

Research paper

Testing audiovisual comprehension tasks with questions embedded in videos as subtitles: a pilot multimethod study

Juan Carlos Casañ Núñez
Universitat Politècnica de València, Spain

juancarloscasan@protonmail.com

Abstract

Listening, watching, reading and writing simultaneously in a foreign language is very complex. This paper is part of wider research which explores the use of audiovisual comprehension questions imprinted in the video image in the form of subtitles and synchronized with the relevant fragments for the purpose of language learning and testing. Compared to viewings where the comprehension activity is available only on paper, this innovative methodology may provide some benefits. Among them, it could reduce the conflict in visual attention between watching the video and completing the task, by spatially and temporally approximating the questions and the relevant fragments. The technique is seen as especially beneficial for students with a low proficiency language level.

The main objectives of this study were to investigate if embedded questions had an impact on SFL students' audiovisual comprehension test performance and to find out what examinees thought about them. A multimethod design (Morse, 2003) involving the sequential collection of three quantitative datasets was employed. A total of 41 learners of Spanish as a foreign language (SFL) participated in the study (22 in the control group and 19 in the experimental one). Informants were selected by non-probabilistic sampling. The results showed that imprinted questions did not have any effect on test performance. Test-takers' attitudes towards this methodology were positive. Globally, students in the experimental group agreed that the embedded questions helped them to complete the tasks. Furthermore, most of them were in favour of having the questions imprinted in the video in the audiovisual comprehension test of the final exam. These opinions are in line with those obtained in previous studies that looked into experts', SFL students' and SFL teachers' views about this methodology (Casañ Núñez, 2015a, 2016a, in press-b). On the whole, these studies suggest that this technique has potential benefits for FL learning and testing. Finally, the limitations of the study are discussed and some directions for future research are proposed.

Keywords: Audiovisual comprehension, listening comprehension, multimethod design, Spanish as a foreign language, subtitles, video listening test.

1. Background

Contrary to speaking and writing, which have observable products, listening comprehension occurs in an internal way, invisible to the eyes of the speaker. Owing to that, it is complex to study its nature and to arrive at a definitive description. In this paper, listening is understood as a process of interpretation of auditory and visual information, as suggested by specialists such as Lynch (2012) and Martín Peris (1991/2007). Rubin (1995b, p. 7) proposes the following definition: "an active process

in which listeners select and interpret information which comes from auditory and visual cues in order to define what is going on and what the speakers are trying to express". Thus, it is considered that factors such as "proxemics, kinesics and deictics are all part of the message. They are not just a sort of gloss on the verbal component" (Riley, 1979, p. 84). Harris (2003), Lynch (2012) and Riley (1979) have suggested that the term *listening* does not reflect the multimodal nature of most listening comprehension situations. From now on, in order to show the dual dimension of this communicative activity, the compound *listening/audiovisual comprehension* will be employed.

According to a number of authors (Lynch, 2009; Mendelsohn, 1994; Rubin, 1995a, Ur, 1999), video materials should prevail over audio recordings to practice listening/audiovisual comprehension. To begin with, it is consistent with a definition of the skill as a process of interpretation of auditory and visual information. Besides, it allows the learner to observe the reality of most speaking interactions. In addition to that, video has a positive effect on motivation (Flowerdew & Miller, 2005; Ur, 1994, 1999; Vandergrift & Goh, 2012). Finally, there are arguments in favour of multimodal learning. According to the cognitive theory of multimedia learning (CTML), multimedia "takes advantage of the full capacity of humans for processing information. When we present material only in the verbal mode, we are ignoring the potential contribution of our capacity to also process material in the visual mode" (Mayer, 2014, p. 6). Thus, as stated by the multimedia principle of the CTML "students learn better from words and pictures than from words alone" (Mayer, 2001, p. 63). This theory was not developed specifically for learning foreign languages, however, some principles are applicable to this field (Plass & Jones, 2005). In relation to the multimedia principle, these authors point out that "it is the combination of both visual and verbal presentations of information that has most strongly and consistently supported listening and reading comprehension and vocabulary acquisition" (p. 479). Video materials should also prevail over audio recordings for testing listening/audiovisual comprehension. First, it is congruent with the double nature of this communicative activity. Second, it is in harmony with common practice in the classroom (Buck, 2001; Gruba, 1997; Pardo-Ballester, 2016). Third, it increases the validity of the test (Bejar, Douglas, Jamieson, Nissan, & Turner, 2000; Wagner, 2007, 2008, 2010a), its authenticity (Alderson, 2005; Bejar et al. 2000; Ockey, 2007; Wagner, 2007, 2008) and its naturalness (Alderson, 2005). Fourth, "seeing the situation and the participants tends to call up relevant schemes" (Buck, 2001, p. 172). Lastly, not using video in language testing may have a negative backwash effect. If the skill is tested only using audio recordings, then there is a pressure to practice this communicative activity mainly with this sort of materials.

