's ability to make sense of visual inputs than a video. Against the more detached feeling of video tests, the visual realism of VR may compel test-takers to be more cognitively engaged in their test.

To date, VR has been used and studied in many fields such as medicine (Riva 2003), culture acquisition (Shin 2015), and tourism (Wagler and Hanus 2018). For example, Shin (2015) studied whether students could improve their culture acquisition ability by using VR technology. Learners walked in the streets of London in VR with an English guide explaining history and architecture, and the participants talked to the guide through voice chat. The results clearly showed that participants benefitted from their cultural immersion in the VR. Another recent study by Wagler and Hanus (2018) compares a video tour, a VR tour, and a real tour. After each tour, the participants' levels of emotional engagement and presence were measured. Levels of emotional engagement and presence were the lowest in the video tour, while similar levels of emotional engagement and presence were found in both the VR tour and the real tour. These findings suggest that a high level of immersion and a strong sense of presence can be beneficial to experience the most realistic environment.

The sense of presence, immersion, and realism (Chung 2016, Dalgarno and Lee 2010, Dolgunsöz et al. 2018, Witmer and Singer 1998) VR provides may give listeners a stronger, more realistic feeling of their spaces and access to visual information than a video. VR, which, as noted above, has not yet been used and tried as a platform for listening assessments, and most of the VR related studies have been primarily conducted at the non-immersive level (De Lucia et al. 2009, Elder et al. 2002, Song 2014). Therefore, employing semi-immersive VR in listening assessments may provide a more valid listening test mode compared to other listening test modes. Taking this potential of VR to task, this study examines whether a VR listening test may be a more valid test format than audio and video listening test modes. More specifically, it evaluates the efficacy of VR listening tests over and against audio and video listening tests for Korean EFL university students. In addition, this study investigates the VR listening test-takers' perceptions of the usefulness of VR in listening assessment through a post-questionnaire. As little research was conducted on test-takers' perceptions of using VR in listening assessments, assessing test-takers' perceptions can inform listening test developers of what additional changes can be made in order

to make the test more acceptable to the test-takers (Song 2014). Therefore, the following research questions were addressed for this study:

1) How does test taker listening performance differ across audio, video, and virtual reality listening tests?
2) How does test taker listening performance differ in terms of test taker proficiency levels?
3) Do the effects of different test modes differ according to different types of questions, namely a main idea question, a detail idea question, and an inference question?
4) What are VR listening test taker perceptions of the usefulness of virtual reality in the listening assessment?

3. Methods

3.1 Participants

The 54 participants were recruited through an online bulletin board at three universities—49 female and five male students of mixed majors between the ages of 20 and 34. Forty nine people are in their 20s, and five people are in their 30s. They had submitted their TOEIC test scores prior to the study. To examine whether different proficiency levels can affect test-takers' performance, they were divided into two groups: high and low. Test-takers in the high proficiency level group had TOEIC scores in the range of 900-990, while those in the low proficiency level group had in the range of 600-700. Three people who had TOEIC scores between 700-900 were excluded from the experiment. According to these criteria, both groups had 27 participants each.

To examine the test mode effects, the 27 in the high proficiency group were further divided into three test mode groups: audio, video, and VR. The same pattern was followed for the 27 in the low proficiency group. Thus, nine students each from both groups were assigned to one test mode group according to their proficiency level. Table 1 shows the complete analysis of the data acquired from the 54 participants.

Table 1. Participants in Different Test Modes

Test Modes	Number of Test-takers in High Proficiency Group	Number of Test-takers in Low Proficiency Group
Audio-only listening test	9	9
Video listening test	9	9
VR listening test	9	9

The 18 students in the VR group were then asked whether they had previous VR experience—nine students did not, while the other nine did, as shown in Table 2. Since watching 360-degree videos is not a must, the nine with no VR experience had no difficulty in using HMD VR goggles.

