Does Modality Matter? On the Modality of Visuals in Computerized Tests of Listening and Its Implications for Design Decisions

Hamed Babaie Shalmani (babaie@iaurasht.ac.ir)
Department of English Language, College of Humanities
Rasht Branch, Islamic Azad University, Rasht, Iran

Abstract

Inspired by the controversial findings of studies suggesting that the modality or redundancy effect is produced when information presented to learners is available in multiple modalities, the present study examined whether the integration of context video or static images in multimedia tests of listening would privilege the participants or adversely impact their listening comprehension in English. To this aim, three study groups were assigned to listen to five aural passages on everyday themes under one of three conditions: an audio-only condition where only auditory information was available; a video-plus-audio condition where the participants listened to the passages while watching video clips of the interlocutors and the settings; and an image-plus-audio condition where both aural information and screenshots of the video counterparts of context visuals were available. The result revealed that both types of context visuals had proved fruitful in aiding the participants' comprehension of the texts, and that while the modality effect was present, the redundancy effect was not produced. Yet, the study found a significant difference in the performance levels of the two experimental groups in favor of context video. Among the proposed explanations is the idea that higher levels of engagement, increased noticing, reduced cognitive load, and in-depth processing and decoding of aural input associated with the use of video modality may demonstrate a preference for the inclusion of video modality in computerized tests of listening.

Keywords*:* context static image, context video, listening comprehension, the modality effect, the redundancy effect

Introduction

Over the past decade, the use of visual stimuli in computerized tests of listening has been growing in popularity largely thanks to Krashen's (1987) Input Hypothesis and social interactionist approaches to language acquisition holding that adding redundancy to input would render it comprehensible to language learners, and that enhanced input is more likely to become intake. While a growing body of research (Babaie, 2008; Hashemi Shahraki & Kassaian, 2011; Kuhl, 2010; Lantolf, 2000; Long, 1996; Pinnow, 2011; Vickers, 2010) has corroborated this view, there is little evidence to show how comprehensible input affects the acquisition process. Convictions are strong that engagement with the input through interaction exerts a substantial impact on the proportion of input becoming intake, and that varying levels of engagement with the input available in different modalities seem to correlate with varying degrees of success with the acquisition process (Amadieu, Lemarié, & Tricot, 2017; Anastopoulou, Sharples, & Baber, 2011; Babaie Shalmani & Khalili

Sabet, 2010; Baragash & Al-Samarraie, 2018; Bezemer, Jewitt, Diamantopoulou, Kress, & Mavers, 2012; Brack, Elliott, & Stapleton, 2004; Cho & Castañeda, 2019; Cloonan, 2010; Lauer, 2009). In other words, learning efficiency is deemed to correlate highly with the number of processes, mechanisms, and channels that are engaged, particularly when the input is available in multiple modalities such as text, image, video, sound, or a combination thereof (e.g. images with embedded texts, also known as *ALT tags*[1]).

When applied to listening comprehension, the idea of presenting input in different modalities to learners seems to underpin Gruba's (1999) connectionist cognitive processing model that suggests that incoming stimuli are processed concurrently by the brain, and that comprehension as a non-additive process involves understanding that is continually modified and revised as more information is presented (Chang, Lei, & Tseng, 2011; Sydorenko, 2010; Wagner, 2007). Along the same line, it has been suggested that learners employ myriad interacting strategies and consult numerous sources of information during the process of comprehension (Buck, 2001). Deep comprehension, then, is believed to be the result of the extent to which different sources of information are combined by the listener in an attempt to reconstruct the original message produced by the speaker. As one medium through which information can be disseminated, visuals and different types of visuals have received considerable attention in the language testing literature, and a number of studies (Chung, 1999; Cross, 2011; Gruba, 2006; Hedge, 2006; Maleki & Safaee, 2011; Sueyoshi & Hardison, 2005) have already shown the positive contributions of visual cues in tests of listening.

For instance, Chung's (1999) study compared the performance scores of Taiwanese students on pre-listening activities under two testing conditions: One group responded to the test items where the aural input was coupled with captions. The second group, on the other hand, responded to the same items under an audio-only condition. The experiment showed that the participants who took the test under a mix of audio and visual stimuli scored significantly higher on the listening test, which suggested that visuals served as a haven for the learners, as they conveyed supplementary data for processing information in dual modalities.

Sueyoshi and Hardison (2005) likewise examined the contributions of gestures and facial cues to second-language learners' listening comprehension of a videotaped lecture by a native speaker of English. The participants consisting of a total of 42 low-intermediate and advanced learners of English as a Second Language were randomly assigned to three stimulus conditions: AV-gesture-face (audiovisual including gestures and face), AV-face (no gestures), and Audio-only. Results of a multiple-choice comprehension task revealed significantly better scores with visual cues for both proficiency levels; however, while the AV-face condition produced the highest scores for the higher level, the AV-gesture-face condition showed the best results for the lower ability students. The researchers contended that learners of varying proficiency levels might need to interact with varying levels of elaboration and details through visuals so as to be able to fully comprehend the aural input.

Similarly, Maleki and Safaee (2011) examined the effects of visuals on participants' comprehension of oral texts through aural stimuli equipped with either paraphrased scripts or static images. The participants were divided into two groups of higher and lower proficiency students and each group was further divided into two subgroups, with one receiving static images with the verbal stimuli and the other receiving script presentations with the oral passages. The study showed that while higher proficiency learners reaped much benefit from textual aids, visual cues proved more advantageous to lower proficiency students. The experiment further revealed that test modules incorporating more static images elicited much better performance than did those including fewer static images. Overall, the study suggested that authentic performance-based

assessment needs to sample behavior within the target language use domain, and this is partly accomplished via the integration of non-verbal components with verbal stimuli in listening comprehension tests.

Capitalizing on Paivio's (1971, 1990, 2007) Dual Coding Theory, Cross's (2011) experiment likewise sought to investigate the role of visual contents in L2 listeners' comprehension of news videotexts. Ten pairs of Japanese English as a Foreign Language (EFL) learners participated in a sequence of tasks requiring them to listen to and discuss various facets of their comprehension of news videotexts. The pairs' dialogue served as the unit of analysis for exploring the effect of visual information on their comprehension. Analysis of the qualitative data suggested that various attributes of the visual content such as audio-visual correspondence impacted their comprehension. Other influences of the visual content that proved to be fruitful were its general utility in facilitating comprehension, its inhibition of attention to and processing of audio information, and its stimulation of learners' expectations and content inferencing.

Whereas there is a great consensus that facial expressions, hand gestures, and body movements combine through visuals to convey meanings extra to the verbal message, which naturally influence both the speaker and the receiver (Baldry & Thibault, 2006; Bateman, 2008; Ventola, Charles, & Kaltenbacher, 2004), precisely how different visual modalities could affect comprehension of aural input is subject to careful investigation. There are studies that have yielded contradictory findings: Baltova (1994), for example, found that visual cues enhanced learners' comprehension in general but did not necessarily stimulate the understanding of the text. Coniam's (2001) study, on the other hand, found that the audio-only group actually scored higher than the video group, albeit the difference was not statistically significant.

Zheng and Samuel (2018) favor a different view by arguing that the benefits afforded by visual stimuli vary depending on some inherent properties of movie clips, such as the distance of the speaker from the listener. In a seminal study, they examined 142 English native speakers' ability to recognize accented speech produced by Chinese speakers of English. The findings revealed that the participants were more accurate at differentiating between words and non-words in English when the video clips zoomed in on the interlocutor's head, but not when they watched the speaker from a far distance. A further finding was that the effect of distance on speech perception proved to be greater when the speech was produced by the person with a strong foreign accent.

Still other scholars like Sweller (2010) argue that the combination of different stimuli in multimedia learning environments might produce the *redundancy effect* where unnecessary or repeated information could increase learners' cognitive load while performing learning tasks. Visuals, then, could have an adverse effect on comprehension of input largely owing to the redundant information they convey, which might ultimately overload learners' working memory. This argument is also reminiscent of what Sweller, Ayres, and Kalyuga (2011) called the *split attention effect* suggesting that presenting material in different modalities could increase the chance of splitting students' attention at the time of information processing, thus maximizing their extrinsic mental load.