When designing listening/audiovisual comprehension tasks, it is essential to take into account that "it is extremely difficult to listen and write at the same time, particularly in a foreign language" (Underwood, 1989, p. 48), and that listening, viewing, reading and writing at the same time can be even more difficult. On the one hand, there is a conflict of visual attention between viewing a video and completing a written activity at the same time. On the other hand, it should be borne in mind that "there is universal agreement that working memory when dealing with novel information is very limited in capacity" (Sweller, Ayres & Kalyuga, 2011, p. 42). As Vandergrift and Goh (2012) highlight, paying attention to the video and the task simultaneously may cause working memory overload. This complexity helps to explain the relatively low degree of attention paid to the video in studies that researched test-takers' viewing rates during foreign language (FL) video-based listening comprehension tests. Ockey (2007) does not supply the average watching rate but, from the data, it can be calculated that it is 44.9%. Wagner (2007) found that examinees made eye contact with the screen 69% of the time, while in a later study (Wagner, 2010b), this figure decreased to 47.9%. These authors recorded the participants while taking the tests and they measured the amount of time informants look towards the monitor. Suvorov (2015) uses eye-tracking technology to investigate how examinees *interact* with two types of videos: context videos and content videos. He discovers that test-takers spend 58% of the time

watching content videos and 51% context videos. These low degrees of attention to the image may have a negative influence both on understanding the video and on the development of the communicative activity. The complexity of the while-viewing phase also contributes to explaining the mixed conclusions in (a) studies that compared the results obtained by a group after conducting a video-based FL test with those achieved by another group that took an audio-only version of the same test, and (b) research into test-taker's attitudes towards the image. Some authors do not find significant differences in performance between taking a test with video and completing the same test with an audio-only version of the audiovisual text (Batty, 2014; Coniam, 2001; Gruba, 1993; Londe, 2009), other authors discover that test-takers achieve higher scores with the video based test (Sueyoshi & Hardison, 2005; Wagner, 2010a, 2013), and Suvorov (2008) finds evidence that results are higher with the audio-only version. Similarly, in some studies, examinees have a positive attitude towards the video image (Sueyoshi & Hardison, 2005; Wagner, 2010b), and in others they have a negative attitude toward the video (Alderson, Clapham & Wall, 1995; Coniam, 2001; Suvorov, 2008). Of course, other elements may have played a role in these mixed results. These include the focus of listening/audiovisual questions, the complexity of the task and the text, the way in which the viewing was carried out, the quality of the input, the stress generated by the tests, the greater or lesser degree of solidarity between visual and verbal information, the type of visuals (content or context visuals), the influence of the video cameras on viewing behaviour, and so on.

In order to keep while-viewing work manageable, it is advisable that written tasks involve scarce reading and writing, and that they require few active elements to be stored in working memory. It is also recommendable to use the technique of paused listening/viewing (see Field, 2008; Stempleski & Tomalin, 2001; Stoller, 1992). Thirdly, it is useful to spatially approximate the video and the activity to reduce the time needed to shift from one stimulus to another (from the task to the video and vice versa). Lastly, as proposed in this study, it may be more beneficial for the learner to see the comprehension questions embedded in the video in the form of subtitles and synchronized with the relevant fragments (Casañ Núñez, 2015a). Basically, comprehension questions appear on screen a few seconds before the beginning of the fragment to which they are related, they remain visible for the duration of the relevant snippet and they disappear when the relevant part of the video finishes. Compared to viewings where the activity is available only on paper, this technique could minimize conflict in visual attention by spatially and temporally approximating the questions and the pertinent scenes, and it could reduce the cognitive strain of the task, since students would only need to pay attention to one subtitled question at a time, instead of different printed questions on paper. According to Field (2008), unskilled listeners may reach working memory overload faster than competent listeners because the former have poorly automatized decoding processes and they spend a great deal of working memory on decoding. Consequently, the procedure is seen as especially beneficial for students with a low proficiency language level. In addition, it can be used occasionally in higher proficiency levels for two main reasons: firstly, it helps learners keep focused on what they are watching compared to viewings with questions on paper; secondly, the study reported in Casañ Núñez (2016a) shows that learners with a high listening/audiovisual comprehension level in SFL (approximately B2+/C1) have positive views on this technique.

This procedure may be employed in different contexts. It is suitable for paper-and-pencil listening/audiovisual comprehension tasks for language learning or testing (see Figure 1) and for CALL or CALT (see Figures 2 and 3). As for learning and testing activities, Alderson et al. (1995, p. 42) state that "the main difference between a test and an exercise is that with exercises learners get support: with tests, they do not". A video demonstration of this technique is available from <https://youtu.be/ALw8XJkrbDQ> (01/02/2017).



Figure 1. Example of an audiovisual comprehension question embedded in the video in the form of a subtitle. "¿De qué temas habla el chico pelirrojo?" [What is the redheaded boy talking about?] From the Spanish film *Los peores años de nuestra vida* by Emilio Martínez Lázaro.

Nombre

Escucha y observa el vídeo. En las preguntas de elección múltiple solo hay una respuesta correcta.

1. ¿Cómo se llama la chica?

- a) Vanesa
- b) Violeta
- c) Verónica
- d) Virginia

2. ¿De qué hablan?

- a) la familia y los estudios de la chica
- b) la familia y el trabajo de la chica
- c) la casa y el trabajo de la chica
- d) la casa y los estudios de la chica

3. ¿Cuál es el teléfono de la chica?


Puedes tomar notas aquí

4. Considerando todo el fragmento, ¿qué sentimientos puede tener el chico por la chica? Justifica brevemente la respuesta.