Table 2. Previous Experience with Different Types of VR

Types of VR	*N*
Total immersive VR	1
Semi-immersive VR	8

Those with VR experience demonstrated the different VR activities they had been exposed to — only one had experienced VR through full immersion at an arcade; the other eight had watched 360-degree videos wearing VR goggles.

3.2 Materials

3.2.1 Listening passages

The listening test comprised three monologs and three dialogs. Five of those — two monologs and three dialogs — were taken from commercially published conversation books and textbooks (e.g. *Person to Person* and *TOEFL iBT*), and one monolog was taken from a *Ted Talk* presentation. Barring the *Ted Talk* presentation, the other five had been recorded by seven recruited actors, who were required to memorize the scripts prior to the study. The range of the listening passages for this study was between 90 seconds and two minutes.

3.2.2 Listening question items

From the passages, 32 listening question items were constructed. These included multiple-choice questions with four options and short-answer items that required the

test-takers to write answers of 10 words or less. There were five to six question items after each passage. With respect to the types of listening questions, the following three were used: a main idea question, a detail idea question, and an inference question, comprising 17, six, and nine questions, respectively. For comparability, all test questions on different test modes were identical, and all test-takers had to write the answers on paper.

3.2.3 VR recordings

With respect to VR recordings, five passages were recorded by the researcher using a 360-degree camera at a gym, lecture room, clothing shop, supermarket, and clinic. For monologs, speakers were filmed at the gym and lecture room, as shown in Figure 1.

Figure 1. Virtual Reality Places for Monologs

Further, the last monolog was extracted from a *Ted Talk*, available on *YouTube* and in VR format, as shown in Figure 2.

Figure 2. *Ted Talk* for Monologue

For dialogs, speakers were filmed at the clothing shop, supermarket, and clinic. There were more than two speakers in each dialog, as shown in Figure 3.

Figure 3. Virtual Reality Places for Dialogs

All the recordings allowed actors' hesitations, false starts, and pauses in a natural way to increase authenticity. Furthermore, for a high level of situational authenticity, background noise was not excluded or controlled.

3.2.4 Post-questionnaire

A post-questionnaire was then given to analyze VR listening test-takers' perceptions of the usefulness of the VR listening test. Questionnaire items comprised three open-ended questions. First, background questions, such as major, name, age, and previous experience with VR, were asked, and then, the three open-ended ones — the first asked about the advantages of the VR listening test, and the second asked about its disadvantages; the last question was whether they would be willing to take the VR listening test if provided.

3.3 Data Collection Procedure

The 54 participants were given the test instructions individually when they entered the classroom; after that, the listening test was provided. Test-takers were allowed to listen or watch only once with 40 minutes to complete the test. They were given a minute to go through the set of comprehension questions before the recording was played. After the recording finished, test-takers answered the questions. No

note-taking was allowed. Answering the 32-item test took approximately 40 minutes.

To elaborate on this study's three different test modes (audio, video, and VR listening), the audio test was provided to 18 test-takers without visual input, while for the video test, recorded 2D videos were played on a computer monitor, and test-takers were asked to answer the questions after each video; and for the VR listening test, *Oculus Rift* head-mounted displays (HMD) were used to provide a more immersive experience, and virtual reality platform *Gizmo* was used to play the VR recordings. Test-takers were asked to wear the goggles and watch each 3D recording, and then remove the goggles and answer the questions. VR listening test-takers had 50 minutes to complete the test, with time to set up the gear taken into consideration. They were seated during the viewing and could turn their heads to look in any direction.

A post-questionnaire was distributed right after the listening test to examine the VR group's perceptions of their VR listening assessment. Including background questions, they were given 15 minutes to answer the three open-ended questions. The data were then recorded in a computer for further analysis.