One explanation for such mixed results can be the type of visual modalities used in the aforementioned studies. Different scholars have proposed different taxonomies for visuals; however, most confer that visuals are available in two main modalities: *content* and *context* (Bejar, Douglas, Jamieson, Nissan, & Turner, 2000; Ginther, 2002). Context or situation visuals are those that give information about the context for verbal exchanges, such as the participants, settings, and text types. An example of a context visual would be a photo depicting a doctor and a patient conversing in a hospital. Two main functions of context visuals as indicated by Ginther (2002) are to set the scene for verbal interactions and to signal a change in speakers' turns in a conversation.

Content visuals, on the other hand, are visuals that primarily supplement the content of the audio portion of verbal stimuli and may include still images, video segments, diagrams, charts, and so on. A photo of nuclear fission following a lecture on this theme represents a striking example of content visuals.

In a study by Bejar et al. (2000), they classified all content visuals into four groups: content visuals that replicate the audio stimulus; content visuals that illustrate the audio stimulus; those visuals that organize information in the audio stimulus; and those that supplement the audio stimulus. All groups of content visuals were then presented to the participants either graphically or textually, or both graphically and textually. Based on the findings of their experiment, the researchers concluded that while the first three types of content visuals enhanced the comprehension of oral stimuli, the last modality of content visuals actually made it harder for the students to fully comprehend the aural stimuli. The researchers, however, argued that further studies would be required to substantiate such claims.

The findings of research to date, then, seem to have been far from conclusive concerning the role of visual modality in appraising listening comprehension. Much needs to be done to explore whether the inclusion of different modalities of visuals would involve rethinking of the construct being assessed through the tests of listening ability. Of particular interest, then, is to examine whether visual input available in different modalities could engage learners' processing mechanisms to varying degrees, thus affecting the participants' performance scores on multimedia tests of listening.

Aims of the Study

Due to a paucity of research on the role of different visual modalities in comprehension of aural texts, and given the stakes of the decisions made on the basis of the test scores, the present study aimed to compare the effects of two visual modalities, that is, context video of the interlocutors and the setting and context static images of the participants and the settings, on the listening comprehension of EFL learners. The overriding objective was to ascertain whether different modalities of visuals would engage the learners' processing mechanisms to a varying extent, thereby affecting the comprehension scores on the measures of aural ability.

These two modalities of context visuals were selected to examine the tenability of the claim suggesting that the redundancy effect and hence cognitive overload would come about when visuals conveyed information extra, and not complementary, to aural stimuli. Provided that these modalities of context visuals appeared to have contributed to the students' comprehension of aural input, the study would further investigate which of the two modalities would privilege the participants any better on multimedia tests of listening.

Research Questions

The present study sought to find an empirically justified answer to the following questions:

1. Is there a statistically significant difference in the performance level of EFL students who take listening tests with embedded audio and that of those who take listening tests with embedded audio and either of the two modalities (video or images) of context visuals?

2. Is there a statistically significant difference in the performance level of EFL learners who take listening tests with embedded audio and context video and that of those who take listening tests with embedded audio and context images?

Method

Participants

The participants were recruited from the constellation of students who were studying EFL at two Iranian universities. They were selected based on the scores they had obtained on a Cambridge IELTS proficiency test. For the purpose of the present study, intermediate-level learners were chosen, as it was hypothesized that they would not be as experienced as higher proficiency learners to capitalize on their pragmatic ability and world schemata to compensate for potential gaps in their comprehension, and that they would draw heavily on visual cues as an aide to assist with their comprehension of aural stimuli. Beginners, by the same token, might not have served as good candidates, as they are not usually equipped with a wide repertoire of vocabulary, and accordingly, their word knowledge could not greatly influence their comprehension of aural texts.

Participation in the study was voluntary; however, to ensure continuous participation, part of the research budget was allocated to rewarding the students with remuneration. The participants were selected first through availability sampling, then through purposive sampling via the administration of an established proficiency test, and finally, through simple random sampling by the help of a digital randomizer. The final pool of the participants comprised 180 intermediate EFL learners who were randomly assigned to six equivalent groups of participants, each consisting of 30 students. Three groups would serve as the pilot groups, and the other three, as the study groups. Each group also comprised a mix of male and female participants.

Materials and Instruments

The materials used in the study comprised a number of instruments and the base material or content of a listening test. Three different versions of a computerized test of listening were developed by the researcher: For one version, only aural information was available. For the other two versions, however, a combination of aural stimuli and either context video or context still pictures was included in the test. To develop an interface for the multimedia test, SWiSH Max[2], a third-party authoring application, was used. The application is renowned for the development of interactive, flash-based multimedia applications and was hence used for building an interface onto which different components of the computerized test would be loaded. Likewise, to develop the test items, Articulate Quizmaker[3] was used. The application features the capability to develop test modules that allow for automatic calculation of students' scores. The points assigned to each correct answer can be logged in a *shared object* and stored on the computer's hard drive. Shared objects act like *cookies* in Hypertext Markup Language (HTML) programming: Just like cookies that store the information of the websites already seen for faster loading and ready access by the user in a latter time, shared objects store the information that they receive from computer users as they enter it into a flash-based application. Finally, iClone[4], a well-renowned puppeteering tool, was used for the design of animated virtual agents that would appear in the video segments of the test modules.

The aural input consisted of five, two-minute excerpts that were optimized for screen view and were then embedded in the test. The excerpts were taken from online news agencies such as MSN and Reuters, and they all centered on everyday themes such as Global Warming, Human Cloning, E-Waste, and so on. Each excerpt was then followed by 10 comprehension questions in the multiple-choice format. To ensure that variations in prosody, accent, phonology, hesitations, rhetorical signaling cues, and so on would not affect the participants' test performance scores, audio files of the news reported by the same reporter(s) or narrator(s) were used.

For the audio-only version of the computerized test, only directions for attempting the test items followed by the items themselves were presented to the students. The participants were allowed to take notes while listening, as the audio files would be played only once. A built-in countdown timer ensured that all the participants would answer the test items within the same time limit. Furthermore, the participants were allowed to modify their answers only within the specified time limit. A confirmation message would pop up to notify the participants that they could no longer modify their answers once they decided on the correct response. When the limit was reached, the program would automatically move on to run the next modules of the test.

For the other two versions, a similar procedure was followed except that, together with the aural input, the participants could also view either a number of video segments or a series of static images related to the content of the aural stimuli. In the video version, an animation of the reporters having eye contacts with the participants and using natural body language and gestures was shown. As the reporters were reporting the scientists' most recent findings about, for example, the pros and cons of human cloning, video clips of how human zygotes could be cloned through genetic engineering and how cloning could be different from simple cell reproduction appeared on the screen. Likewise, in the still-image version, the same reporters talked about the same topic except that a series of static graphics illustrating identical cells reproduction as well as the steps involved in cloning would pop up every few seconds. Like the audio-only version, a countdown timer would provide control for the amount of time spent on the test items.

In addition to the computerized test of listening, there were other instruments that were employed for participant selection as well as for reporting the degree of the participants' engagement with the listening task. These consisted of a Cambridge IELTS language proficiency test that was used to recruit learners of an intermediate level of language proficiency and a feedback module embedded in the computerized test that would enable the participants to indicate whether, how, and to what extent their mental processes were engaged.

The provision of feedback by the participants essentially aimed to examine their views on the utility of or the affordances, if any, offered to them by context visuals at the time of listening. Likewise, the participants' feedback would allow for estimation of the degree to which topic familiarity, as a potential covariate, could have affected the participants' test performance scores.

Specifically, the participants would be given five minutes at the end of each of the five listening subtests to indicate whether they had found the visual modality used beneficial or distracting. They would also be asked to explain whether and how they were helped or distracted. This was made possible with the help of a pop-up window requiring the participants to input their comments as well as *shared objects* to log their reports in a file on their computers' hard drives for later analysis.