Figure 2. Prototype of an audiovisual comprehension task for CALL. Notice that playback controls are available to learners. "¿Cómo se llama la chica?" [What is the girl's name?]. From the Spanish film *Ópera prima* by Fernando Trueba.

Nombre

Escucha y observa el vídeo. En las preguntas de elección múltiple solo hay una respuesta correcta.



1. ¿Cómo se llama la chica?

a) Vanesa

b) Violeta

c) Verónica

d) Virginia

2. ¿De qué hablan?

a) la familia y los estudios de la chica

b) la familia y el trabajo de la chica

c) la casa y el trabajo de la chica

d) la casa y los estudios de la chica

3. ¿Cuál es el teléfono de la chica?

4. Considerando todo el fragmento, ¿qué sentimientos puede tener el chico por la chica? Justifica brevemente la respuesta.

Puedes tomar notas aquí

Figure 3. Prototype of an audiovisual comprehension task for CALT. Notice that only the play control is available to test-takers. From the Spanish film *Ópera prima* by Fernando Trueba.

Previously, a theoretical framework describing the use of this technique and its potential benefits and limitations has been proposed (Casañ Núñez, 2015a). The framework was developed out of a literature review, the teaching experience with this procedure and the comments of a group of experts in teaching Spanish as a Foreign Language (SFL), Spanish linguistics and/or the use of technology. Also, small studies have been carried out to investigate what SFL university teachers and university students with a high listening/audiovisual comprehension level in SFL (approximately B2 +/C1) think about this technique (Casañ Núñez, 2016a, in press-b). The results suggest that, overall, teachers and learners have positive views about this methodology. In addition to experts', teachers' and students' views, it is fundamental to find out what effect this technique has on learners' audiovisual comprehension and viewing behaviour.

The main purposes of the current pilot study were to investigate if the technique had an impact on SFL students' audiovisual comprehension test performance and to find out what examinees thought about imprinted questions. Moreover, it explored some learner preferences regarding listening/audiovisual comprehension. The study used datasets from previous research that described the development of a listening/audiovisual test (Casañ Núñez, 2016b, pp. 36-51). The study reported in this paper, however, had different objectives; it took into account data that was not inspected and it analysed the data addressing the following research questions:

Research question 1: How important is it for learners to practise listening/audiovisual comprehension in the classroom? What type of recordings (audio or video) do students prefer for practising listening/audiovisual comprehension in the classroom? Do learners think that the visual input helps them to understand what speakers are saying? How do students practise listening/audiovisual comprehension outside the classroom?

Research question 2: Does the use of questions embedded within the video in the form of subtitles and synchronized with the relevant fragments in a FL audiovisual comprehension test facilitate test-takers' performance? Do the test-takers of the experimental group score higher or lower than the test-takers of the control group?

It was hypothesized that examinees that take the audiovisual comprehension test with questions embedded in the video, that is, students in the experimental group, would outperform examinees that took the same test without them, i.e., those in the control group. As described in the introduction, imprinted questions could minimize the conflict in visual attention between watching the video and completing the task, by spatially and temporally approximating the questions and the relevant fragments. Moreover, this technique could diminish the cognitive strain of the activity, because learners would only need to pay attention to one request each time, instead of several printed questions. Nonetheless, a null hypothesis of no difference was tested.

Research question 3: What are test-takers' attitudes towards the use of questions embedded in the video as subtitles?

- Research question 3.1: To what extent do test-takers in the experimental group agree or disagree that embedded questions helped them to complete the tasks?
- Research question 3.2: Is there any subtitled question consistently considered more or less helpful than the others by the experimental group?
- Research question 3.3: Are students in the experimental group in favour of having the questions embedded in the video (in addition to having them on paper) in the audiovisual comprehension test of the final exam?

As mentioned in the introduction, the use of subtitled questions has potential benefits. In addition, a previous study (Casañ Núñez, 2016a) showed that SFL learners had positive attitudes towards this methodology. Therefore, in this study, it was hypothesized that students in the experimental group would agree that the questions embedded in the form of subtitles aided them, and that they would be in favour of having them in the test of the final exam.

2. Method

2.1. Study design

A multimethod design (Morse, 2003) was employed. It involved the sequential collection of three quantitative datasets that were used basically to answer different subquestions. First, participants were surveyed with the purpose of getting to know the sample and some of their preferences regarding listening/audiovisual comprehension. Second, an audiovisual comprehension test with two variants was administered to find out if there were differences in performance between test-takers of the control and experimental groups. Third, attitudinal data towards the use of questions embedded in the form of subtitles from the experimental group was collected through a questionnaire.

2.2. Participants

SFL students were selected by a convenience, non-probabilistic sampling method (Dörnyei, 2007, pp. 98-99). The instruments were administered in different lessons. Some students did not complete the first questionnaire because they missed the lessons where they were handed out (see Table 1).

Table 1. Number of informants that completed each instrument.