3.4 Data Analyses

After completion, scoring for the listening test was done dichotomously — right or wrong. All 32 items were scored either 1 or 0, with 32 being the maximum score given. Based on the scores obtained from the different test mode groups, the Friedman test, SPSS version 21, was used as a non-parametrical test to examine whether the test modes and proficiency levels had any effect on the test-takers' listening performance. Then, the Wilcoxon signed-rank test was conducted to examine the differences between the groups.

Furthermore, the Friedman test was used to investigate whether the effects of the test modes differed according to the types of questions, namely understanding the main idea, understanding detailed ideas, and drawing inferences. In case of differences, the Wilcoxon signed-rank test was conducted to find out where they lay.

As for the perceived usefulness of virtual reality in listening assessment, post-questionnaire responses were categorized by themes and their frequency measured, and the advantages and disadvantages of using VR in listening assessment were explored.

4. Results

4.1 Test Mode Effects on Test-takers' Listening Test Performance

Scores from each test mode were compared to find out the effects of the test modes on the test-takers' listening test performance. Table 3 presents the descriptive statistics of listening test scores from the three test modes—high proficiency level test-takers in the audio group (n = 9) have a mean score of 21.75 out of 32, the maximum score, those in the video group have 26.50, and those in the VR group have 27; low proficiency test-takers in the audio group have a mean score of 15.75, those in the video group have 22, and those in the VR group have 23.75.

Table 3. Descriptive Statistics of Listening Test Scores from Different Test Modes

Test Mode	Proficiency Level	*M*	*SD*	*N*
Audio	High	21.75	2.217	9
	Low	15.75	2.217	9
Video	High	26.50	4.203	9
	Low	22.00	1.414	9
VR	High	27.00	.816	9
	Low	23.75	3.500	9
Total		22.79	4.482	54

To find out whether the differences between the test modes are statistically significant, a non-parametric Friedman test was conducted, rendering a Chi-square value of 11.29, which was significant (p = .004) at the significance level of .05, as shown in Table 4.

Table 4. Test Mode Effects on Test-takers' Listening Test Performance

X^2	*df*	*Sig.*
11.29	2	.004

To closely investigate where those differences occur, the Wilcoxon signed-rank test was conducted on the different combinations of the three groups; scores from these combinations were compared and statistically examined, as shown in Table 5. As a result, scores between the audio and video groups (Z = −2.375, p = .018) and audio and VR groups (Z = −1.199, p = .012) were revealed to have statistically significant

differences. However, there was no significant differences between the VR and video groups (Z = −2.527, p = .230).

Table 5. The Results of Wilcoxon signed-rank Test for the Test Mode Effects

Types of Test Mode	*Z*	*Sig.*
Audio − Video	−2.375	.018
Video − VR	−2.527	.230
VR − Audio	−1.199	.012

The results show that test groups with visual input significantly outperformed the audio group, thus revealing that receiving visual input leads to a better listening performance.

4.2 Effects of Proficiency Levels on Listening Test Performance

To examine the effects of proficiency levels on test-takers' listening test performance, a non-parametric Friedman test was conducted between test-takers with low and high levels, rendering a Chi-square value of 14.55, which was significant (p = .012), as shown in Table 6.

Table 6. The Effect of Proficiency Levels on Test-takers' Listening Test Performance

X^2	*df*	*Sig.*
14.55	1	.012

The results indicate that test-takers' listening performance varies according to their proficiency level. In other words, test-takers with a high proficiency level outperformed those with a low level.

4.3 Test Mode Effects on Listening Test Performance by Question Types

Next, effects of test modes on listening test performance by question types were examined. Table 7 shows the comparison of mean scores of different question types — main idea, detail idea, and inference — across the three test modes.

