Konin Language Studies
Faculty of Philology, State University of Applied Sciences in Konin, Poland
KSJ 7 (1). 2019. 75-98
http://ksj.pwsz.konin.edu.pl
doi: 10.30438/ksj.2019.7.1.4

The effect of viewing comprehension questions as video captions on test-takers' performance and visual behavior in a second language test

Juan Carlos Casañ-Núñez
University of Valencia, Valencia, Spain
https://orcid.org/0000-0001-6918-9604
juan.casan@uv.es

Abstract

The complexity of while-viewing activities (listening, viewing, reading, and writing simultaneously) has been mostly ignored in the fields of teaching and testing L2 listening. To address this gap, an innovative technique has been proposed: the use of audiovisual comprehension questions imprinted in the video in the form of subtitles. In essence, questions appear on screen a few seconds ahead of the fragment to which they are associated, and they continue on screen until the end of the pertinent scene. This work reports on an approximate replication of Casañ-Núñez's (2017c) study on this methodology. The original study researched if imprinted questions had an impact on second language students' audiovisual comprehension test performance and what examinees thought about this technique. The aim of the replication study is twofold. Firstly, it was designed to confirm or not the results of the original study. Secondly, it investigates for the first time whether imprinted questions have an effect on second language learners' viewing behavior with regard to the video image. As in the original study, imprinted questions do not have a statistically significant effect on test performance, but participants' attitudes towards this technique are positive. The results also suggest that this technique could be an effective way of reducing the conflict of visual attention between watching a video and completing a written task simultaneously. Finally, the limitations of the study are addressed and some possible directions for future research are proposed.

Keywords: audiovisual comprehension; language teaching; language testing; listening comprehension; replication study

1. Introduction

From a theoretical point of view, the use of visuals in education is supported by Mayer's cognitive theory of multimedia learning. According to this theory, people learn better from words and images than from words alone (Mayer, 2014). Audiovisual materials are a common practice in the second language (L2) classroom, and there is a growing number of authors that advocate the use of video in L2 testing (e.g., Suvorov, 2015; Wagner, 2010). Surprisingly, little attention has been paid to what L2 learners do while they complete an audiovisual task. Alderson, Clapham, and Wall (1995) observe that "many students were not actually watching the video monitors: rather, they were reading their answer sheets whilst listening to the soundtrack and responding on the basis of what they heard" (p. 219). Elmankush (2017), Ockey (2007), Suvorov (2015), and Wagner (2007, 2010) quantify the amount of time that examinees watch the videos while they complete the test, and in most cases, the average viewing time is less than 60% of the playing time. Obviously, there is a conflict in visual attention between viewing a video and completing a written task at the same time. As far as we know, this difficulty has been mostly ignored in the fields of teaching and testing L2 listening. To start addressing this gap on research, Casañ-Núñez (2018) has proposed an innovative technique: the usage of audiovisual comprehension questions imprinted in the video in the form of subtitles and synchronized with the relevant fragments, for L2 learning and testing. The current study is part of an ongoing investigation on this technique. Specifically, the current work researches whether imprinted questions have an effect on test-takers' performance and visual behavior in an L2 viewing comprehension test.

2. Literature review

2.1. Definition of listening comprehension

There is not an all-encompassing definition of listening (Worthington & Bodie, 2018). The International Listening Association (1996) defines the skill as "the process of receiving, constructing meaning from and responding to spoken and/or nonverbal messages". This description accounts for the nature of most listening situations, that is, with few exceptions (e.g., listening to the radio), the listener/viewer must attend to both verbal and visual information. This fact has led authors such as Harris (2003) and Lynch (2012) to suggest that the term listening does not reflect the multimodal nature of listening. From now on, in this paper, the compound *listening/audiovisual comprehension* will be utilized to encompass both listening-only and viewing comprehension situations, *listening comprehension* will be used to

refer to listening-only contexts, and both *viewing comprehension* and *audiovisual comprehension* will be employed to describe multimodal conditions.

2.2. Complexity of audiovisual comprehension activities

Many authors suggest the pedagogic exploitation of audio and/or audiovisual texts for L2 learning in three stages: before listening/viewing, while listening/viewing, and after listening/viewing (e.g., Richards & Burns, 2012; Rost, 2016). With regard to the while-listening phase, Underwood (1989) points out that "it is extremely difficult to listen and write at the same time, particularly in a foreign language" (p. 48). The while-viewing phase can be even more demanding. On the one hand, as Vandergrift and Goh (2012) highlight, paying attention to the images, the audio and the task at the same time can lead to working memory overload. On the other hand, there is a conflict in visual attention between looking at a video and completing a written activity simultaneously. Elmankush (2017), Ockey (2007), Suvorov (2015) and Wagner (2007, 2010) have explicitly researched test-takers' viewing behavior in L2 audiovisual comprehension tests. Ockey (2007) and Wagner (2007, 2010) videotape test-takers and measure the amount of time that they take to look at the image while the video is playing. Ockey (2007) finds out that participants (N = 6) watch the video 44.9% of the playing time, Wagner (2007) 69% (N = 36), and Wagner (2010) 47.9% (N = 56). Suvorov (2015) and Elmankush (2017) employ eye-tracking technology to investigate examinees' viewing behavior. Suvorov (2015) researches how test-takers interact with two types of videos: *content* videos (visuals are related to the verbal information) and *context* videos (the image offers information about the communicative situation). The percentage of time that test-takers (N = 33) spend looking at the image in content videos (57.9%) is statistically significantly higher than in context videos (50.7%). Suvorov (2015) suggests that the examinees find content videos more interesting and informative than context ones. In Elmankush's (2017) study, test-takers (N = 30) view the videos 57.4% of the playing time.

The degrees of attention to the image found in these studies are rather low in most cases, and they may have a negative effect on the comprehension of the audiovisual text. According to Sidaty, Larabi, and Saadane (2017, p. 97), "auditory signals can influence our visual perception and vice versa". A well-known example of the influence of vision upon speech perception is the McGurk and MacDonald effect (McGurk & MacDonald, 1976). In this experiment, a woman was filmed while she repeated ba-ba, ga-ga, pa-pa or ka-ka. From the original recording four videos were produced in which the original audio and lip movements were combined in the following way: (1) ba-voice/ga-lips, (2) ga-voice/ba-lips; (3) pa-voice/ka-lips; (4) ka-voice/pa-lips. A hundred and three people were tested

under two conditions: audiovisual, where they watched the videos and repeated what they heard, and audiotory, where they looked away from the screen and repeated what they heard. Under the auditory condition accuracy was high, whereas under the audiovisual condition mistakes were noteworthy. In the field of cinema, Chion (1994, p. 21) explains that sound and image influence each other in audiovisual perception: "sound shows us the image differently than what the image shows alone, and the image likewise makes us hear sound differently than if the sound were ringing out in the dark". For instance, "the same sound can convincingly serve as the sound effect for a crushed watermelon in a comedy or for a head blown to smithereens in a war film" (Chion, 1994, p. 23).