Instruments	Groups	
	Control	Experimental
First questionnaire	18	18
Audiovisual comprehension test	22	19
Second questionnaire (only for the experimental group)	n/a	19

All participants were enrolled in *Spanish II*, a foreign language course delivered at the Universidade de Coimbra (Portugal). *Spanish II* had three shifts. Two of them had fewer persons registered. Thus, it was decided that the smaller groups completed the same version of the test. Randomly, the less numerous shifts were designated control groups and the remaining one, the treatment group. All participants had the same Spanish teacher and studied Human or Social Sciences degrees. All but two informants were between eighteen and twenty-four years old. All were Lusophones except for one participant in the experimental group who was of Ukrainian origin. Roughly speaking, students had been studying Spanish for a similar amount of time and they had spent an analogous amount of time in Spanish-speaking countries. 83.3% of the participants of the control group and 88.9% of the informants of the experimental group reported that they were very interested or very much interested in cinema. The others answered "neutral". That aspect was relevant because the videotexts employed in the audiovisual comprehension test were film scenes and a low interest in cinema could have some negative influence on performance. A pre-test to determine the informants' level of audiovisual comprehension in Spanish was not administered. From the observation of the texts and tasks employed for practising listening/audiovisual comprehension in the classroom, it was estimated that most students had a B1+ level in this skill (according to the *Common European Framework for Languages*). As described in Casañ Núñez (2016b, pp. 39-40), in order to estimate the degree of equivalence between the groups regarding audiovisual comprehension, participants completed slightly modified versions of the first and second tasks from a B1 level listening test designed by Hidalgo de la Torre (Coord., 2013: pp. 52-53). Both groups obtained similar high scores. The average mark in the control group was 9.566 out of 12 ($SD = 1.861$) and in the experimental one, it was 9.684 out of 12 ($SD = 1.827$). That suggested that there were few differences between the groups and that the B1 level listening tasks were easy for test-takers. The Mann-Whitney test for two independent samples confirmed that there were no statistically significant differences between the scores of the groups ($U = 146.5$, $z = -.185$, $p = .853$, $r = -.031$).

2.3. Instruments

To collect the data, three instruments were developed: two questionnaires and an audiovisual comprehension test with two versions. The first questionnaire aimed at gathering specific information about the sample group: sociodemographic aspects, what importance was attributed to the exercise of listening/audiovisual comprehension in the classroom to learn Spanish, some learning preferences about listening/audiovisual comprehension, to what extent video images were considered helpful or unhelpful to understand the interlocutors, and how listening/audiovisual comprehension was practiced outside the classroom. It included twenty-six items and, following the classification of Saris and Gallhofer (2014), it employed open requests and closed categorical requests. The development of the instrument involved expert review, a pilot, and a study where the repeated-surveys method (Brown, 2001, pp. 171-172) was used to calculate its reliability. The questionnaire and a detailed account of its elaboration can be found in Casañ Núñez (in press-a).

The audiovisual comprehension test had two variants: a traditional one for the control group and an experimental one for the treatment group. An extensive account of the planning, design and trialling of both versions of the test can be found in Casañ Núñez (2016b). In both instances, tasks were available on paper. Besides, in the experimental one, questions were embedded in the video in the form of subtitles and were synchronized with the relevant fragments (see Figures 4 and 5).



Figure 4. Screenshots of questions 6 and 7. The photographs belong to *Los peores años de nuestra vida* by Emilio Martínez Lázaro.

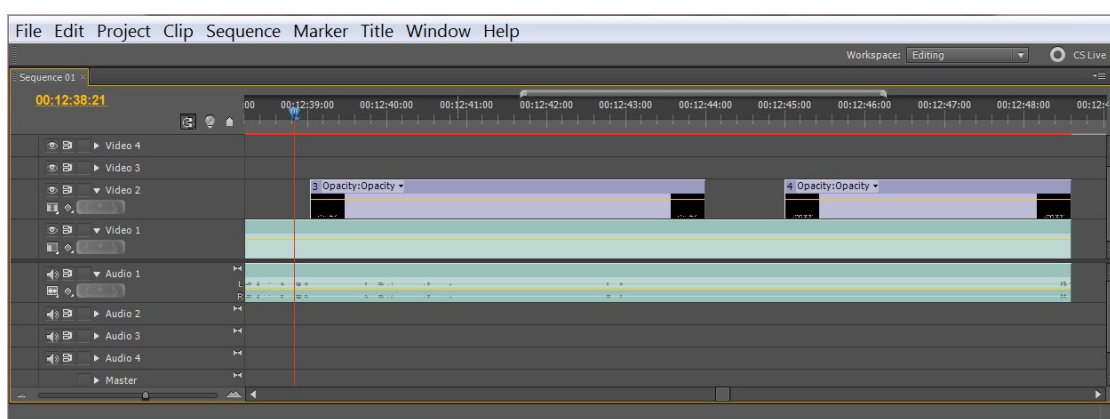


Figure 5. Timeline in Adobe Premiere Pro. Video 1 and Audio 1 tracks correspond to the film. Video 2 track shows the timing of questions 6 and 7.

The target language use domain was the comprehension of informal conversations pertaining to the personal domain in Spanish romantic comedies. The construct measured three skills: extracting specific information, identifying general ideas and recognizing feelings in face-to-face informal conversations. The test was composed of two tasks, two texts and seven items. The main features of the texts and the experimental items can be seen in Tables 2 and 3. To indicate the film scene, an eight-digit time code is used. The first three pairs of digits correspond to hours, minutes and seconds, respectively, and the last pair, to frames. Thus, 00:02:36:12 designates the following point: 2 minutes, 36 seconds and 12 frames of *Ópera prima*.

Table 2. Main characteristics of the input texts.