Table 7. Descriptive Statistics of Listening Test Performance by Different Question Types

Question Types	Test Modes	*M*	*SD*	*N*
Detail idea	Audio	4.00	1.069	6
	Video	4.88	1.356	6
	VR	5.38	.744	6
Inference	Audio	7.00	1.927	9
	Video	8.13	.991	9
	VR	8.50	.756	9
Main idea	Audio	9.63	3.021	17
	Video	11.38	2.825	17
	VR	12.50	3.295	17

For detail-type questions, the audio group had a mean score of 4, video group of 4.88, and VR group of 5.38; for inferences, the audio group had a mean score of 7, video group of 8.13, and VR group of 8.5; and for the main idea-question type, a similar pattern — of the audio group having the lowest score — was observed: 9.63 for the audio group, 11.38 for the video group, and 12.50 for the VR group. The overall pattern observed was the VR group outperforming the other two and the audio group scoring the lowest regardless of the question types, as shown in Figure 4.

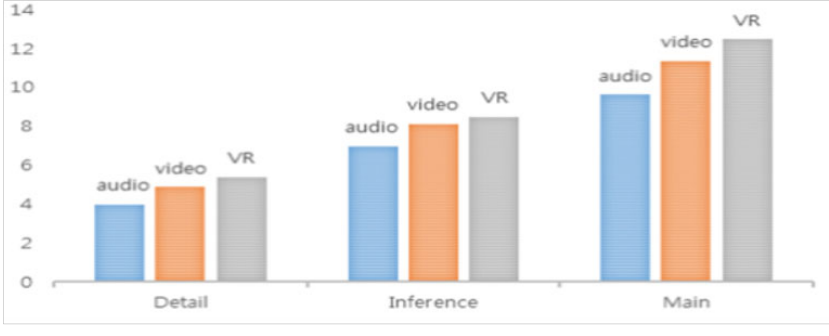

Figure 4. Group Differences of Listening Test Performance by Question Types

To find out whether the differences depended on the test modes, further analysis was conducted. The Friedman test was carried out to examine whether the score differences depended on the different question types.

4.3.1 Test mode effects of test-takers' performance on detail-type questions

The result is a Chi-square value of 4.923, which is significant (p = .045), as shown in Table 8. There were significant group differences in listening test performance with respect to detail-type questions. The Wilcoxon signed-rank test was performed to determine which groups showed differences.

Table 8. Test Mode Effects on Detail-Type Questions

X^2	*df*	*Sig.*
4.923	2	.045

Table 9. The Post-hoc Test for the Test Mode Effects on Detail-Type Questions

Types of test mode	Z	*Sig.*
Audio - Video	-1.294	.196
Video - VR	-.850	.395
VR - Audio	-2.232	.026

As shown in Table 9, a statistically significant difference was found between the VR and audio groups (Z = -2.232, p = .026) but not between the audio and video groups (Z = -1.294, p = .196) and video and VR groups (Z = -.850, p = .395). Although no difference is found between the video and VR groups, it is assumed that the VR group test-takers were able to better search for detailed information.

4.3.2 Test mode effects of test-takers' performance on inference-type questions

To understand whether there were any differences in performance with respect to inference-type questions, the Friedman test was conducted.

Table 10. Test Mode Effects on Inference-Type Questions

X^2	*df*	*Sig.*
3.308	2	.191

As shown in Table 10, no statistically significant differences in listening test performance across the different test modes were found—a Chi-square value of 3.308 was found, which was not significant (p = .191). This result shows that the different test modes did not affect the listening performance in inference-type questions.

4.3.3 Test mode effects of test-takers' performance on main idea-type questions

The Friedman test was performed to investigate any group differences resulting from different test modes on main idea-type questions. According to Table 11, there were no significant mean differences. A Chi-square value of 5.586 was found, which was not significant (p = .061).

Table 11. Test Mode Effects on Main Idea-Type Questions

X^2	*df*	*Sig.*
5.586	2	.061

In conclusion, the group differences between different test modes were observed by question types, and the VR and audio groups showed significant differences when it came to searching for details. Furthermore, statistically significant differences in the listening test performance in inference- and main idea-type questions were not found. Through visual inputs in VR, test-takers seemed to better search for detailed information.