Procedure

The study began with the recruitment and assignment of prospective participants into pilot and study groups. Participants' selection was a tedious process that took approximately two months to

complete. The participants were selected based on availability sampling first: All students who happened to be available and agreed to stay with the researcher over the course of the experiment were enrolled in the study. They were drawn from available EFL classrooms. A purposive sampling technique was then employed to choose the qualified candidates based on their overall band score on a sample copy of the Cambridge IETLS examination papers. One hundred and ninety-two participants, who had obtained 4.5 or 5 on the test, were identified as intermediate-level learners who would be entitled to participate in the study.

An initial estimation of the available facilities (e.g. number of computer terminals, budget constraints, etc.), however, suggested that the number of prospective participants should not exceed 180; therefore, a digital randomizer called SuperCool Random Number Generator[5] was employed to randomly select only 180 participants from the pool of qualified candidates. The same randomizer was then employed to randomly assign the qualified candidates in the final pool to six pilot and six study groups. The participants were required to come to the researcher's computer lab hosting 30 computer terminals.

Next, three versions of the computerized test mentioned earlier, each consisting of five aural passages and a total of 50 comprehension questions in the multiple-choice format, were administered to the pilot groups to undergo standardization. At the researcher's signal, all the students wore their headsets and launched the application's executable by clicking on its desktop icon. The computers came with a copy of the application pre-installed; accordingly, it would take only a while for them to run the computerized tests. The participants were also required to write their names and email addresses in dedicated fields that would pop up on the intro screen. As the participants were attempting the comprehension questions, their scores were updated and stored in a log file on their computers' hard drives.

Once the students took the tests, the statistics of the items were estimated using a digital item analyzer called Test Analysis Program[6] (TAP). The application features the capability to mark defective items with an asterisk, that is, items whose item facility (IF) and item discrimination (ID) indices are not within the desirable ranges ($0.37 \leq IF \leq 0.63$ & $ID \geq 0.40$). Using TAP, one item on the audio-only version, four items on the video-plus-audio version, and two items on the image-plus-audio version were identified as malfunctioning items and were hence excluded from the tests. The participants' scores were then recalculated and the stage was set for the second step of analysis, that is, factor analysis.

Running an exploratory factor analysis (EFA) and drawing on the Principal Components extraction technique (Brown, 2010; Osborne & Banjanovic, 2016), the researcher employed the Statistical Package for the Social Sciences (SPSS) so as to extract all hypothetical factors whose *eigenvalues* fell well above unity. The analysis revealed that, for all three versions of the test, only one factor made the greatest contribution to the test scores variance, suggesting that, in all likelihood, the tests had good construct validly. Precisely, 44.02% of the total variance in the audio-only version, 51.04% in the video-plus-audio version, and 40.26% in the image-plus-audio version of the test was accounted for by a single factor, which was promising. Further evidence supporting this claim came from the analysis of scree plots of eigenvalues. The point at which the screes begin to level off can be deemed as the cut-off point. As can be seen in the following figures, only one scree (factor) has the highest eigenvalue, and this was the case with all three tests:

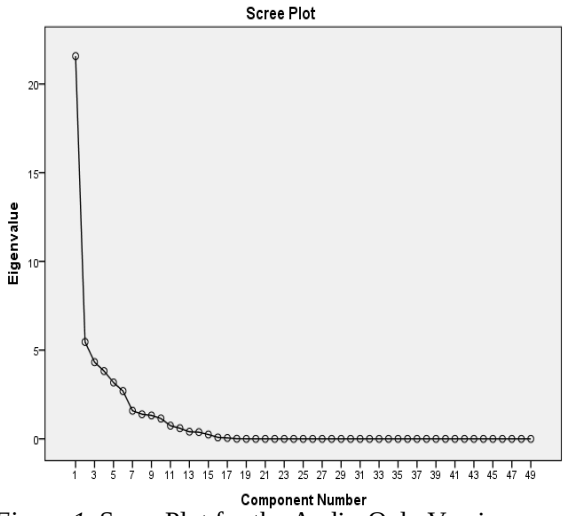

Figure 1. Scree Plot for the Audio-Only Version

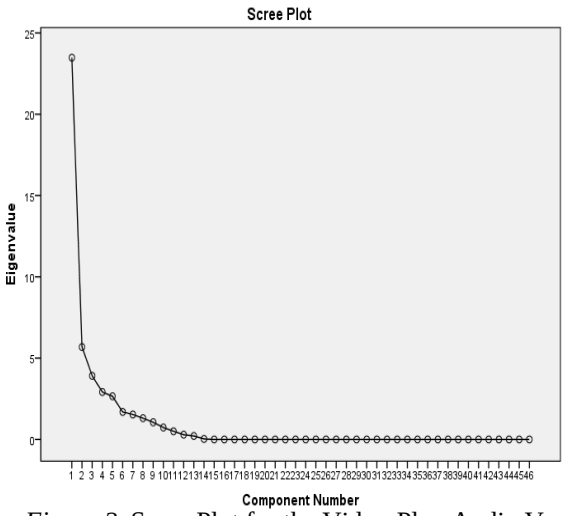

Figure 2. Scree Plot for the Video-Plus-Audio Version

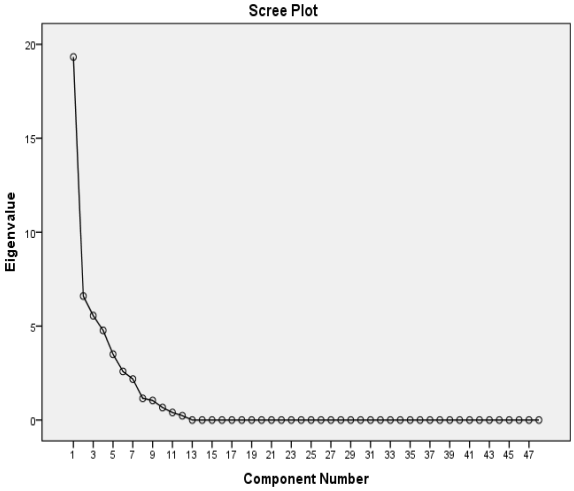

Figure 3. Scree Plot for the Image-Plus-Audio Version

Once the construct validity of the tests was established, their reliability coefficient was calculated through using a Cronbach's alpha. It turned out to be 0.96 for the audio-only version, 0.97 for the video-plus-audio version, and 0.95 for the image-plus-audio version, which was significant. Next, the tests were administered to the three study groups, each receiving one of the three versions. Like the pilot groups, the participants sat at the computer terminals and wore headphones. At the beginning, a robot guide appeared on the screen explaining to the participants how to take the test. The groups then listened to five, two-minute aural texts based on the excerpts taken from online news agencies.

For the experimental groups, one or two virtual agents embarked on reporting the latest news on everyday themes to the participants. For one group, the participants watched an animation of the reporter(s) together with a series of short video clips with relevant but extra content. The content of these videos would supplement, and not complement, the aural stimuli. For the other experimental group, screenshots taken from the animation movies and the accompanying video clips shown to the first group were used instead. Both groups, however, watched two modalities of context visuals, as the main objective was to ascertain whether extra, and not complementary,

information would produce an adverse effect on the participants' comprehension. For the control group, on the other hand, only the voices of the reporters could be heard; no visual cues were available. The following figures show the main interface for the three versions of the computerized test:

Figure 4. The Video-Plus-Audio Version of the Computerized Test. iClone, a powerful puppeteering tool, enabled the researcher to create stunning virtual characters reporting on everyday themes. In this figure, the agents are reporting on Human Cloning when a video clip of doctors working on human cells is shown in the background.