Cognitive load theory (Jiang, Kalyuga, & Sweller, 2017; Sweller, Ayres, & Kalyuga, 2011) is an instructional theory grounded in the knowledge of the human cognitive architecture. Its purpose is "to maximize construction of organized knowledge by managing the cognitive load experienced by the learners" (Jiang et al., 2017, p. 2). According to this theory, the cognitive load required by learning tasks can be divided into two categories: intrinsic and extraneous. The intrinsic cognitive load is imposed by the inherent nature of the materials, and the extraneous cognitive load by the way in which the information is presented. The former is fixed, whereas the latter can be altered "by changing instructional designs to increase learning efficiency and to maximize learning outcomes" (Jiang et al., 2017, p.3). Intrinsic and extraneous cognitive load are additive and they define the total cognitive load needed to process learning materials (Sweller et al., 2011). If the total cognitive load surpasses "the available resources of working memory, the cognitive system will fail, at least in part, to process necessary information" (Sweller et al., 2011, p. 58). We consider that this theory can be applied to L2 while-viewing activities. As watching video materials in an L2 already imposes a high intrinsic cognitive load, especially to non-competent L2 viewers, it follows that while-viewing activities must be planned so as to generate a low extraneous cognitive load.

2.3. Imprinted questions

Casañ-Núñez (2018) proposes an innovative technique that may help reduce the difficulty of while-viewing activities in the fields of L2 teaching and testing: the use of audiovisual comprehension questions imprinted in the video in the form of subtitles and synchronized with the relevant fragments. In essence, questions appear on screen a few seconds ahead of the fragment to which they are associated and they continue on screen until the end of the pertinent scene. Questions involve scarce reading: up to two lines and up to 37 characters per line. Formal features (lines, position, font type, reading time, etc.) are based on established

conventions in the use of subtitles according to Díaz Cintas (2013), Díaz Cintas and Remael (2007) and the Spanish Association for Standardization and Certification (AENOR, 2012). A film scene with imprinted questions is available at https://youtu.be/ALw8XJkrbDQ. This technique may have some benefits over viewings where the comprehension activity is available only on paper:

1. Theoretically, it reduces the conflict in visual attention between viewing the video and completing the task, by temporally and spatially approximating the questions and the relevant fragments.
2. Hypothetically, the fact that questions are synchronized with the relevant fragments helps learners or test-takers to focus their attention on the pertinent scenes of the video.
3. Theoretically, it simplifies the activity, as students would only need to pay attention to one imprinted question at a time, rather than having to concentrate on several questions simultaneously on paper.
4. Hypothetically, in the event that a student loses focus, he/she is helped to continue the activity when the next imprinted question appears on screen.
5. Theoretically, the previous benefits lower the extraneous cognitive load of the activity.
6. In other words, this technique may free up cognitive resources that some learners may need to process the audiovisual text and answer the comprehension questions.
7. Overall, L2 students have positive opinions about this technique (Casañ-Núñez, 2017c; Torres-Salvador, 2019).

This methodology is considered especially beneficial for adult unskilled listeners/viewers (A1, A2 and B1 levels according to the *Common European Framework for Languages*, Council of Europe, 2018). This is so because, according to Field (2008), non-competent listeners/viewers employ a great amount of working memory in decoding, and they may reach working memory overload faster than competent listeners/viewers. The strategy does not aim to replace other practices with video but to complement them.

The proposal is the result of an ongoing multiphase mixed-methods investigation (Creswell, 2014) composed of several studies. As part of this investigation, Casañ-Núñez (2017c) conducted a multimethod study (Morse, 2003) with two main objectives: (1) enquire whether imprinted questions had an impact on Spanish L2 students' audiovisual comprehension test performance, and (2) ascertain what test-takers thought about this methodology. 41 Spanish L2 learners were involved in the study (22 in the control group and 19 in the treatment group). As for the first research objective, the author hypothesized that participants who took a viewing comprehension test with questions imprinted in the

video (experimental group) would outperform participants that sat the same test without them (control group) thanks to the first five theoretical benefits of imprinted questions. Nevertheless, a null hypothesis of no difference was tested. The results did not confirm the hypothesis that the experimental group would outperform the control group in the test. Concerning the second research objective, on the grounds of the first five theoretical benefits of imprinted questions, the author hypothesized that examinees of the experimental group would have positive attitudes towards this methodology, and the results confirmed this hypothesis. Casañ-Nuñez's (2017c) study was the first and only that researched whether imprinted questions had any impact on L2 learners' audiovisual comprehension test performance, and there is no other work to compare the results to. As a future research line, the researcher suggested investigating whether imprinted questions had an effect on L2 learners' viewing behavior with regard to the video image. On the basis of the first benefit of imprinted questions, the author hypothesized that this technique would make it easier for students to watch a video while they simultaneously conduct a task. This is an important matter because, as mentioned before, studies that researched examinees' visual behavior in L2 audiovisual comprehension tests found rather low degrees of attention to the video image.

2.4. Purpose statement and research questions

In light of the difficulty of L2 audiovisual comprehension while-viewing activities, little attention paid to this issue, the theoretical benefits of imprinted questions, the scarce research on the impact of imprinted questions on comprehension and the lack of research of its effect on students' viewing behavior, it was decided to conduct an approximate replication (Porte, 2012) of Casañ-Nuñez's (2017c) study. The aim was twofold. Firstly, it was designed to confirm or not the results of the original study. Secondly, it investigated for the first time whether imprinted questions had an effect on L2 learners' viewing behavior with regard to the video image. The following research questions (RQ) were addressed:

1. RQ 1: Does the use of questions imprinted within the video in the form of subtitles and synchronized with the relevant fragments in an L2 audiovisual comprehension test have an impact on test-takers' performance?
2. RQ 2: To what extent do test-takers agree or disagree that the synchronized video-imprinted questions helped them to complete the test?
3. RQ 3: Does the use of synchronized video-imprinted questions have an impact on test-takers' viewing behavior?