4.4 Learner Perceptions on Virtual Reality Listening Test

Post-questionnaire questions asked to the VR listening test group were analyzed to explore test-takers' perceived usefulness of the VR listening test. With respect to the perceived benefits of using VR in the listening assessment, most participants gave positive responses, as shown in Table 12.

Table 12. Comments on Perceived Advantages and Disadvantages of Using Virtual Reality in Listening Assessment

	Comments	Frequency
Advantages	VR makes me highly immersive in the artificial worlds, helping me easily understand the situations and atmosphere.	14
	VR listening test is interesting and fun.	12
	VR makes it easier to identify speakers' attitudes, facial expressions, and gestures.	8
Disadvantages	VR does not allow note-taking.	6
	I experience virtual reality sickness.	4
	I can hear background noise.	4

VR test-takers' comments on the advantages of VR listening test were categorized into three opinions. A total of 14 test-takers commented that VR made them highly immersive so that they could easily understand the situations and atmosphere—being highly immersive as one of the advantages of VR is mentioned in other VR studies (Ornberg 2003, Song 2014) as well; for example, they could immediately identify the situations where the conversations had taken place and the atmosphere, such as peaceful or crowded. Furthermore, 12 test-takers thought experiencing VR in a listening test was fun and interesting. Eight test-takers indicated that VR helped them identify speakers' attitudes, facial expressions, and gestures; in fact, one test-taker said the speakers appeared to talk to her like in real life in monolog-type listening settings, enabling her to identify their facial expressions easily. Thus, VR seems to provide better detailed information about the speakers (Buck 2001, Coniam 2001, Ockey 2007). Similar comments have been observed in previous VR studies (Ornberg 2003, Song 2014).

However, there were negative remarks as well, which too were categorized into three opinions. Six test-takers were dissatisfied with the VR listening test owing to the fact that note-taking is not possible; as academic listening tests normally allow test-takers to take notes, lack of it can be a serious issue (Dunkel and Davy 1989, Hale and Courtney 1994). However, this study did not allow that, and note-taking did not seem to affect the study results. As enabling note-taking in VR listening test does not seem easy to achieve technically, other alternatives can be suggested—first, test-takers might be asked to listen to passages twice to reduce the cognitive burden, and second, questions requiring too specific information may be excluded from the VR listening test, thus calling for a restructuring of the listening test items.

Furthermore, four test-takers expressed that they felt dizzy during the VR test. Dizziness during VR, also called as virtual sickness or motion sickness, has been mentioned as one of the disadvantages in existing literature as well (Dolgunsöz et al. 2018, Lawrence and Gary 1992, Magaki 2017, Song 2014). Virtual sickness can occur when there is a gap between the intended senses in VR and test-takers' actual senses. It cannot be solved perfectly. However, to minimize the gap, if the hardware is improved, for example, increase in resolution, better refresh rates, and better lenses, this sickness can be reduced. VR contents should be designed with caution by considering movement of test-takers' views and the change of locations. As the level of presence and immersion depends heavily on the type of device and feature, more research seems necessary for application of this emerging technology in assessment.

Four other test-takers stated that background noise, due to the recordings being taped in everyday outdoor locations to increase authenticity, made it difficult for them to focus on conversations. When it comes to authenticity, tasks should reflect real-life situations. These negative comments reflect the fact that EFL Korean test-takers seem too familiar with controlled listening recordings.

When test-takers were asked whether they would be willing to take a VR listening test over other listening test modes, all answered in the affirmative. Considering the larger number of test-takers responding positively, VR listening test seems to provide test-takers with richer visual cues to comprehend listening. However, this does not necessarily mean it is a more valid test, because the positive comments, i.e. advantages, do not seem to directly show a better performance, based on the non-significant performance differences between the VR and video test modes.