1. Text source	Videotext (source: <i>CEFR</i> a p. 49) Scene from the comedy <i>Ópera prima</i> 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 <i>Los peores años de nuestra vida</i> 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: <i>CEFR*</i> 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

* Common European Framework of Reference for Languages: Learning, Teaching and Assessment.

Table 3. Description of the embedded questions in the scenes from *Ópera prima* (1-4) and *Los peores años de nuestra vida* (4-7).

Subtitle	Focus	Question type	Timing	No. of lines	Font and size	Colour	Estimated level
1. ¿Cómo se llama la chica?	ESI	Multiple-choice	00:01:10:00 00:01:37:09	1	Arial Narrow 36	White (hex colour code #FFFFFF)	A1
2. ¿De qué hablan?	IGI	Multiple-choice	00:01:44:14 00:02:02:02	1			A2/B1
3. ¿Cuál es el teléfono de la chica?	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?	RF	Open	00:02:27:08 00:02:36:12	2			A2/B1
5. ¿De qué temas habla el chico pelirrojo?	IGI	Multiple-choice	00:11:40:00 00:12:25:24	1			B2
6. ¿Cómo se llama la chica?	ESI	Multiple-choice	00:12:39:04 00:12:43:24	1			A1
7. ¿Qué sentimientos puede tener el chico pelirrojo por la chica?	RF	Open	00:12:45:00 00:12:48:15	2			B2

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

The second questionnaire aimed at answering the third group of research questions. It was a development of a survey used in a previous exploratory study (Casañ Núñez, 2016a). A questionnaire design was chosen because they are "uniquely capable of gathering large amounts of information quickly" (Dörnyei, 2007, p. 101). The instrument had three parts: title, introductory text and items (see appendix). The title tried to be as informative as possible. The introductory text specified the objective of the questionnaire, it mentioned that there were no correct or incorrect answers, and it stated that the data collected would be treated confidentially and would only be used for academic purposes. Following the classification of Saris and Gallhofer (2014), the instrument was made up of two types of items: closed categorical requests and open

requests. The former ones were of two subtypes: requests with nominal response categories and requests with ordinal response categories.

As recommended by Saris and Gallhofer (2014), the response options of the closed categorical requests attempted to offer a range of options, broad enough to encompass the possible answers of participants, and they were mutually exclusive. Closed categorical requests with ordinal response categories consisted of Likert-type items (1, 2.1, 2.2, 2.3, 2.4, 5, 6.1, 6.2 and 6.3). They made up a scale that attempted to answer research question 3.1 *To what extent do test-takers in the experimental group agree or disagree that embedded questions helped them to complete the tasks?* The scale was named *helpfulness scale*. Additionally, individual items tried to respond to research question 3.2 *Is there any subtitled question ranked consistently higher or lower than the others by the experimental group?* Test-takers expressed their opinion about the tasks of the test in a global way (items 1 and 5) and about each subtitled question individually (items 2.1, 2.2, 2.3, 2.4, 6.1, 6.2 and 6.3). There are discrepancies on the number of anchors of Likert scales: Likert (1932) employs mainly five alternatives; Jacoby and Matell (1971) suggest that three categories are sufficient; Allen and Seaman (2007) advise a minimum of five options, and Bisquerra and Pérez-Escoda (2015) recommend eleven. Three anchors were opted for, primarily for two reasons. First, it was enough to find out if the attitude towards subtitled questions was positive, negative or neutral. Second, the size of the sample was small and it would be necessary to collapse the categories. The alternatives chosen were: agree, neither agree nor disagree and disagree. In that selection there was a balance between positive and negative alternatives and a neutral central point could be identified. In addition, terms that indicated total agreement or disagreement were avoided because, according to Cohen, Manion and Morrison (2011), people often prefer to skip them so as not to appear extremist. Closed categorical requests with nominal response categories (items 3 and 7) sought to answer research question 3.3 *Are students in the experimental group in favour of having the questions embedded in the video (in addition to having them on paper) in the audiovisual comprehension test of the final exam?*

“By permitting greater freedom of expression, open-format items can provide a far greater richness than fully quantitative data. The open responses... can also lead us to identify issues not previously anticipated” (Dörnyei, 2007, p. 107). Their main drawback lies in the complexity of analysing them (Cohen et al., 2011). Two items of this nature were included so that participants could express their opinion freely (items 4 and 8). In order for test-takers’ production level in Spanish not to restrict responses, the possibility of writing both in Spanish and Portuguese was offered.

2.4. Procedures

All instruments were administered by the researcher during *Spanish II* lessons. The fact that the author was also the teacher may have favoured the fact that participants took the questionnaires and the test seriously (an essential requirement for the results to be valid). The first questionnaire was administered for the first time in February 2014 to students from the three shifts. Precise instructions were provided and students were told that there were no correct or incorrect replies. No time limit was imposed and questions arising from any doubts were answered. Informants took up to 8 minutes to complete the questionnaire and none of them requested help.

The test and the second questionnaire were administered in the same classroom in March 2014. Treatment groups completed the experimental version of the test and the questionnaire, whereas the control group took the traditional variant of the test. Test-takers were informed that they could take as long as they needed to answer. The slowest test-taker of the control group employed 12 minutes and 27 seconds to finish the test. The slowest examinees of the two experimental groups needed 20 minutes and 14 seconds, and 19 minutes and 42 seconds to respond to the test and the survey. A detailed account of the procedures followed can be found in Casañ Núñez (2016b, pp. 42-43).