RQs 1 and 2 were drawn from the original study, whereas RQ 3 was an addition to the replication study. Similarly to Casañ-Nuñez's (2017c) study, on the grounds of the first five potential benefits of imprinted questions, it was foreseen that examinees that took an audiovisual comprehension test with the items available on paper and the stems of the questions imprinted in the video (the experimental group) would a) outperform test-takers that sat the same test with the questions available only on paper (the control group) and b) have positive attitudes towards this methodology. On the grounds of the first theoretical benefit, it was hypothesized that test-takers would view the audiovisual texts longer than participants in the control group. Nonetheless, as in the original study, null hypotheses of no difference were tested.

If the data in the current study supported the hypotheses, it would have a considerable impact on L2 learning and testing. It would suggest that this technique is useful for lowering the visual attention conflict and the extraneous cognitive load of while-viewing activities. This would be particularly beneficial for L2 non-competent listeners/viewers.

3. Method

This approximate replication study (Porte, 2012) closely complied with the research procedures in Casañ-Nuñez's (2017c) study with one exception: the investigation of students' visual behavior was added in the replication.

3.1. Study design

As in the original study, a multimethod design (Morse, 2003) was used. It involved the collection of four quantitative datasets that were employed to respond to the various research questions. First of all, participants were surveyed to get to know the sample and some of their preferences concerning listening/audiovisual comprehension. After that, a viewing comprehension test was administered to investigate if test-takers of the control and experimental groups performed differently. Participants were video recorded while they took the test to find out if there were differences in visual behavior between the control and experimental groups. Lastly, attitudinal data towards the usage of imprinted questions from the treatment group was gathered through a questionnaire.

3.2. Participants

As in the original study, the sample was selected by a convenience, non-probabilistic sampling method (Battaglia, 2008). Whereas in the original study participants

were Spanish L2 learners at the Universidade de Coimbra (Portugal), informants in the replication were Spanish L2 students at the Universität Rostock (Germany). The sample was composed of 28 persons: 24 were taking the Teaching Degree in Spanish (*Lehramt Fach Spanisch)* and 4 the BA in the Spanish language, Literature and Culture (*Bachelor Spanische Sprache*, *Literatur u. Kultur)*. Fourteen were enrolled in *German-Spanish Translation III* and 14 in *Grammar and Writing I*. On the grounds of the teachers' observations, it was estimated that most students attending *German-Spanish Translation III* had a B2 level (according to the *Common European Framework of Reference for Languages*, Council of Europe, 2018) in viewing comprehension and most participants enrolled in *Grammar and Writing I,* a B1 level. Randomly, those enrolled in each subject were split into two clusters of seven. Next, in an aleatory way, one cluster of each subject was assigned to the control group and the remaining one to the treatment group. All participants were between 18 and 29 years of age except for one in the experimental group who was between 45 and 49. All were German native speakers except for one student in the control group whose mother tongue was French. Roughly speaking, the groups had been studying Spanish for a similar amount of time (control group, M = 5.78 years, SD = 3.167 years; experimental group, M = 5.43 years, SD = 3.050 years). On average, the control group had spent less time in Spanish speaking countries than the treatment group (control group, M = 9.14 weeks, SD = 8.87 weeks; experimental group, M = 12.86 weeks, SD = 7.65 weeks).

As for learners' preferences concerning listening/audiovisual comprehension, some points can be highlighted. First, almost all participants (26/92.8%) reported that practicing listening/audiovisual comprehension in class was either "very" important or "very much" important. The results were very similar to those reported in the original study (100% answered "very" important or "very much" important). Second, most learners did not show a preference for audio or video materials for practicing listening/audiovisual comprehension in class (13/46.4%), and many more preferred video (12/42.9%) over audio (3/10.7%). The results are in line with those of the original study: 58.3%, 30.6% and 11.1% answered "both equally", "video" and "audio", respectively. Finally, most participants (19/67.9%) reported that the visual information was "very" or "very much" helpful to comprehend the speakers and 8 (28.6%) had no clear opinion and answered "neutral". The results are close to those of the original study (the matching percentages were 72.3% and 25%).

Table 1 summarizes the demographics in the replication and the original studies. Table 2 shows the number of participants who completed each instrument and were video recorded in the replication.

Table 1 Main characteristics of participants in the replication and the original studies

	Studies	
	Replication	Original
Selection	Convenience samplings	
Number	28	41
Context	Universität Rostock	Universidade de Coimbra
Age	Mostly between 18 and 29	Mostly between 18 and 24
Mother tongues	27 German speakers and 1 French speaker	40 Portuguese speakers and 1 Ukrainian speaker
Estimated proficiency levels in Spanish L2 audiovisual comprehension	B1 and B2	B1+

Table 2 Number of participants who answered each instrument and were videotaped in the replication

	Group	
	Control	Experimental
Answered pre-test questionnaire	14	14
Completed the audiovisual comprehension test	14	14
Responded to the post-test questionnaire (only for the experimental group)	n/a	14
Were video recorded	14	12[a]

[a]. The video camera was placed at the back of the classroom, and it did not fully record two participants in the front row because others were blocking the view.

3.3. Materials

Three instruments were used to collect the data: two questionnaires and a viewing comprehension test. As in the original study, materials were written in the L2 being studied (i.e., Spanish) for two reasons: first, participants may have different L1s; and second, it was considered that they were proficient enough to understand Spanish. However, they were allowed to answer either in Spanish or their L1.

3.3.1. Pre-test questionnaire

The pre-test questionnaire (available on request) was employed to gather specific information about the sample group: sociodemographic information, time learning Spanish, time spent in Spanish speaking countries, and some learning preferences regarding listening/audiovisual comprehension. It was composed of twenty-four items. Following the proposal of Saris and Gallhofer (2014), it used open and closed questions. The survey was very close to the one administered in the original study. The only differences lay in the fact that some sociodemographic items were adapted to the new context and that the translations of some of the words that may be difficult for German learners were added in brackets. These potentially problematic words were identified and translated

with the help of Claudia Ascher (Universidade de Coimbra), a German lecturer with knowledge of Spanish. The design and elaboration of the original questionnaire involved expert review, a pilot study, and a reliability analysis. A detailed account of its construction is available in Casañ-Núñez (2017a).