5. Conclusion and Discussion

This study examined whether a VR listening test is a more valid test mode than audio and video listening test modes. In order to examine the efficacy of VR listening tests over audio and video listening tests for Korean EFL learners, the results of test scores from three different test modes were compared. As VR can provide a strong sense of presence and highly immersive feelings in the test, test-takers were expected to make the most of visual inputs by their own choice. VR was further expected to increase the authenticity of tests and be more valid in listening assessments.

The results revealed that differences in test-takers' performance for the VR and the video groups were not statistically significant; however, they were statistically significant between these two groups and the audio group. The inclusion of visual information in listening assessments helps Korean test-takers' listening comprehension; this is consistent with the results of previous studies (Baltova 1994, Shin 1998, Wagner 2010). However, the performance differences between VR and video groups were not statistically significant. It is thus assumed that the possible advantages of using VR in listening assessments were not adequately played out at the semi-immersive level.

The mean scores of different question types across the three test modes were then compared in order to asses if test-takers' performance differs according to question

types, namely, main idea-type, detail idea-type, and inference-type. The results showed that differences in test-takers' performance on a detail idea question were not statistically significant between audio and video groups and video and VR groups, respectively; however, the differences in test-takers' performance for the audio and VR groups were statistically significant. It was found that the VR group seemed to better search for detailed information than the audio group. However, when it came to the performance on main idea- and inference-type questions, the differences in scores from all three groups were not statistically significant.

The test-takers' perceptions of the usefulness of VR in listening assessments were revealed through a post-test questionnaire; they generally expressed positive opinions. Test-takers commented that they could get quickly immersed in VR, and could closely see the speakers' facial expressions and gestures. Moreover, all of the test-takers thought the VR listening test was fun and interesting, and they are willing to choose the VR listening test over other listening test modes. Although the majority of participants expressed the benefits of the VR listening test, there were some negative remarks. Some test-takers commented that they experienced virtual sickness, and others were dissatisfied with the VR listening test since note-taking is not possible. However, those challenges seemed closely linked to technical issues. If VR is more technologically advanced, then there is a possibility that these challenges will be solved.

Contrary to the potential effect of VR on task performance, the VR listening test did not show the expected results. However, as test-takers gave positive responses and they seemed to get help from VR, for example, understanding detailed information, VR seemed to clearly provide more dynamic conversations and immersive experience than the video listening test. And if more diverse listening question items, other than multiple-choice comprehension items, are developed that utilize the advantages of semi-immersive VR as much as possible, the VR listening test can be a more valid evaluation than other listening test formats.

In order to create a more valid listening test, total-immersive VR can also be implemented. Total-immersive VR goes beyond semi-immersive VR; test-takers can be participants of artificial worlds. With test-takers as participants, a variety of test questions and assessment methods can be developed. For example, test-takers can be expected to act real in VR, such as listening to specific directions and act it out. Test-tasks can meet target language-use tasks more closely than any test, and thus potentially increase test usefulness.

Although VR listening tests should not be considered as the perfect listening test, they do have the potential of being a more valid listening assessment. VR has been studied and utilized in several fields and has shown positive results. Thus, it is necessary for test developers to study these possibilities. Continuous efforts are required for better and more valid listening assessments.

There are several limitations to this study. The sample size (n = 54) was small, and the interaction effect between the test mode and proficiency level was not explored. More participants need to be recruited for further studies. Next, as most of the previous studies (Baltova 1994, Gruba 1993, Shin 1998, Sueyoshin and Hardison 2005, Suvorov, 2013) utilized multiple-choice comprehension items, this study also primarily included multiple-choice comprehension items. Other types of question items need to be developed to assess test-takers' performance more effectively. In addition, in-depth interviews inquiring into the VR test-takers' level of immersion can be conducted for better results. Given the significance of this topic for the future of listening comprehension testing, additional research is needed using different sampling and test materials.

References

Alderson, J., C. Clapham. and D. Wall. 1995. *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Anderson, A. and T. Lynch. 1998. *Listening.* Oxford: Oxford University Press.