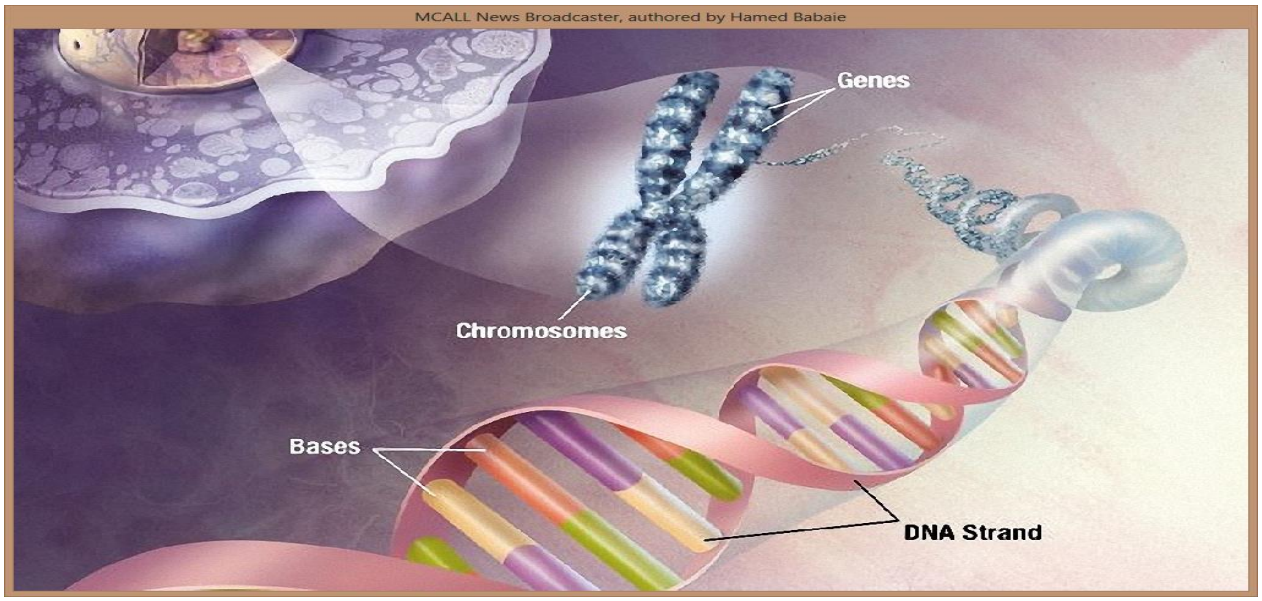

Figure 5. The Image-Plus-Audio Version of the Test. Screenshots relevant but extra to the content of the aural stimuli are being shown. The participants could see these images changing in tandem with the aural information at regular intervals.

All three groups were also given a piece of paper, a pencil, and the option of taking notes while listening. At the end of listening, all three groups answered the comprehension questions. Each passage would be followed by 10 items, and each item correctly answered would receive a score of one mark. Students would spend two minutes listening to the aural input or watching the

avatars and 10 minutes answering the questions followed by each listening passage. Therefore, for five passages, they would spend 60 minutes in total in the testing session. The participants would also have an additional five minutes only at the end of the testing session to review and probably modify all their responses once more before pressing the next button and submitting the answers all at once to signal the end of the session. For each listening session, however, once the time limit was reached, the application would automatically move on to open a pop-up window requiring the participants to indicate whether they had found the visual modality used helpful in aiding their comprehension, or useless or distracting. They had to specify the reason(s) for their responses to this question, too. The following figures show two samples of the comprehension questions:

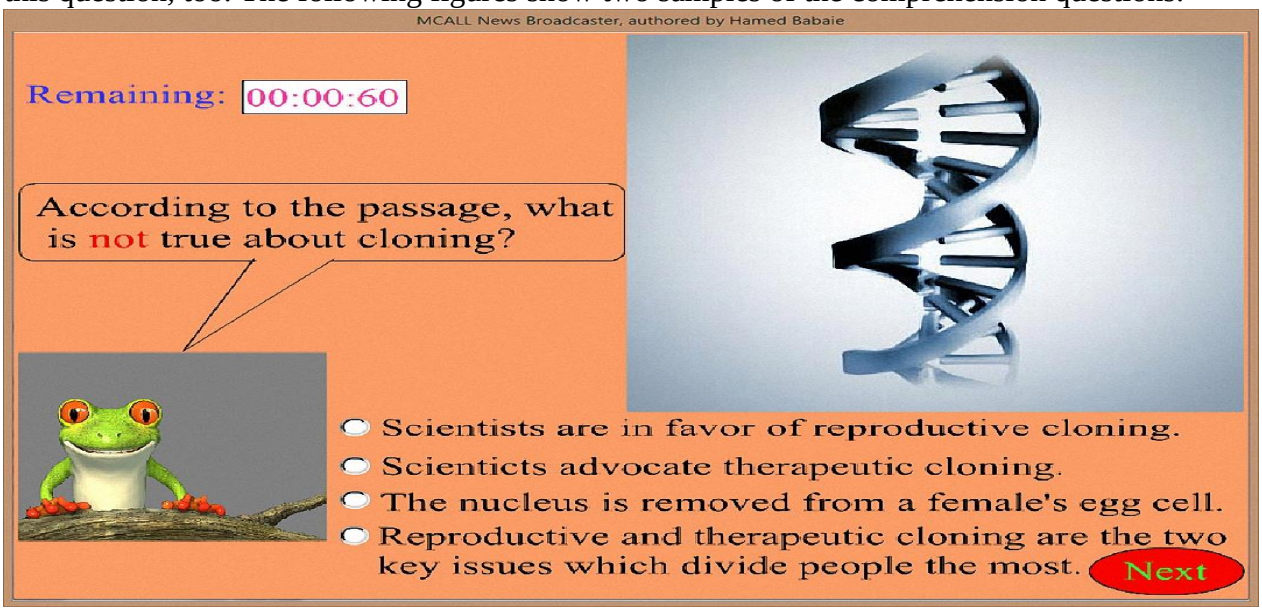

Figure 6. An Example of the Comprehension Questions. The participants could select from among the options and even modify their choices within the time limit. When the Next button was pressed or the limit was reached, they would be automatically taken to the next module.

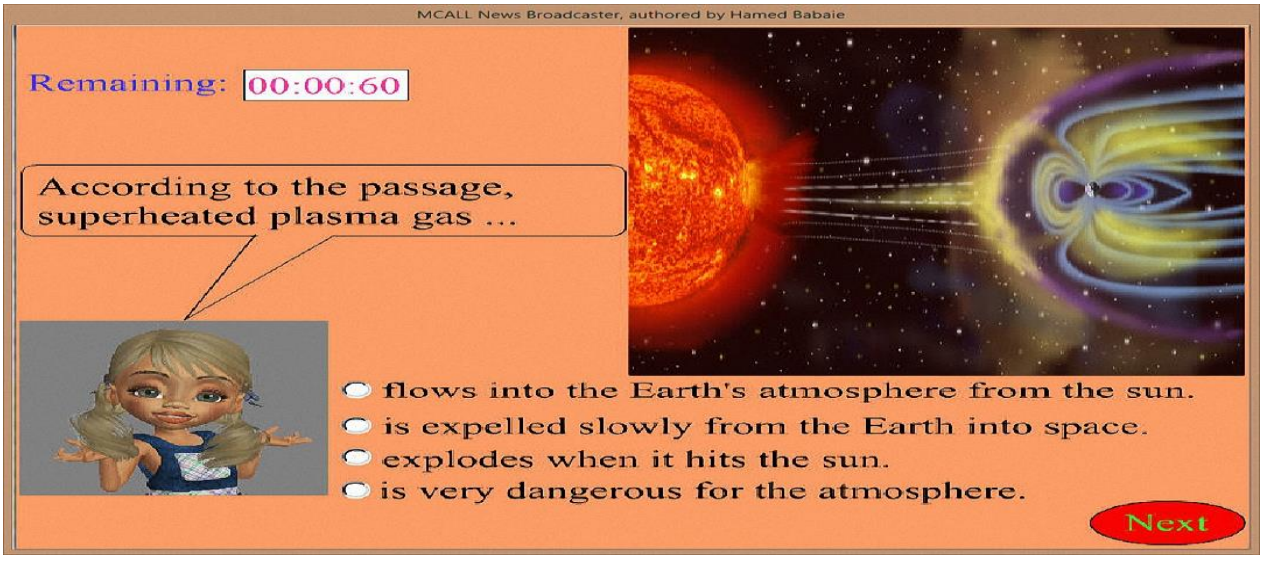

Figure 7. Another Example of the Items Used in the Computerized Test of Listening. ActionScript Programming enabled the researcher to author the computerized test in such a way that it could automatically calculate and store the participants' scores in a log file.