2.5. Data analyses

All quantitative statistical analyses were carried out using SPSS version 21, except for effect sizes. For that purpose, the Windows scientific calculator was used. Data was double-checked for accuracy to avoid input errors. Furthermore, a frequency analysis of all variables was carried out to verify that there were no missing or anomalous values. Table 4 sums up the quantitative data analyses that were followed to answer each research question.

As for the open-ended items in the second questionnaire (4 and 8), only 5 out of 19 participants in the experimental group made comments, and those observations were short (one sentence). This might have to do with Dörnyei's cautionary advice (2007, p. 105): questionnaires "are unlikely to yield the kind of rich and sensitive description of events and participant perspectives that qualitative interpretations are grounded in". As the amount of qualitative data was so small, the only analysis consisted in quantizing related responses.

Table 4. Quantitative data analyses used for answering research questions

Research question	Research instrument	Analyses
RQ1	First questionnaire (instrument developed by Casañ Núñez, in press-a)	Frequencies
RQ2	Audiovisual comprehension test (instrument developed by Casañ Núñez, 2016b)	Scores of each version of the test: descriptive statistics, analyses of facility, discrimination and reliability, and correlations with two tasks from a listening exam (see Casañ Núñez, 2016b, pp. 44-51) Tests scores: Shapiro–Wilk test Comparison of tests scores: Mann–Whitney test
RQ3	Second questionnaire (based on a survey used by Casañ Núñez, 2016a)	RQ 3.1: Reliability analysis of the helpfulness scale (Cronbach's <i>alpha</i> , corrected item-total correlation, Cronbach's alpha if item deleted) and descriptive statistics of the scale RQ 3.2: Friedman test (items 2.1, 2.2, 2.3, 2.4, 6.1, 6.2 and 6.3) and two Wilcoxon Signed Rank tests using a Bonferroni adjusted alpha value (item 2.4 with item 2.2, item 2.4 with item 2.3). RQ 3.3: frequencies (items 3 and 7)

3. Results and discussion

3.1. Research question 1

The first research question addressed learners' thoughts and preferences regarding listening/audiovisual comprehension. Informants' answers can be found in Tables 5 to 8.

Table 5. How important is it for learners to practise listening/audiovisual comprehension in the classroom?

		Frequency	Percent	Valid Percent
Valid	Very*	16	39.0	44.4
	Very much	20	48.8	55.6
	Total	36	87.8	100.0
Missing**	System	5	12.2	
Total		41	100.0	

*Participants could choose one of five responses: (a) very little, (b) a little, (c) neutral, (d) very, and (e) very much.

**Missing values correspond to informants that did not complete the first questionnaire.

All informants considered that practising listening/audiovisual comprehension in the classroom was either “very” important (16 / 44.4%) or “very much” important (20 / 55.6%). These results are congruent with the weight of listening/comprehension in communication (“we listen to twice as much language as we speak, four times as much as we read, and five times as much as we write”, Celce-Murcia & Olshtain, 2000, p. 102) and in acquisition (“in order for acquirers to progress to the next stage in the acquisition of the target language, they need to understand input language that includes a structure that is part of the next stage”, Krashen & Terrel, 1983, p. 32).

Table 6. What type of recordings (audio or video) students prefer for practising listening/audiovisual comprehension in the classroom?

		Frequency	Percent	Valid Percent
Valid	Audio*	4	9.8	11.1
	Video	11	26.8	30.6
	Both equally	21	51.2	58.3
	Total	36	87.8	100.0
Missing**	System	5	12.2	
Total		41	100.0	

*Participants could choose one of three options: (a) audio, (b) video, and (c) both equally.

**Missing values correspond to informants that did not complete the first questionnaire.

Most students (21 / 58.3%) reported that they did not have a preference for audio or video materials for practising listening/audiovisual comprehension in the classroom. In addition, many more participants showed a predilection for video (11 / 30.6%) instead of audio (4 / 11.1%). As participants were not asked to justify their answer, it is not possible to know why they made their choices. Possibly, the preference for video over audio could be due to the positive effect of video on motivation (Flowerdew & Miller, 2005; Ur, 1994, 1999; Vandergrift & Goh, 2012). Despite this, from the point of view of the learners’ preferences, the results provide support for the idea, previously defended in the introduction, that video should prevail over audio to teach listening/audiovisual comprehension, or that at least both types of materials should have a similar weight.

Table 7. Do learners think that the visual input helps them to understand what speakers are saying?

		Frequency	Percent	Valid Percent
Valid	A little*	1	2.4	2.8
	Neutral	9	22.0	25.0
	Very	20	48.8	55.6
	Very much	6	14.6	16.7
	Total	36	87.8	100.0
Missing**	System	5	12.2	
Total		41	100.0	

*Participants could choose one of five responses: (a) very little, (b) a little, (c) neutral, (d) very, and (e) very much.