3.3.2. Audiovisual comprehension test

The viewing comprehension test was identical to the one administered in the original study. A detailed account of its elaboration and the instrument itself can be found in Casañ-Núñez (2016). The control and the experimental groups were administered the same test in different ways. In both groups, the items were available on paper. In addition, in the experimental scenario, the stems of the questions were imprinted within the video in the form of subtitles and synchronized with the relevant fragments. The target language use domain was to understand informal conversations belonging to the personal domain in Spanish romantic comedies. The test evaluated extracting specific information, identifying general ideas and recognizing feelings in face-to-face informal talks. The test was made up of two tasks, two film scenes and seven question items. The key characteristics of the input texts and the experimental items can be found in Tables 3 and 4. In order to specify the fragments, an 8-digit time code was employed. The first three pairs of digits relate to hours, minutes and seconds, respectively, and the last pair, to frames.

Table 3 Main features of the audiovisual texts

1. Text source	Videotext (source: *CEFR* p. 49) Scene from the comedy *Ópera prima* by Fernando Trueba where an informal conversation takes place. Location: from 00:00:00:00 to 00:02:36:12.	Videotext Scene from the comedy *Los peores años de nuestra vida* by Emilio Martínez Lázaro where an informal conversation takes place. Location: from 00:11:10:00 to 00:12:48:15.
2. Authenticity	Genuine	Genuine
3. Domain type (source: *CEFR* page 45)	Personal	Personal
4. Text length	2 min and 36 s	1 min and 38 s
5. No of participants	2	4
6. Text speed (global impression)	Normal	Fast
7. Accent (all participants)	Standard	Standard
8. How often played	Once	Once
9. Estimated level	A2/B1	B2/C1

Abbreviation: *CEFR*, *Common European Framework of Reference for Languages*

Table 4 Description of the imprinted questions in the fragments from *Ópera prima* (1-4) and *Los peores años de nuestra vida* (4-7)

Imprinted question	Focus	Question type	Timing	No- of lines	Font and size	Color	Estimated level
1. ¿Cómo se llama la chica? [What is the girl's name?]	ESI	Multiple-choice	00:01:10:00 00:01:37:09	1	Arial Narrow 36	White (hex color code #FFFFFF)	A1
2. ¿De qué hablan? [What are they talking about?]	IGI	Multiple-choice	00:01:44:14 00:02:02:02	1			A2/ B1
3. ¿Cuál es el teléfono de la chica? [What is the girl's telephone number?]	ESI	Open question with only one possible answer	00:02:04:04 00:02:22:00	1			A2
4. ¿Qué sentimientos puede tener el chico por la chica? [What feelings may the boy have for the girl?]	RF	Open	00:02:27:08 00:02:36:12	2			A2/ B1
5. ¿De qué temas habla el chico pelirrojo? [What is the redheaded boy talking about?]	IGI	Multiple-choice	00:11:40:00 00:12:25:24	1			B2
6. ¿Cómo se llama la chica? [What is the girl's name?]	ESI	Multiple-choice	00:12:39:04 00:12:43:24	1			A1
7. ¿Qué sentimientos puede tener el chico pelirrojo por la chica? [What feelings may the red-haired boy have for the girl?]	RF	Open	00:12:45:00 00:12:48:15	2			B2

Abbreviations: ESI, extracting specific information; IGI, identifying general ideas; RF, recognizing feelings.

3.3.3. Post-test questionnaire

The post-test questionnaire (available on request) was designed to find answers to the second group of research questions. It closely resembled the questionnaire used in the original study. The questionnaire had three parts: title, introduction, and items. The title tried to be as informative as possible. The introduction outlined the purpose of the instrument, pointing out that there were no right or wrong responses, and that answers would be treated confidentially. The instrument employed two types of questions: 9 Likert-type items that made up a scale and 2 open queries. The alternatives chosen for all the Likert-type questions were agree, neither agree nor disagree and disagree. The 9 Likert-type items made up a scale designed to answer research question 2: *To what extent do test-takers agree or disagree that the synchronized video-imprinted questions helped them to complete the test?* As in the original study, the scale was named "helpfulness" scale. Finally, the questionnaire contained two open text items so that participants could express their views freely. A detailed description of the questionnaire can be found in the original study.

3.4. Video-recording equipment and software

In order to study participants' visual behavior during the test, they were videotaped with a digital camera placed on a tripod. Videos were recorded at a resolution of 1280 by 720 at 29 frames per second. VLC media player on a Windows computer was used by the researcher to reproduce the videos and analyze viewing behavior.

3.5. Procedures

The procedures in the original and the replication studies were closely comparable with the exception that participants were video-recorded while they completed the test in the replication study. The study involved separate administrations of the instruments to the control and the experimental group. All materials were administered by the author. To begin with, participants of both groups filled out the pre-test questionnaire. They had unlimited time to respond and the researcher was available to answer questions related to the questionnaire itself. After that, the viewing comprehension test was administered. The control group took the test with the items available on paper, while the treatment group completed the test with the items available on paper and with the stems of the questions imprinted in the video in the form of subtitles. All of the participants received directions on the test structure and a short presentation of the input texts. Besides, the experimental group received an explanation of how imprinted questions worked. A full account of these instructions can be found in Casañ-Núñez (2016, pp. 21-22). Test-takers were also informed that the input texts would be played once and that there was no time limit to answer the questions. A laptop, two speakers, a video-projector and screen were used to reproduce the input texts. Participants were video-recorded while they took the test. The camera was placed at the back of the classroom and not in front because some participants preferred it this way, thus safeguarding anonymity. Finally, the second questionnaire was administered to the experimental group. Again, participants had unlimited time to answer, and the author was available to clear any doubts arising from the questionnaire.

The test was hand-scored by the author. Multiple-choice items and the open question with one possible answer (1, 2, 3, 5, 6; see Table 4) were scored right or wrong. In order to determine if the answers to open-response items (4 and 7, see Table 4) were satisfactory or unsatisfactory, the strategy proposed by Buck (2001) was employed: administering the test to a group of skilled listeners and using their answers as a basis for evaluating if responses were acceptable (p. 141). This trialing is described in Casañ-Núñez (2016). Correct and satisfactory responses were assigned 1 point and incorrect or unsatisfactory answers 0 points. The test was graded a second time to check accuracy.