Except for the control group students who were never asked to report on their mental processes, their level of engagement with the aural stimuli, as well as their views on the overall utility, if any, of context visuals for their enhanced comprehension of the message, all the participants in the experimental groups, invariably, were required to answer this question for each passage they listened to and write their answers in the required field before moving on. They were given five minutes to do so. This limit was applied, as a countdown timer would control the total time spent by the participants on answering the test items across the groups such that they would all finish the test at the same time. For the control group, however, these extra minutes served as a short break. Drawing on ActionScript Programing, the computerized test also featured the capability to calculate the students' total scores. This was accomplished by obtaining the aggregate of the points assigned to correct answers for individual students once they attempted the comprehension questions.

At the end of the experiment, the students in all three groups were also asked to indicate their prior familiarity, if any, with the themes of the listening tasks before leaving the lab. This could help with the estimation of the degree to which topic familiarity influenced the participants' test performance scores.

Data Analysis

In an attempt to explore the potential differences between the mean scores of the study groups, a one-way Analysis of Variance (ANOVA) was employed using the latest version (v26.0) of the Statistical Package for the Social Sciences (SPSS). A Scheffé's test was also used to show which means were significantly different from each other.

Results and Discussion

Results

Table 1 presents the descriptive statistics of the participants' scores on the three versions of the computerized test of listening:

Table 1
Descriptive Statistics of the Scores on Computerized Tests of Listening

Descriptives	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Group A	30	41.53	5.361	.979	39.53	43.54	22	45
Group B	30	37.10	5.222	.953	35.15	39.05	25	46
Group C	30	31.57	6.877	1.256	29.00	34.13	19	45
Total	90	36.73	7.104	.749	35.25	38.22	19	46

Note. Group A = video plus audio, Group B = image plus audio, Group C = audio only

As can be seen in the table, the three groups performed quite well on the tests; however, the mean scores varied, suggesting that the three treatment conditions might have differentially impacted the test performance scores of the participants. The mean difference is in favor of the two

experimental groups who listened to multimedia tests of listening with embedded audio and context visuals. A further glimpse at the means, nevertheless, reveals that the participants who watched the animation movies of the scenes as well as of the interlocutors were at a greater privilege, performance-wise, compared to their counterparts in the context images group. To check for the meaningfulness of the mean difference across the study groups and its statistical significance, however, a one-way ANOVA was employed in the latter stage of analysis.

Table 2
Levene's Statistic

	df1	df2	Sig.
1.648	2	87	.198

Note. The Sig. value for the Levene's test is higher than the alpha value ($p > 0.05$); the assumption for using parametric tests is justified.

Homoscedasticity of variances is a precondition for the use of parametric tests such as ANOVA. As can be seen in Table 2 above, the probability value reported for the Levene's test of equality of variances is higher than the preset alpha level ($p > 0.05$), suggesting that the groups' variances were equal.

Table 3
ANOVA Results Reported for the Mean Scores of the Three Study Groups

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1496.067	2	748.033	21.725	.000
Within Groups	2995.533	87	34.431		
Total	4491.600	89			

Note. The Sig. value for ANOVA is smaller than the alpha level ($p < 0.05$); the assumption of homogeneity of variances is violated.

Table 3 shows the results of an ANOVA employed to ascertain where the observed mean difference was statistically significant. As can be seen, the observed value reported for the ANOVA is well beyond unity, and the p-value also is smaller than the preset alpha value ($p < 0.05$), suggesting that the difference between the means were statistically significant and hence meaningful.

Table 4
Results of the Post Hoc Comparison of the Mean Scores

(I) Exp. Groups	(J) Exp. Groups	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
A	B	4.433*	1.515	.017	.66	8.21
	C	9.967*	1.515	.000	6.19	13.74
B	A	-4.433*	1.515	.017	-8.21	-.66
	C	5.533*	1.515	.002	1.76	9.31
C	A	-9.967*	1.515	.000	-13.74	-6.19
	B	-5.533*	1.515	.002	-9.31	-1.76

Note. The Sig. value for each comparison set is smaller than the alpha value ($p < 0.05$), and no confidence interval contains zero; all three means are significantly different.

To show which means were different and which mean difference was statistically significant, a Scheffé's test was also used. As can be seen in Table 4, all three means are significantly different, as the mean difference is statistically significant for each comparison set ($p < 0.05$), which implies that the participants' test performance scores were differentially impacted

by all three versions of the test. Given that the same aural passages together with their follow-up comprehension questions were administered to the three study groups, the format of the test, then, can account for the participants' varying performance on the measures of aural ability.

Discussion

The present study sought to fulfil a two-fold objective: On the one hand, it sought to ascertain whether a statistically significant difference would be found in the performance level of EFL students who took listening tests under an audio-only condition and that of those who took the tests combining audio with either modality (video or images) of the visual stimuli. Analysis of the mean scores obtained on the three versions of the test revealed that the participants had performed differently on these measures, and that those students who sat the visual-plus-audio versions outperformed their counterparts in the control group who took the audio-only version of the test. One explanation that can be given to examiners for their having a potential preference for listening tests with embedded visuals is that a combination of visual and auditory stimuli seems to prove comparatively more fruitful in aiding the students' comprehension of aural texts, as the availability of input in dual modalities is likely to increase individuals' level of engagement in higher-order comprehension processes.

In an attempt to examine whether topic familiarity, too, might have influenced the students' performance on the three measures of aural ability, the participants in the two experimental groups were prompted to indicate whether and to what extent their comprehension of the texts had been affected by this potential covariate by entering their responses in a pop-up window, which would be activated by the feedback module of the tests.

Analysis of the participants' profiles revealed that they had very little familiarity with the topics of the passages they had listened to. Out of the five aural passages, the degree of participants' familiarity reported in percentage for Human Cloning was 11.11% (10 out of 90 students in the three groups); 7.12% for E-Waste; 16.66% for Battling Diabetes; 8.88% for Solar Winds; and finally, for Global Warming, it was reported as 22.22%. Based on the reported percentages, it can be argued that topic familiarity could not have been an issue or could have made very little contribution to the participants' performance on the tests. The idea that the availability of information in multiple modalities might positively influence learning or cognitive processes has been taken up by a number of scholars.

Yanguas' (2009) experiment, for example, compared the effects of different gloss modalities on text comprehension and vocabulary learning among EFL learners. The participants were exposed to multimedia texts under one of four conditions: a no-gloss condition, textual annotation, pictorial gloss, and a combination of textual and pictorial definitions. The study revealed that all the participants who had worked with multimedia glosses could more effectively decipher the meaning of the key terms in the texts and hence arrived at a deeper understanding of the passages largely owing to the availability of information in more than one modality.

The information available in textual or pictorial modality or a combination thereof is assumed to have increased the learners' awareness of unknown terms that were key to text comprehension. Yet, whereas increased noticing, as also indicated by the participants themselves when their think-aloud protocols were analyzed, had focused the learners' attention primarily on the meaning of key words, the study found no significant difference between different annotation modalities in terms of their varying effects on text comprehension. The author, then, suggested that while modality matters (such as when vocabulary information is fetched through both contextual

clues and from any one of annotation modalities), the type of gloss modality could have no sizable effect on the levels of noticing and hence the degree of comprehension.

While experiments like Yanguas' (2009) suggest increased noticing as one likely explanation for the promising effects of modality, some studies (e.g. Babaie Shalmani & Khalili Sabet, 2010) have contended that the use of different modalities may reduce the cognitive demand required to process stretches of texts, when processing information through different memory channels can help learners allocate a greater portion of their working memory capacity to higher-order comprehension processes. The released capacity, due to better management of memory channels, then, accounts for learners' deeper understanding of L2 texts. This argument suggests that while the modality effect comes about when texts are coupled with information in other modalities, the redundancy effect is less likely to happen, in antithesis to mainstream assumptions.

Along the same line, Chang, Lei, and Tseng (2011), in their seminal study on the effects of media presentation modes on listening comprehension and learners' cognitive load, examined Taiwanese students' listening comprehension under two conditions: A single mode when only voices could be heard, and a double mode when aural passages were presented to learners together with their texts. The study found that the participants working with the double mode outperformed those working with the single mode, and that they were subjected to a reduced cognitive load than were those for which only aural modality was available.