Baltova, I. 1994. The impact of video on the comprehension skills of core French students. *Canadian Modern Language Review* 50, 507-531.

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

Brett, P. 1997. A comparative study of the effects of the use of multimedia on listening comprehension. *The system* 25, 39-53.

Brindley, G. 1998. Assessing listening abilities. *Annual Review of Applied Linguistics* 18, 171-191.

Buck, G. 2001. *Assessing Listening.* Cambridge: Cambridge University Press.

Burgoon, J. K. 1994. Nonverbal signals. In M. L. Knapp and G. R. Miller, eds., *Handbook of Interpersonal Communication,* 229-285. CA: SAGE Publications.

Chen, C. J. and C. S. Teh. 2013. Enhancing an instructional design model for virtual reality-based learning. *Australian Journal of Educational Technology* 29, 699-716.

Chung, B. 2016. VR market trends and prospect, *Dong Hyang* 28, 7-13.

Chung, U. K. 1994. *The Effect of Audio, a Single Picture, Multiple Pictures, or Video on Second Language Listening Comprehension*. Doctoral dissertation, University of Illinois, Urbana-Champaign, IL, USA.

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, 1-14.

Dalgarno, B. and M. Lee. 2010. What are the learning affordances of 3D virtual environments? *British Journal of Educational Technology* 41(1), 10-32.

De Lucia, A., R. Francese., I. Passero and G. Tortora. 2009. Development and evaluation of a virtual campus on Second Life: The case of Second DMI. *Computers & Education* 52, 220-233.

Dolgunsöz, E., G. Yıldırım and S. Yıldırım. 2018. The effect of virtual reality on EFL writing performance. *Journal of Language and Linguistic Studies* 14, 278-292.

Dunkel, P. and S. Davy. 1989. The heuristics of lecture notetaking: Perceptions of American & international students regarding the value & practice of notetaking. *English for Specific Purposes* 8, 33-50.

Elbder, C., N. Iwashita and T. McNamara. 2002. Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing* 19, 347-368.

Field, J. 2004. An insight into listeners' problems: Too much bottom-up or too much top-down. *System* 32, 363-377.

Field, J. 2008. *Listening in the Language Classroom*. Cambridge: Cambridge University Press.

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

Grabowski, A. and J. Jankowski. 2015. Virtual reality based pilot training for underground coal miners, *Safety Science* 72, 310-314.

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

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

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

Hadar, U. 1989. Speech production: Role of head movements. *Language and Communication* 9, 245-257.

Hale, G. and R. Courtney. 1994. The effect of note taking on listening comprehension in the TOEFL. *Language Testing* 11, 29-47.

Harris, T. 2008. Listening with your eyes: The importance of speech-related gestures in the language. *Foreign Language Annals* 36, 80-187.

Hughes, A. 2003. *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Janusik, L. A. 2004. *Researching Listening from the Inside Out: The Relationship Between Conversational Listening Span and Perceived Communicative Competence.* Doctoral dissertation, University of Maryland, MD, USA.

Jennett, C. A., L. Cox., P. Cairns., S. Dhoparee., A. Epps., T. Tijs and A. Walton. 2008. Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies* 66, 641-661.

Kim, J. 2004. *Second Language English Listening Comprehension Using Different Presentations of Pictures and Video Cues*. Doctoral dissertation, University of New South Wales, Sydney, NSW, Australia.

Kintsch, W. 1998. *Comprehension: A Paradigm for Cognition*. Cambridge: Cambridge University Press.

Lawrence, J. H. and E. R. Gary. 1992. Visually induced motion sickness in virtual environment. *Presence: Virtual and Augmented Reality* 1, 306-310.

Lynch, T. 1998. Theoretical perspectives on listening. *Annual Review of Applied Linguistics* 18, 3-19.