**Missing values correspond to informants that did not complete the first questionnaire.

Most informants (26 / 72.3%) responded that the visual input was “very” or “very much” useful to understand the speakers. One participant commented on item 4 of the second questionnaire the following: “Pienso que la imagen es una ayuda para la comprensión” (“I think the image aids comprehension”, author’s translation). The results are in line with some of the findings in Sueyoshi and Hardison (2005, p. 682) and Wagner (2010b, p. 287). Sueyoshi and Hardison asked 42 English as a second language (ESL) students if they agreed with three different statements: “it is easier to understand English when I can see the speaker’s face”, “it is easier to understand English when I can see the speaker’s gestures” and “it is easier to understand English conversations on TV than on the radio” (p. 697). Participants’ answers revealed agreement with the three utterances. 14 of the 42 ESL learners were exposed to a video where it was possible to see the speakers’ gestures and face. Afterwards, they were asked if they agreed that watching the speaker’s gestures, on the one hand, and watching the speaker’s face, on the other hand, helped their understanding. Globally, students sympathized with both ideas. Wagner (2010b) administered a post-test questionnaire to 56 ESL test-takers to explore their attitudes towards the use of video. The survey consisted of seven 5-point Likert-type items. Globally, participants’ attitudes were positive. One of the items requested informants to express agreement or disagreement towards the statement “being able to see the video made the test easier” (p. 286) and most informants “agreed”. Other authors found evidence that learners had a negative view of the video in listening/audiovisual comprehension tests (Alderson et al., 1995; Coniam, 2001; Suvorov, 2008). As mentioned in the introduction, those mixed results could be due to several reasons (the conflict in visual attention between watching the video and completing the task, the focus of the questions, the complexity of the task and the text, etc.).

Table 8. How do students practise listening/audiovisual comprehension outside the classroom?

	Percent				
	0*	1-2	3-4	5-6	7
Listening to audio recordings intended for language learning	58.3	22.2	16.7	2.8	0
Listening to audio recordings intended for native speakers of	25	13.9	36.1	25	0

Spanish					
Talking to native speakers in Spanish	27.8	41.7	25	2.8	2.8
Talking to non-native speakers in Spanish	19.4	41.7	30.6	2.8	5.6
Watching videos intended for language learning	58.3	16.7	19.4	2.8	2.8
Watching videos intended for native speakers of Spanish	25	27.8	27.8	13.9	5.6
*0 = they do not, 1-2 = once or twice a week, 3-4 = 3 or 4 times a week, 5-6 = 5 or 6 times a week, 7 = everyday					

The most frequent ways to practise listening/audiovisual comprehension outside the classroom were listening to audio recordings and watching video materials intended for native speakers. The results can be related to a study on audiovisual materials by González-Vera and Hornero Corisco (2016). The authors asked 37 English as a foreign language learners what materials they used to practise listening outside the language classroom. Among the five options provided (Podcasts, video clips, DVDs, CDs and other), the most chosen categories were video clips (30%), CDs (28%) and DVDs (23%). In addition, the results revealed that students preferred authentic materials rather than materials intended for language learning. This is in line with the value of realia in some current language teaching methods, such as communicative language teaching and task-based language teaching. As for the results regarding “talking to native speakers in Spanish” and “talking to non-native speakers in Spanish”, they seem logical because, on the one hand, Spanish as a foreign language was being learnt and, on the other hand, in Coimbra there were opportunities to interact with Spanish exchange students, Spanish tourists and Spanish residents.

3.2. Research question 2

The second research question was *Does the use of questions embedded within the video in the form of subtitles and synchronized with the relevant fragments in a FL audiovisual comprehension test facilitate test-takers' performance? Do the test-takers of the experimental group score higher or lower than the test-takers of the control group?*

As described in Casañ Núñez (2016b), the scores of both versions of the test were subject to descriptive statistics, analyses of facility, discrimination and reliability, and correlations with two tasks from a listening test. Overall, the results showed that the instruments were adequate for the research purposes for which they were designed. Just to mention some aspects, Cronbach's alpha was .650 in the traditional test and .705 in the experimental one (according to Suhr & Shay [2009], alphas over .60 are acceptable for instruments developed for research purposes); in both tests, the mean inter-item correlation for the items was in the .2 to .4 range (as recommend by Briggs & Cheek [1986]); and the audiovisual comprehension tests were significantly correlated in a positive way with two tasks from a listening exam (experimental test: $r^s = .592$; $p < .01$; traditional test: $r^s = .833$, $p < .01$).

Table 9. Descriptive Statistics for the tests

Test	N	Min	Max	M	SD	Mdn	Skewness	SE	Kurtosis	SE
Trad.	22	1.00	7.00	5.182	1.652	5.500	-1.016	.491	.731	.953
Exp.	19	2.00	7.00	5.053	1.779	5.000	-.618	.524	-.814	1.014

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

Table 10. Test Statistics*

	Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)	Exact Sig. (2-tailed)	Exact Sig. (1-tailed)	Point Probability
Test	203.500	393.500	-.147	.883	.885	.444	.009

*Grouping Variable: type of test (traditional test, experimental test).