Chief among the explanations is the possibility that the presentation of transcripts together with the aural input enabled the participants to more efficiently build up schemas when one channel was dedicated to processing aural information and the other to schema building; cognitive load was lowered and hence their higher scores on the test. This explanation and other similar arguments, then, challenge Sweller's (2010) and Sweller et al.'s (2011) construal of the redundancy and split-attention effects on the grounds that it is not necessarily the case that extra information available in multiple modalities would increase the processing demand and hence cognitive load. When different channels are engaged, processing can be facilitated even tough redundant information exists in one or two modalities.

Extending these assumptions to the present study, it can be argued that redundant information would not necessarily lead to the redundancy effect, and that students' engagement with the listening task can be augmented by the availability of information in multiple modalities that in turn is processed by different memory channels. Three possibilities, then, may help explain the findings of the present study: (a) increased learners' noticing thanks to high levels of engagement caused by visuals; (b) efficient decoding of the aural message due to efficient processing when different channels are addressed, and hence a greater portion of the working memory capacity is allocated to comprehension processes; and (c) reduced cognitive load when one memory channel is allotted to schema building with the help of visuals defining the context and the setting, and the other, to aural stimuli.

Some of these assumptions also reflect the participants' views when exposed to either of visual modalities. In analyzing their profiles, it was found that around 94% of the students in the video-plus-audio group and approximately 82% of the participants in the image-plus-audio group found visuals helpful in aiding their comprehension. Some students declared that, while listening, they could not pick up all the ideas and hence relied heavily on visuals to grasp the message. The following vignette represents one student's idea about the visual modality used:

> *When listening, I couldn't understand some parts of the message; so, I paid attention to many details by looking at the gesture of the interlocutors, the movements of their*

lips, the diagrams presented, and it helped a lot to visualize the context and all the message in mind…

This description suggests that when learners fail to process aural information efficiently, the visual modality may compensate for the failure in some way, perhaps by revealing something more about the topic, or by reducing the cognitive demand in a way that allows for better decoding of the message. While this assumption may hold, some other students believed that they had been more influenced by visuals themselves, as they had found them more engaging, while having no special difficulty understanding the message through the aural modality. One student wrote:

It was really nice when you saw how virtual characters could speak like human beings! The graphics, images, video clips, etc. all were fascinating; everything was great. I was so fascinated I almost forgot I was listening…

This description also implies that while high levels of engagement were involved, the processing of information in the auditory modality could not have been disrupted dramatically by students' noticing the visual stimuli, which further corroborates the idea that learners' cognitive load was reduced owing to the modality effect.

The present study also sought to determine whether the participants' test performance scores would be differentially impacted if different modalities of context visuals were integrated with the aural stimuli in multimedia tests of listening. The study showed that the experimental group who listened to aural passaged with embedded context video had obtained a higher mean on the listening test. This finding can be explained in light of Al-Seghayer's (2001) postulation in whose study the relative effects, if any, of textual and image modalities on learners' both vocabulary learning and comprehension of English narratives were examined. The participants were assigned to one of the three gloss conditions: the textual gloss, video gloss, and pictorial gloss. The results showed that both visual modalities appeared comparatively more fruitful than the textual modality, and that the video modality was even more helpful than the image modality.

Among the suggested explanations was the idea that video builds a more vivid mental image of the concept being introduced and is more likely to arouse students' curiosity, thereby leading to their deep concentration (Al-Seghayer, 2001). Extending these assumptions to the present study, it can be argued that the greater affordances offered by the video-plus-audio version of the test are due to higher levels of student engagement and noticing caused by dynamic visuals where dynamism and lifelikeness might have aroused a deeper sense of curiosity in the participants and in turn contributed to their higher levels of concentration, more efficient processing, and better decoding of information on their part.

While dynamism and lifelikeness are characteristics typically inherent in dynamic visuals, the type of context video employed in the present study contained an extra feature assumed to have further added to its caliber to engender higher levels of motivation in the participants. The virtual characters appearing as reporters in different modules served as efficient attention-getting devices that could have further sustained learners' attention during the experiment (Ahmadi, Sahragard, & Babaie Shalmani, 2017). While these characters also appeared in the context images presented to the second experimental group, they lacked dynamism and hence might have failed to create in the learners the impression that they were listening to real human beings. One ramification of this might have been the participants losing interest in watching them over time.

The smaller percentage reported on the utility of context images (around 82%) by the second experimental group, then, can be ascribed to the lack of dynamism and lifelikeness of the

interlocutors involved in the screenshots. This clearly suggests that context video may serve as a more efficient visual modality to be embedded in tests of listening, where it offers the potential to enhance noticing, levels of engagement, concentration, processing, and decoding of aural input. While all these assumptions may hold, investigation of the relative effects of different visual modalities on students' listening comprehension is still a relatively new line of enquiry that demands careful scrutiny and warrants closer inspection.

Conclusion and Implications

The present study revealed that, thanks to the modality effect, the integration of visuals in computerized tests of listening had proved highly beneficial in aiding learners' comprehension of aural input. Increased noticing contingent upon high levels of engagement; reduced cognitive load resulting from dual memory channels engaged in information processing; and efficient processing and decoding of information are but a few advantages offered by context visuals. Context images, nevertheless, had not proved as fruitful as context videos in sustaining learner motivation and ensuring continuous engagement; lack of dynamism and lifelikeness are among possible explanations.

Considering the greater privilege that can be offered by context videos in multimedia learning environments, the major implications of the present study can be discussed in light of viable theories of multimedia learning and for, at least, three constellations of stakeholders, namely ELT practitioners, language learners, and materials or courseware designers.

As for ELT practitioners, for example, it is recommended that they capitalize on strategies that are tailored to the needs of students in relation to the demands of the learning tasks in which they are engaged (Reichelt, Kämmerer, Niegemann, & Zander, 2014). When it comes to reading and listening comprehension activities, there is a great consensus that effective understanding of the input requires building of schemas as well as the alignment of information gleaned from the incoming stimuli with prior knowledge, if any, of the subject matter (Joh & Plakans, 2017). As such, encouraging learners to draw on mental imagery through using visualization strategies can partly help with schema formation and relatively good understanding of the message; however, more satisfactory results could be obtained when learners watch (and not are asked to only imagine) the scene and the characters involved in the event in a live drawing or an animation movie (Bétrancourt & Benetos, 2018).

It has been contended that the incorporation of context video into tests of L2 listening can increase the authenticity of these measures, largely owing to the fact that learners spend most of their time interacting with their peers or other speakers in face-to-face exchanges, where their understanding is more or less mediated by verbal and non-verbal cues provided by the parties involved in the communicative event (Emerick, 2019). They are, then, helped to demonstrate their ability to effectively reconstruct the message using the cues that can be heard and seen at the same time, in much the same way as they do so in real-life, daily communication.

An additional pedagogical benefit of using multimedia tests with embedded video for language teachers is that multiple forms of assessment can match students' different learning styles, thereby facilitating "student performance at their level of competence by removing the barriers that uncomfortable test formats can create" (Rich, Colon, Mines, & Jivers, 2014, p.7). Multimedia tests are relatively a new form of assessment that has been growing in popularity over the past decades (Hao, 2010). Drawing on Fleming and Mills' (1992) VAK (visual, aural, kinesthetic) learning styles model, Ocepek, Boanic, Serbec, and Rugelj (2013) appreciate the pedagogical value of

multimedia tests, contending that the integration of aural stimuli with other media such as video segments holds a great promise for accommodating diverse learning styles, especially for visual and auditory learners.

As for language learners, the potential benefits of using multimedia tests with embedded video can be discussed in light of the Cognitive Load Theory (Sweller, 2016) suggesting that instruction should be delivered in such a way as to reduce the *extrinsic mental load* and maximize the *germane load* required for efficient internalization of knowledge. When applied to multimedia tests of listening, it can be argued that schema building to help with effective comprehension of the message can be aided partly by the verbal memory, which is responsible for processing aural stimuli, and partly by the visual memory, which is responsible for processing the non-verbal cues available in the visual modality (Guan, Song, & Li, 2018). This would reduce the extrinsic load to the extent that a greater portion of the working memory capacity could be allocated to higher-order comprehension processes as well as seamless integration of information available in dual modalities.