3.6. Data analyses

In order to quantify for how long test-takers looked at the video screen while the videos were playing, the following strategy was used. The researcher monitored each examinee on the video recording using a chronometer. When a test-

taker oriented his/her head towards the screen while the video was playing, the timer counted the time, and when the examinee looked away from the video screen, the stopwatch was paused. The procedure was repeated twice for each test-taker, and the average of both measures was taken into consideration. The mean was regarded as an improved accuracy measure because human reaction time varies "from moment-to-moment" (Birmingham & Taylor, 2005, p. 375), and the chronometer was manipulated by the researcher. This counting strategy assumed that when a test-taker oriented his/her head towards the screen, he/she was watching the video. 28 participants took the test. However, only 26 could be analyzed because the camera did not adequately record 2 of them due to the fact that others were blocking the view.

Table 5 Quantitative data analyses used for answering research questions

Objective/ Research question	Data	Analyses
Getting to know the sample	Pre-test questionnaire items	Frequencies
RQ1	Audiovisual comprehension test data	
	Items scores in the traditional and the experimental variants of the test	Reliability analyses (Cronbach's alpha, Cronbach's alpha if item deleted, mean inter-item correlation values)
	Test scores in the traditional and the experimental variants of the test	Outlier detection (boxplots), descriptive statistics, tests of normality (Shapiro-Wilk tests), independent samples *t*-test comparing test scores, effect size (Pearson's correlation coefficient)
RQ2	Post-test questionnaire data	
	Helpfulness scale items	Reliability analysis (Cronbach's alpha, Cronbach's alpha if item deleted, and mean inter-item correlation value)
	Helpfulness scale score	Outlier detection (boxplot) and descriptive statistics
RQ3	Visual behavior data	
	Variable *percentage of time viewing the videos* in control and experimental groups	Outlier detection (boxplots), descriptive statistics, tests of normality (Shapiro-Wilk tests), independent samples *t*-test comparing visual behavior patterns, effect size (Pearson's correlation coefficient)

All quantitative statistical analyses (see Table 5) were computed in SPSS version 21, except for effect sizes, which were calculated using Windows 7 calculator in scientific mode. Data was double-checked to ensure accuracy. In addition to that, frequency analyses of all variables were performed to detect missing or anomalous values. None was found. Table 5 sums up the quantitative data

analyses that were undertaken to answer each research question. The post-test questionnaire contained two open-ended items. As few participants made comments and these remarks were short, the only strategy employed to analyze qualitative data involved quantizing related answers. Statistical analyses were in line with the original study with the exception of the analysis of the visual behavior data, not present in Casañ-Núñez's (2017c) study.

4. Results and discussion

4.1. Research question 1

The first research question was *Does the use of questions imprinted within the video in the form of subtitles and synchronized with the relevant fragments in an L2 audiovisual comprehension test have an impact on test-takers' performance?* First, Cronbach's alpha and Cronbach's alpha if item deleted (CAID) values were calculated. Alpha value was .657 in the traditional variant of the test and .641 in the experimental one. According to Suhr and Shay (2009), alpha values for instruments designed for research purposes can be as low as .60 (p. 3). CAID values were between .591 and .669 in the traditional test and between .518 and .641 in the experimental one. This means that none of the items would considerably augment the reliability of the tests if they were deleted. The alpha value is affected by the number of items on the scale (Cortina, 1993) and "with short scales (e.g., scales with less than ten items) it is common to find quite low Cronbach values (e.g., .5)" (Pallant, 2013, p. 101). For this reason, Pallant (2013)

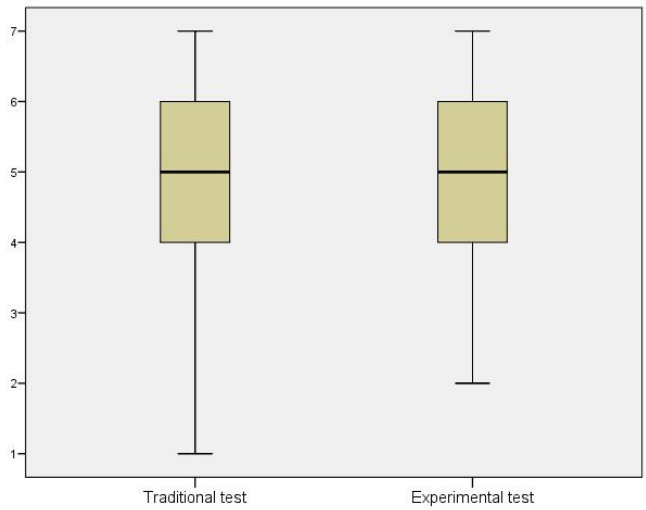

Figure 1 Boxplots of the test scores

suggests reporting the mean inter-item correlation value (MICV) for scales with few items. The MICW was .243 in the traditional test and .216 in the experimental one. These values were appropriate according to Briggs and Cheek (1986): "the optimal level of homogeneity occurs when the mean inter-item correlation is in the .2 to .4 range" (p. 115). As the reliability analyses were satisfactory, scores containing the sum of the component items were created. Next, the test scores were checked for outliers and none was found (see Figure 1).

As can be seen in Table 6, on average, examinees that took the test with the questions imprinted in the video as subtitles (M = 5.07, SD = 1.542) scored higher than the test-takers that took the same test without them (M = 4.64, SD = 1.598) by 0.43 points. Since the distribution of the test scores was normally distributed both in the traditional test ($S\text{-}W$ = .932, df = 14, p = .329) and in the experimental one ($S\text{-}W$ = .933, df = 14, p = .336), it was legitimate to use an independent-samples t-test to find out if the difference was statistically significant. The t-test revealed that the difference was not statistically significant: t (26) = -.722, p (two-tailed) = .477, r = 0.14. The results were similar in the original study: the difference between the test scores was not statistically significant (U = 203.500, z = -.147, p [exact, two-tailed] = .885, r = -0.023). Thus, the results of both studies did not validate the hypothesis that the treatment group would score higher than the control one in the viewing comprehension test thanks to the theoretical benefits of imprinted questions. Let us examine the foundations of these hypothetical benefits. The first one consists in reducing the conflict in visual attention between viewing the video and completing the task, by temporally and spatially approximating the questions and the relevant fragments. This issue was researched for the first time in the current study, and the results suggest that imprinted questions may reduce the visual attention conflict (see section 3.3.). The second and the fourth benefits have not been proven to be true, but researchers and Spanish L2 teachers agree with them (see Casañ-Núñez, 2015, 2017b). The third and the fifth possible benefits (imprinted questions simplify the activity and lower the extraneous cognitive load of the activity) have not been proven to be true, but they seem plausible. Finally, test-takers' view about this technique both in the current study (see section 3.2) and in the original study was positive. Thus, there is an apparent mismatch between test performance and attitude towards this technique. All this suggests that the hypothesis was well-grounded and that it did not hold in the original and the replication studies due to a given reason. First, some L2 listeners/viewers are better than others at making use of visual information (Wagner, 2008). It could be the case that test-takers in the control group were better at using non-verbal information than participants in the experimental group. In fact, in the replication study, there was a lot of variation in viewing behavior: between 43.30% and 80.78% of the playing time in the control group and between 58.10% and 79.09% in the experimental

group (see table 8 in section 4.3). Second, perhaps imprinted questions neither aid nor hinder comprehension when presented to test-takers together with a printed version because they ignore the former.