Magaki, T. 2017. Measuring reduction methods for VR sickness in virtual environments. *International Journal of Virtual and Personal Learning Environment* 7, 27-43.

McGurk, H. and Y. Macdonald. 1976. Hearing lips and seeing voices. *Nature* 264, 746-748.

McNeill, D. 1985. So you think gestures are non-verbal? *Psychological review* 92, 350-371.

Morley, J. 1991. Aural comprehension instruction: Principles and practices. In M. Celce-Murcia, ed., *Teaching English as a second or foreign language,* 69-85. MA: New bury House.

Ockey, G. J. 2007. Construct implications of including still image or video in

computer-based listening tests. *Language Testing* 24, 517-537.

Ornberg, T. 2003. *Linguistic Presence on the Internet: Communication, World View and Presence in Online Virtual Environments*. Master's thesis, University of Umea, Sweden.

Peterson, P. W. 1991. Skills and strategies for proficient listening. In M. Celce-Murcia, ed., *Teaching English as a Second or Foreign Language,* 87-100. MA: New bury House.

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

Richards, J. C. 1983. Listening comprehension: Approach, design, and procedure. *TESOL Quarterly* 17, 219-240.

Riseborough, M. 1981. Physiographic gestures as decoding facilitators: Three experiments exploring a neglected facet of communication. *Journal of Non-verbal Behavior* 5, 172-183.

Riva, G. 2003. Applications of virtual environments in medicine. *Methods Inf Med* 42, 524-534.

Rost, M. 2011. *Teaching and Researching Listening*. UK: Pearson.

Sheflen, A. 1964. The significance of posture in communication systems. *Psychiatry* 27, 316-331.

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

Shin, Y. C. 2015. A virtual walk through London: Culture learning through a cultural immersion experience. *Computer Assisted Language Learning* 28, 407-428.

Shih, Y. C. and M. T. Yang. 2008. A Collaborative virtual environment for situated language learning using VEC3D. *Educational Technology & Society* 11, 56-68.

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

Song, J. 2014. *A Study of ESL Students' Performance and Perceptions in Face-to-face and Virtual-world Group Oral Tests.* Doctoral dissertation, The University of Texas at Austin, TX, USA.

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 and I. Katz, eds., *Developing and Evaluating* L*anguage Learning Materials,* 53-68. IA: Iowa State University.

Underwood, M. 1997. *Teaching Listening*. New York: Longman.

von Raffler-Engel, W. 1980. Kinesics and paralinguistic: A neglected factor in second-language research and teaching. *Canadian Modern Language Review* 36, 225-237.

Wagler, A. and M. D. Hanus. 2018. Comparing virtual reality tourism to real-life experience: Effects of presence and engagement on attitude and enjoyment, *Communication Research Reports* 35, 456-464.

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

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

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

Wang, C. X., H. Song., F. Xia and Q. Yan. 2009. Integrating Second Life into an EFL program in China: Research Collaboration across the continents. *TechTrends* 53, 14-19.

Weir, C. J. 2005. *Language Testing and Validation: An Evidence-based Approach*. Basingstoke, UK: Palgrave Macmillan.

White, G. 1998. *Listening*. Oxford: Oxford English.

Witmer, B. G. and M. J. Singer. 1998. Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments* 7, 225-240.

Zhang, H., X. He., B. Nie and H. S. Mitri. 2016. *A new virtual reality training system for cool underground mines*. Paper presented at the 3rd International Symposium on Mine Safety Science and Engineering. Montreal, Canada.

Examples in: English
Applicable Languages: English
Applicable Level: Tertiary

Lee, Areum (Ph.D. Candidate)
Ewha Womans University
Department of English Education
52, Ewhayeondae-gil, Seodaemun-gu
Seoul, Korea
Tel: 02-3277-2647
E-mail: areum0504@gmail.com

Received: Nov. 10, 2019
Revised: Dec. 12, 2019
Accepted: Dec. 23, 2019