Finally, as for courseware designers aiming to integrate instruction with assessment of language skills in one application suite, it is recommended that the test module include: (a) a video segment to improve student motivation and also to aid with processing of information available in dual modalities, and (b) a video segment that contains information that would complement the audio portion of the listening material. Non-verbal cues, such as various expressions worn by the speaker, character's emotions, forms of address (frontal vs. lateral), the physical appearance or the dressing style (professional- vs. plain-looking) of the speaker, gesticulation, and the like (Beege, Nebel, Schneider, & Rey, 2019) are but a few examples of complementary information not necessarily conveyed or are partially transferred through the audio portion of the listening material. Reduced extrinsic mental load, which would result from the visual memory handling paralinguistic information, provides strong support for the incorporation of context videos into measures of aural ability.

In sum, the findings of the present study favor the incorporation of context videos into multimedia tests of listening, suggesting that high levels of engagement and increased noticing of aural input are assumed to aid with deeper understanding of passages on the part of the learners. Comparative research on the contributions of different modalities of visuals to student learning, however, is still in its infancy; therefore, further studies are required to provide grounds for firmer judgments to be made as to the overall utility of context visuals for language learning, in general, and for L2 listening comprehension, in particular.

Limitations and Suggestions for Future Studies

The present study suffered from some limitations: The sample of participants consisted of a mix of male and female students; however, the number of male and female students who were present in individual groups was not equal, and hence their gender could have played as a covariate, potentially moderating the treatment effect. Future experiments may employ a balanced mix of male and female participants across the study groups or expose the participants of individual groups to different treatment conditions in a counterbalancing design so as to discount the possibility for gender jeopardizing the internal validity of the experiment. Alternatively, a factorial design with groups and subgroups of participants could be employed to allow for the study of both the main and interaction effects, if any, of the variables and also to allow for the examination of the amount of contribution of individual factors to students' comprehension in multimedia tests of listening.

Likewise, it could be intriguing if a similar experiment is replicated with groups of participants comprising mixed-proficiency EFL learners: Since the affordances offered by context visuals were only examined among learners of the intermediate level of language proficiency, future investigations may seek to explore if the effects of visuals in multimedia tests of listening are also moderated by students' linguistic ability.

Of equal interest would be examination of the issue of authenticity in multimedia tests of listening featuring animated agents in video clips: Precisely how and to what extent virtual characters can mimic and indeed emulate human behavior in actual classroom settings demands the design of an experimental study with human teachers and their virtual counterparts interacting in real time, discussing topics across diverse situations, and lecturing under two different treatment conditions. The results can then be compared to examine whether the gain reaped would be comparable across the treatment conditions. The same results can also be obtained and compared in a triangulated study where a comparison could be made across the performance scores and views of different study groups on the utility of context visuals featuring either human or virtual interlocutors. The interview probes may ask about issues such as likability of characters (compared to their human counterparts), lifelikeness and their believability, their cognitive benefits (i.e. the extent to which they engage learners' attention, or the degree to which they aid in the comprehension of the aural message), and other similar issues.

Several scholars, however, have also shown interest in the degree of the *embodiment* of virtual interlocutors in multimedia tutorials and the extent to which they influence students' learning from pedagogical agents. The *embodiment effect* (Mayer & DaPra, 2012) represents the extent to which a virtual tutor can assimilate the behaviors of human teachers in classrooms. It has been contended that a fully embodied affective virtual tutor may create in learners a feeling of partnership comparable to that they develop when interacting with a human teacher (Guo & Goh, 2016). Along the same line, it has been argued that different degrees of agents' embodiment may correlate with varying degrees to which they are perceived as conversational aides, mediating and regulating students' talks like a real partner (Griol, Sanchis, Molina, & Callejas, 2019). In the present study, using iClone, an attempt was made to create highly embodied anthropomorphic virtual characters that would resemble real interlocutors and also demonstrate a wide array of human behaviors, simulating interactions and modeling human non-verbal forms of communication in real-life situations. Future experiments, however, may explore whether degrees of agents' embodiment in multimedia tutorials can correlate with varying levels of learning success, and whether the results can also apply to students' comprehension of aural input in multimedia tests of listening featuring low- and high-embodied virtual characters in context visuals (Lusk & Atkinson, 2007; Mayer, 2014). In a similar vein, it could be interesting to examine whether other attributes such as "androgyny" (Nowak & Rauh, 2005, p. 153) of conversational agents would also affect students' perceptions and hence their performance on multimedia tests of listening in tutorial applications.

Notes

[1] https://www.commonplaces.com/blog/writing-alt-tags-for-images/
[2] http://www.swishzone.com/downloads/SetupSwishmax4.exe?af_id=
[3] http://www.articulate.com/products/quizmaker.php
[4] http://www.reallusion.com/iclone/iclone_trial.aspx

[5] http://www.supercoolbookmark.com/download/supercoolrandom104.zip
[6] http://www.ohio.edu/people/brooksg/downloads/tap.exe

References

Ahmadi, A., Sahragard, R., & Babaie Shalmani, H. (2017). Anthropomorphism–matters or not? On agent modality and its implications for teaching English idioms and design decisions. *Computer Assisted Language Learning*, *30*(1-2), 149-172. doi: http://dx.doi.org/10.1080/09588221.2017.1284132

Al-Seghayer, K. (2001). The effect of multimedia annotation modes on L2 vocabulary acquisition: A comparative study. *Language Learning & Technology*, *5*(1), 202-232. Retrieved from http://llt.msu.edu/vol5num1/alseghayer/default.pdf

Amadieu, F., & Lemarié, J., Tricot, A. (2017). How may multimedia and hypertext documents support deep processing for learning? *Psychologie Française*, *62*, 209–221. doi: http://dx.doi.org/10.1016/j.psfr.2015.04.002

Anastopoulou, S., Sharples, M., & Baber, C. (2011). An evaluation of multimodal interactions with technology while learning science concepts. *British Journal of Educational Technology*, *42*(2), 266-90.

Babaie Shalmani, H., & Khalili Sabet, M. (2010). Pictorial, textual, and picto-textual glosses in e-reading: A comparative study. *English Language Teaching*, *3*(4), 195-203.

Babaie, H. (2008). On the effects of help options in MCALL programs on the listening comprehension of EFL learners. *Journal of Teaching English Language and Literature Society of Iran*, *2*(6), 27-47.

Baldry, A., & Thibault, P. J. (2006). Multimodal corpus linguistics. In G. Thompson, & S. Hunston (Eds.), *System and corpus: Exploring connections* (pp. 164–183). London, UK: Equinox.

Baltova, I. (1994). Impact of video on the comprehension skills of core French students. *Canadian Modern Language Review*, *50*(3), 506-531.

Baragash, R. S., & Al-Samarraie, H. (2018). Blended learning: Investigating the influence of engagement in multiple learning delivery modes on students' performance. *Telematics and Informatics*, *35*(7), 2082-2098. doi: https://dx.doi.org/10.1016/j.tele.2018.07.010

Bateman, J. A. **(**2008). *Multimodality and genre: A foundation for the systematic analysis of multimodal documents.* Basingstoke: Palgrave Macmillan.

Beege, M., Nebel, S., Schneider, S., & Günter, D. R. (2019). Social entities in educational videos: Combining the effects of addressing and professionalism. *Computers in Human Behavior*, *93*, 40-52. doi: https://dx.doi.org/10.1016/j.chb.2018.11.051

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper*. Princeton, NJ: ETS.

Bétrancourt, M., & Benetos, K. (2018). Why and when does instructional video facilitate learning? A commentary to the special issue "Developments and trends in learning with instructional video". *Computers in Human Behavior*, *89*, 471-475. doi: http://dx.doi.org/ 10.1016/j.chb.2018.08.035

Bezemer, J., Jewitt, C., Diamantopoulou, S., Kress, G., & Mavers, D. (2012). Using a social semiotic approach to multimodality: Researching learning in schools, museums and hospitals. *NCRM Working Paper*, *1*(12), 1-14.