Table 6 Descriptive statistics for the tests

Test	*N*	Min	Max	*M*	*SD*	Median	Skewness	SE Skewness	Kurtosis	SE Kurtosis
Trad.	14	1	7	4.64	1.598	5	-.764	.597	.577	1.154
Exp.	14	2	7	5.07	1.542	5	-.431	.597	-.488	1.154

Abbreviations: Trad. = traditional; Exp. = Experimental.

4.2. Research question 2

The second research question was: *To what extent do test-takers agree or disagree that the synchronized video-imprinted questions helped them to complete the test?* The helpfulness scale was employed to answer this. First of all, its reliability was checked. Cronbach's alpha was .767. According to Cohen, Manion, and Morrison (2011), this alpha value suggests that the scale is reliable. Cronbach's alpha if item deleted values were between .692 and .773. In other words, none of the items would significantly increase the reliability of the scale if they were removed. The mean inter-item correlation value was .246, which is appropriate according to Briggs and Cheek (1986). As the reliability was adequate, a new variable containing the sum of the component items was generated. To that end, each "agree" was added up as 3 points, each "neither agree nor disagree" as 2 points, and each "disagreement" was as 1 point. Afterwards, the scale was checked for outliers. None was found (see Figure 2).

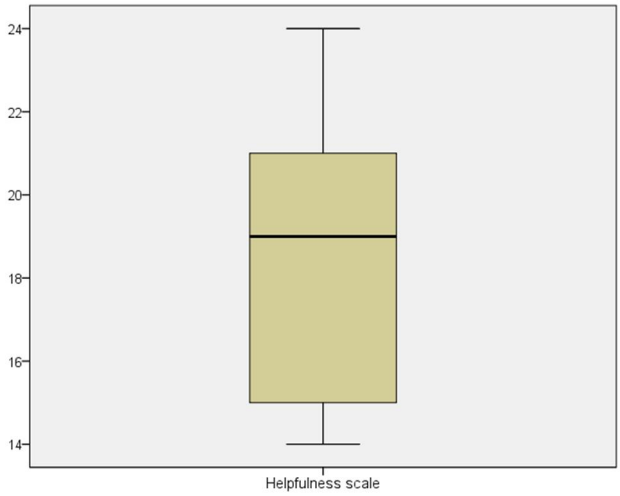

Figure 2 Boxplot of the helpfulness scale

Descriptive statistics for the scale are available in Table 7. The scale could range between 9 (for responding "disagree" to all 9 questions) and 27 (for answering "agree" to all 9 items). Thus, the closer the total to 27, the greater the agreement that imprinted questions helped examinees take the test. The average of the scale (*M* = 18.36, *SD* = 3.272) suggests that, overall, participants considered that imprinted questions were somewhat useful to complete the test. Certain participants wrote comments about the imprinted questions in the post-test questionnaire. Some of their views were related to the helpfulness of the technique ("it was easier to focus on the red-haired boy talk because I was waiting for the answer"; "imprinted question 1 helped me", "it is easier to answer question 4 by seeing the speakers than by listening only. The imprinted question helps to stay focused on the images", "imprinted question 5 helped me"[1]), and others limited the usefulness to low language proficiency students ("I imagine that embedded questions help students beginning to study Spanish a lot. I like the aid, but they did not help me to complete the test"; "they could help low proficiency students") or difficult questions ("they help when dealing with difficult questions"). The results can be compared to two other investigations. First, in the original study (Casañ-Núñez, 2017c) the average of the helpfulness scale was 21.47 out of 27 (*SD* = 2.577), hence, suggesting that test-takers found imprinted questions helpful to complete the test. Second, Torres-Salvador (2019) researched the use of imprinted questions to teach viewing comprehension in the English as second language classroom, and she found out that 78,6% of the learners agreed that this technique was helpful and that "students' motivation levels were considerably incremented by this practice" (p. 42). To sum up, overall, the results of previous studies and the current study support the hypothesis that test-takers of the treatment group would have positive views towards this technique.

Table 7 Descriptive statistics for the helpfulness scale

N	*M*	Median	*SD*	Skewness	SE Skewness	Kurtosis	SE Kurtosis	Min	Max
14	18.36	19.00	3.272	-.063	.597	-1.129	1.154	14	24

4.3 Research question 3

The third research question was *Does the use of synchronized video-imprinted questions have an impact on test-takers' viewing behavior?* First, the variable *percentage of time viewing the videos* was checked for outliers using a boxplot, and none was identified (Figure 3). After that, descriptive statistics were calculated. As can be seen in Table 8, on average, examinees that took the test with

[1] All comments have been translated into English by the author.

the stem of the questions embedded in the video as subtitles looked towards the video a higher percentage of the time (*M* = 70.034, *SD* = 6.248) than test-takers that took the same test without imprinted questions (*M* = 62.052, *SD* = 9.871). As the distribution of the variable was not significantly different from a normal distribution in the control group (*S-W* = .958, *df* = 14, *p* = .693) and in the experimental one (*S-W* = .974, *df* = 12, *p* = .949), further analysis with a *t*-test was appropriate. The *t*-test showed that the difference in the percentage of time looking at the videos was statistically significant: *t*(24) = -2.413, *p* (two-tailed) = .024. Besides, according to Cohen's criteria (Cohen, 1988), the effect was medium to large (*r* = 0.442) and so it represented a fairly substantive finding. The results support the hypothesis that test-takers in the experimental group would watch the videos longer than participants in the control group. This suggests that this technique may be an effective procedure to reduce the visual attention conflict between viewing a video and completing a task at the same time. Other authors have quantified the percentage of time that examinees watch the video in L2 viewing comprehension tests. In Ockey's (2007) study test-takers looked at the video screen 44.9% of the playing time, in Wagner's (2007) study 69%, in Wagner's (2010) study 47.9%, in Suvorov's (2015) study 58% in content videos and 51% in context videos, and in Elmankush's (2017) study 57.4%. Setting aside the differences among the studies, if the viewing rates are compared to one another, the result is favorable because in all previous studies test-takers watched the video screen a lesser percentage of the playing time than the experimental group in the current study (less than 70.034%). This also suggests that imprinted questions may reduce the visual attention conflict of while-viewing activities.

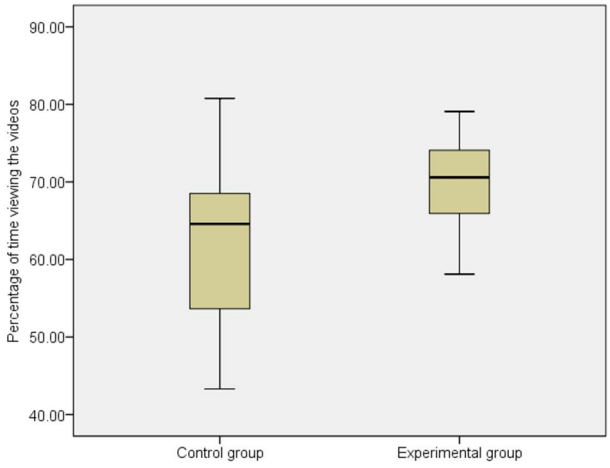

Figure 3 Boxplots of the variable *percentage of time viewing the videos* in the control and experimental groups

Table 8 Descriptive statistics for the variable *percentage of time viewing the videos*

Group	*N*	Min	Max	*M*	*SD*	Skewness	SE Skewness	Kurtosis	SE Kurtosis
Control	14	43.30	80.78	62.0524	9.87084	-.222	.597	-.020	1.154
Experimental	12	58.10	79.09	70.0337	6.24782	-.327	.637	-.193	1.232

5. Conclusion

This approximate replication study (Porte, 2012) of Casañ-Núñez's (2017c) study had two primary purposes. On the one hand, it was intended to confirm or otherwise the results of the original study. On the other hand, it investigated for the first time whether imprinted questions had an impact on L2 students' viewing behavior regarding the video image. Overall, the results of the original and the replication studies are similar. As for test-takers' view of imprinted questions, in the present study, participants consider imprinted questions to be useful to complete the test. This is in line with views expressed in the original study and Torres-Salvador´s (2019) study. Furthermore, this is in line with experts' and Spanish L2 teachers' opinions about this technique (Casañ-Núñez, 2015, 2017b), since both groups regard this proposal as beneficial for the teaching of listening/viewing comprehension. With regard to the effect of imprinted questions on test-takers' performance, neither the original nor the replication study support the hypothesis that the experimental group would score higher than the control group in the viewing comprehension test thanks to the theoretical benefits provided by the imprinted questions. However, on the grounds that these theoretical benefits seem to be well-grounded, that the current study suggests that this technique may reduce the visual attention conflict, and that test-takers consider that this procedure is useful, it is suggested that the hypothesis does not hold in the current and the original studies for a given reason and two possible explanations are discussed. As for the impact of imprinted questions on examinees' viewing behavior, test-takers in the experimental group watch the videos longer than participants in the control group, and the *t*-test shows that the difference is statistically significant. This suggests that imprinted questions may be an effective procedure to reduce the conflict in visual attention between watching a video and completing a written activity simultaneously. This represents a significant finding because studies that have researched students' visual behavior in L2 viewing comprehension tests (e.g., Elmankush, 2017; Suvorov, 2015; Wagner, 2010) have encountered relatively low degrees of attention to the video image, which is negative for both the comprehension of the video and the development of the skill. To sum up, research on imprinted questions suggests that this proposal is potentially beneficial for L2 audiovisual comprehension learning and testing, particularly for non-competent listeners/viewers.

The replication study has various shortcomings. Firstly, as in many empirical studies in the social sciences, participants were selected by non-probabilistic sampling, which limits the extent of generalizability of the findings. Second, as in other studies (Ockey, 2007; Wagner, 2007, 2010), in order to investigate the impact of imprinted questions on test-takers' visual behavior, eye-tracking technology was not employed. Instead, those that participated were video recorded while they took the test with a camera placed on a tripod. When an examinee oriented his/her head towards the video screen while the video was playing, it was estimated that he/she was viewing the video. It must be acknowledged, however, that looking towards the video screen does not necessarily imply paying attention to the video image.

The following directions for future research are suggested. First, it would be beneficial to set up a new replication study with random sampling. Second, in order to study the effect of imprinted questions on L2 learners' visual behavior with regard to the video image, it would be worthwhile to rely on eye-tracking technology rather than on a digital video camera. Third, there was a considerable amount of variation in participants' viewing behavior, especially in the control group. This extensive variation was also found in Ockey's (2007), Wagner's (2007, 2010) and Suvorov's (2015) studies. In order to research the causes of this variation, it would be useful to supplement the visual behavior data in the control and the experimental conditions with verbal reports after the tests. Finally, further research is needed to investigate this technique in other contexts: a classroom learning situation, computer-assisted language learning and computer-assisted language testing.

Acknowledgments

I would like to thank the Universität Rostock students for participating in this study. I also express my gratitude to Dr. Rafael Arnold, Minerva Peinador and Susana Rodríguez for facilitating the study at the Universität Rostock. This research was possible thanks to an Erasmus teaching staff mobility grant at the Universität Rostock.

References

AENOR (2012). *Norma UNE 153010: Subtitulado para personas sordas y personas con discapacidad auditiva*. Madrid, Spain: AENOR.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Battaglia, M. P. (2008). Nonprobability sampling. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 523-526). Thousand Oaks, CA: SAGE Publications, Inc.

Birmingham, H. P., & Taylor, F. V. (2005). A design philosophy for man-machine control systems. In N. Moray (Ed.), *Ergonomics: Skill, displays, controls, and mental workload*, Volume II (pp. 360-382). London, England: Taylor & Francis. (Original work published 1954)

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality*, *54*(1), 106-148.