Brack, C., Elliott, K., & Stapleton, D. (2004). *Visual representations: Setting contexts for learners*. Paper presented at the 21st ASCILITE Conference, Perth, WA.

Brown, J. D. (2010). Statistics corner. Questions and answers about language testing statistics: How are PCA and EFA used in language research? *Shiken: JALT Testing & Evaluation SIG Newsletter*, *14*(1), 19-23. Retrieved from http://hosted.jalt.org/test/PDF/Brown32.pdf

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

Chang, C., Lei, H., & Tseng, J. (2011). Media presentation mode, English listening comprehension and cognitive load in ubiquitous learning environments: Modality effect or redundancy effect? *Australian Journal of Educational Technology*, *27*(4), 633-654.

Cho, M-H., & Castañeda, D. A. (2019). Motivational and affective engagement in learning Spanish with a mobile application. *System*, *81*, 90-99. doi: https://dx.doi.org/10.1016/j.system.2019.01.008

Chung, J. (1999). The effects of using video texts supported with advance organizers and captions on Chinese college students' listening comprehension: An empirical study. *Foreign Language Annals*, *32*(3), 295-308.

Cloonan, A. (2010). Technologies in literacy learning: A case study. *E–Learning and Digital Media, 7(*3), 248-257.

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.

Cross, J. (2011). Comprehending news videotexts: The influence of the visual content. *Language Learning & Technology*, *15*(2), 44-68.

Emerick, M. R. (2019). Explicit teaching and authenticity in L2 listening instruction: University language teachers' beliefs. *System*, *80*, 107-119. doi: https://dx.doi.org/10.1016/j.system.2018.11.004

Fleming, N. D., & Mills, C. (1992). Not another inventory, rather a catalyst for reflection. *To Improve the Academy*, *11*, 137-155. doi: https://dx.doi.org/10.1002/j.2334-4822.1992.tb00213.x

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

Griol, D., Sanchis, A., Molina, J. M., & Callejas, Z. (2019). Developing enhanced conversational agents for social virtual worlds. *Neurocomputing*, *354*, 27–40. doi: https://doi.org/10.1016/j.neucom.2018.09.099

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

Gruba, P. (2006). Playing the videotext: A media literacy perspective on video-mediated l2 listening. *Language Learning & Technology*, *10*(2), 77-92.

Guan, N., Song, J., & Li, D. (2018). On the advantages of computer multimedia-aided English teaching. *Procedia Computer Science*, *131*, 727-732. doi: https://dx.doi.org/10.1016/j.procs.2018.04.317

Guo, Y. R., & Goh, D. H.-L. (2016). Evaluation of affective embodied agents in an information literacy game. *Computers & Education*, *103*, 59-75. doi: https://doi.org/10.1016/j.compedu.2016.09.013

Hao, Y. (2010). Does multimedia help students answer test items? *Computers in Human Behavior*, *26*, 1149-1157. doi: http://dx.doi.org/10.1016/j.chb.2010.03.021

Hashemi Shahraki, S., & Kassaian, Z. (2011). Effects of learner interaction, receptive and productive learning tasks on vocabulary acquisition: An Iranian case. *Procedia - Social and Behavioral Sciences, 15*, 2165-2171. doi: https://dx.doi.org/10.1016/j.sbspro.2011.04.073

Hedge, T. (2006). *Teaching and learning in the language classroom*. Oxford: Oxford University Press.

interaction. *Journal of Pragmatics, 42*, 116-138.

Joh, J., & Plakans, L. (2017). Working memory in L2 reading comprehension: The influence of prior knowledge. *System, 70*, 107-120. doi: http://dx.doi.org/10.1016/j.system.2017.07.007

Krashen, S. D. (1987). *Principles and practice in second language acquisition*. Hemel Hempstead: Prentice-Hall International.

Kuhl, P. K. (2010). *Brain mechanisms in early language acquisition*. Seattle, WA 98195, USA: Institute for Learning & Brain Sciences, University of Washington.

Lantolf, J. P. (2000). *Sociocultural theory and second language learning.* Oxford: Oxford University Press.

Lauer, C. (2009). Contending with terms: "Multimodal" and "Multimedia" in the academic and public spheres. *Computers and Composition*, *26*, 225-239.

Long, M. H. (1996). The role of linguistic environment in second language acquisition. In W. Ritchie, & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413-468). San Diego, CA: Academic Press.

Lusk, M. M., & Atkinson, R. K. (2007). Animated pedagogical agents: Does their degree of embodiment impact learning from static or animated worked examples? *Applied Cognitive Psychology*, *21*, 747-764. doi: https://dx.doi.org/10.1002/acp.1347

Maleki, A., & Safaee, M. (2011). The effect of visual and textual accompaniments to verbal stimuli on the listening comprehension test performance of Iranian high and low proficient EFL learners. *Theory and Practice in Language Studies*, *1*(1), 28-36. Retrieved from http://www.academypublication.com/issues/past/tpls/vol01/01/05.pdf

Mayer, R. E. (Ed.). (2014). *The Cambridge handbook of multimedia learning*. New York, NY: Cambridge University Press.

Mayer, R. E., & DaPra, C. S. (2012). An embodiment effect in computer-based learning with animated pedagogical agents. *Journal of Experimental Psychology: Applied, 18*(3), 239-252.

Nowak, K. L., & Rauh, C. (2005). The Influence of the avatar on online perceptions of anthropomorphism, androgyny, credibility, homophily, and attraction. *Journal of Computer-Mediated Communication*, *11*, 153-178. doi: https://doi.org/10.1111/j.1083-6101.2006.tb00308.x

Ocepek, U., Bosni, Z., Serbec, I. N., & Rugelj, J. (2013). Exploring the relation between learning style models and preferred multimedia types. *Computers & Education*, *69*, 343-355. doi: http://dx.doi.org/10.1016/j.compedu.2013.07.029

Osborne, J. W., & Banjanovic, E. S. (2016). *Exploratory factor analysis with SAS*. Cary, NC: SAS Institute Inc.

Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart and Winston.

Paivio, A. (1990). *Mental representations: A dual coding approach*. Oxford: Oxford University Press.

Paivio, A. (2007). *Mind and its evolution: A dual coding theoretical interpretation.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Pinnow, R. J. (2011). “I’ve got an idea”: A social semiotic perspective on agency in the second language classroom. *Linguistics and Education*, *22*, 383-392.

Reichelt, M., Kämmerer, F., Niegemann, H. M., & Zander, S. (2014). Talk to me personally: Personalization of language style in computer-based learning. *Computers in Human Behavior*, *35*, 199-210. doi: http://dx.doi.org/10.1016/j.chb.2014.03.005

Rich, J. D., Colon, A. N., Mines, D., & Jivers, K. L. (2014). Creating learner-centered assessment strategies for promoting greater student retention and class participation. *Frontiers in Psychology*, *5*(595), 1-13. doi: http://dx.doi.org/10.3389/fpsyg.2014.00595

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

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, *22*, 123–138. doi:10.1007/s10648-010-9128-5

Sweller, J. (2016). Cognitive Load Theory, evolutionary educational psychology, and instructional design. In D. Geary, & D. Berch (Eds.), *Evolutionary perspectives on child development and education* (pp. 291-306). Switzerland: Springer.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory*. New York, NY: Springer-Verlag.

Sydorenko, T. (2010). Modality of input and vocabulary acquisition. *Language Learning & Technology*, *14*(2), 50-73.

Vandergrift, L. (2003). From prediction through reflection: Guiding students through the process of L2 listening. *The Canadian Modern Language Review*, *59*, 425-440.

Ventola, E., Charles, C., & Kaltenbacher, M. (2004). *Perspectives on multimodality*. Amsterdam: John Benjamins.

Vickers, C. H. (2010). Language competence and the construction of expert-novice in NS-NNS

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

Yanguas, I. (2009). Multimedia glosses and their effects on L2 text comprehension and vocabulary learning. *Language Learning & Technology, 13*(2), 48-67.

Zheng, Y., & Samuel, A. G. (2018). How much do visual cues help listeners in perceiving accented speech? *Applied Psycholinguistics*, *40*(1), 93-109. doi: https://dx.doi.org/10.1017/S0142716418000462