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

Casañ-Núñez, J. C. (2015). Un marco teórico sobre el uso de preguntas de comprensión audiovisual integradas en el vídeo como subtítulos: un estudio mixto. *MarcoELE*, *20*, 1-45. http://marcoele.com/comprension-audiovisual-y-subtitulos/

Casañ-Núñez, J. C. (2016). Desarrollo de una prueba de comprensión audiovisual. *MarcoELE*, *22*, 1-70. http://marcoele.com/descargas/22/casan-prueba_audiovisual.pdf

Casañ-Núñez, J. C. (2017a). Diseño y fiabilidad de un cuestionario sobre la comprensión auditiva/audiovisual. *Bellaterra Journal of Teaching & Learning Language & Literature*, *10*(3), 47-65 https://doi.org/10.5565/rev/jtl3.686

Casañ-Núñez, J. C. (2017b). Tareas de comprensión audiovisual con preguntas subtituladas: valoraciones de cinco profesores universitarios de espanol como lengua extranjera. *E-JournALL*, *EuroAmerican Journal of Applied Linguistics and Languages*, *4*(1), 20-39. http://dx.doi.org/10.21283/2376905X.6.77

Casañ-Núñez, J. C. (2017c). Testing audiovisual comprehension tasks with questions embedded in videos as subtitles: A pilot multimethod study. *The EUROCALL Review*, *25*(1), 1-25. https://doi.org/10.4995/eurocall.2017.7062

Casañ-Núñez, J. C. (2018). Preguntas de comprensión audiovisual sobreimpresas: propuesta para el aprendizaje de segundas lenguas. *Revista Latinoamericana de Tecnología Educativa*, *17*(1), 105-120. https://doi.org/10.17398/1695-288X.17.1.105

Chion, M. (1994). *Audio-vision: Sound on screen* (C. Gorbman, Trans.). New York, NY: Columbia University Press. (Original work published 1990).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, L., Manion, L., & Morrison, K. (2011): *Research methods in education* (7th ed.). New York, NY: Routledge.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98-104. https://doi.org/10.1037/0021-9010.78.1.98

Council of Europe (2018). *Common European framework of reference for languages: Learning, teaching, assessment. Companion volume with new descriptors*. Strasbourg, France: Council of Europe Publishing. https://www.coe.int/en/web/common-european-framework-reference-languages

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Thousand Oaks, CA: Sage Publications.

Díaz Cintas, J. (2013). Subtitling: Theory, practice and research. In C. Millán & F. Bartrina (Eds.), *The routledge handbook of translation studies* (pp. 273-287). London, London: Routledge.

Díaz Cintas, J., & Remael, A. (2007). *Audiovisual translation: Subtitling*. Manchester, London: St. Jerome Publishing.

Elmankush, I. (2017). Investigating the impact of including videos or still images in computer-based academic listening comprehension tests (PhD thesis). University of York. http://etheses.whiterose.ac.uk/id/eprint/18811

Field, A. (2013). Discovering statistics using SPSS (4th ed.). London, England: Sage Publications.

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

Harris, T. (2003). Listening with your eyes: The importance of speech related gestures in the language classroom. *Foreign Language Annals*, *36*(2), 180-187. https://doi.org/10.1111/j.1944-9720.2003.tb01468.x

International Listening Association (1996). Home page. http://listen.org/

Jiang, D., Kalyuga, S., & Sweller, J. (2017). The curious case of improving foreign language listening skills by reading rather than listening: An expertise reversal effect. *Educational Psychology Review*. https://doi.org/10.1007/s10648-017-9427-1

Lynch, T. (2012). Traditional and modern skills. Introduction. In M. Eisenmann & T. Summer (Eds.), *Basic issues in EFL teaching and learning* (pp. 69-81). Heidelberg, Germany: Winter.

Mayer, R. (2014). *The Cambridge handbook of multimedia learning* (2nd ed.). Cambridge: Cambridge University Press.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.

Morse, J. M. (2003). Principles of mixed methods and multimethod research design. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 189-208). Thousand Oaks, CA: Sage.

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, *24*(4), 517-537. https://doi.org/10.1177/0265532207080771

Pallant, J. (2013). *SPSS survival manual: A step by step guide to data analysis using SPSS* (5th ed.). Maidenhead, Australia: Open University Press/McGraw-Hill.

Porte, G. (2012). Introduction. In G. Porte (Ed.), *Replication research in applied linguistics* (pp. 1-17). Cambridge: Cambridge University Press.

Richards, J. C., & Burns, A. (2012). *Tips for teaching listening*. New York, NY: Pearson Education.

Rost, M. (2016). *Teaching and researching listening* (3rd ed.). New York, NY: Routledge.

Saris, W. E., & Gallhofer, I. N. (Eds.). (2014). *Design, evaluation, and analysis of questionnaires for survey research* (2nd ed.). Hoboken, NJ: Wiley & Sons https://doi.org/10.1002/9781118634646

Sidaty, N., Larabi, M.-C., & Saadane, A. (2017). Toward an audiovisual attention model for multimodal video content. *Neurocomputing*, *259*, 94-111. https://doi.org/10.1016/j.neucom.2016.08.130

Suhr, D., & Shay, M. (2009). Guidelines for reliability, confirmatory and exploratory factor analysis. In conference proceedings of the *Western Users of SAS Software* (pp. 1-15). San Jose, CA: www.lexjansen.com/wuss/2009/anl/ANL-SuhrShay.pdf

Suvorov, R. (2015). Interacting with visuals in L2 listening tests: An eye-tracking study. In V. Berry (Ed.), *ARAGs research reports online* (Report #AR-A/2015/1). http://www.britishcouncil.org/sites/default/files/interacting_with_visuals_in_l2_listening_tests_suvorov.pdf

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. London: Springer.

Torres-Salvador, M. (2019). *Imprinted questions: An innovative approach on audiovisual materials in the EFL classroom* (Unpublished MA thesis). University of Valencia, Spain.

Underwood, M. (1989). *Teaching listening*. London: Longman.

Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening*. New York, NY: Routledge.

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning and Technology, 11*(1), 67-86. http://llt.msu.edu/vol11num1/pdf/wagner.pdf

Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly*, *5*(3), 218-243. https://doi.org/10.1080/15434300802213015

Wagner, E. (2010). Test-takers' interaction with an L2 video listening test. *System*, *38*(2), 280-291. https://doi.org/10.1016/j.system.2010.01.003

Worthington, D. L., & Bodie, G. D. (2018). Defining listening: A historical, theoretical and pragmatic assessment. In D. L. Worthington & G. D. Bodie (Eds.), *The sourcebook of listening research: Methodology and measures* (pp. 3-17). Hoboken, NJ: John Wiley & Sons.