As can be seen in Table 9, the mean scores were very similar in both tests, suggesting that questions embedded within the video did not have an effect on test performance. Since the distributions of the test scores were not normally distributed either on the experimental test ($S-W = .874$, $df = 19$, $p = .017$) or the traditional test ($S-W = .888$, $df = 22$, $p = .017$), it was not appropriate to use an independent-samples t -test. Instead, the Mann-Whitney test was employed (see Table 10). The results revealed no statistically significant difference between the test scores obtained by the control group ($Mdn = 5.500$) and the experimental group ($Mdn = 5.000$), $U = 203.500$, $z = -.147$, p (exact, two-tailed) = .885, $r = -0.023$, which indicated that the subtitled questions did not have any impact on test performance. The results did not confirm the hypothesis that the experimental group would outperform the control group thanks to the help provided by the questions embedded in the video. One can think of two different explanations for that. First, perhaps the hypothesis was wrong. It could be the case that questions embedded in the video as subtitles do not constitute either a help or a hindrance when they are combined with questions on paper because learners do not pay attention to them. Second, the hypothesis and the reasoning behind it were right but they did not hold in the study for some reason. It could be that the groups were not equivalent. Although they had a similar profile, participants did not complete a pre-test to determine their level of audiovisual comprehension in Spanish.

3.3. Research question 3

The third research question investigated test-takers' attitudes towards the use of questions embedded in the video as subtitles. Research question 3.1. inquired to what extent test-takers in the experimental group agreed or disagreed that embedded questions helped them to complete the tasks. A helpfulness scale (items 1, 2.1, 2.2, 2.3, 2.4, 5, 6.1, 6.2 and 6.3) was employed to try to answer this matter. To measure its reliability, Cronbach's alpha, corrected item-total correlation (CITC) and Cronbach's alpha if item deleted (CAID) values were calculated. Cronbach's alpha was .778. According to Cohen et al. (2011: p. 640), alpha values between .70 and .80 indicate that the scale is "reliable". All but one CITC values were above .30 (between .365 and .763), as recommend by De Vaus (2002) and Pallant (2011). Item 6.3 had a CITC level of .254, which is acceptable for Henning (1987, cited in Green, 2013). CAID values were between .709 and .781. In other words, none of the items would substantially increase the reliability if they were deleted. As the analysis was satisfactory, a new variable containing the sum of the component items was created. For that purpose, each "disagreement" was counted as 1 point, each "neither agree nor disagree" as 2 and each "agree" as 3. The scale score was checked for outliers and two were identified (see Figure 6). However, as Osborne and Overbay (2004, p. 3) point out, "there is a great deal of debate as to what to do with identified outliers". Following Larson-Hall's (2010) proposal, statistics with and without them were calculated (see Tables 11 and 12).

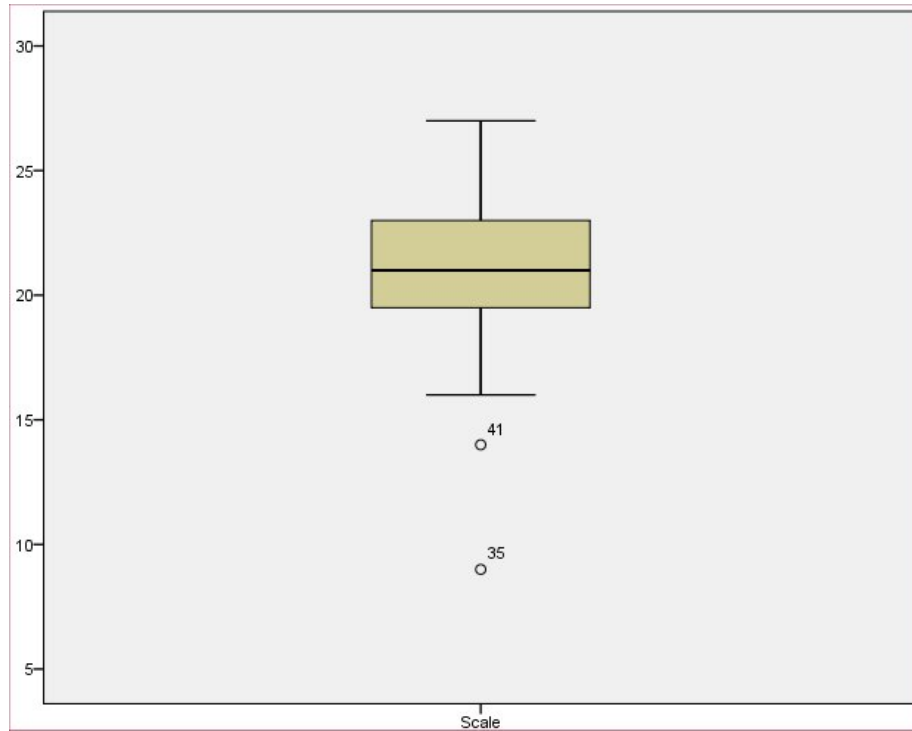


Figure 6. Boxplot of the scale scores.

Table 11. Descriptive Statistics for the scale with two outliers

	N	Min	Max	M	Mdn	SD	Skewness	SE	Kurtosis	SE
Scale	19	9	27	20.42	21	4.059	-1.366	.524	2.563	1.014
Valid N	19									

Table 12. Descriptive Statistics for the scale without outliers

	N	Min	Max	M	Mdn	SD	Skewness	SE	Kurtosis	SE
Scale	17	16	27	21.47	22	2.577	-.297	.550	1.167	1.063
Valid N	17									

The minimum score for the scale could be 9 (for answering “disagree” to all 9 items) and the maximum 27 (for responding “agree” to all 9 items). Therefore, the higher the score, the greater the agreement that embedded questions helped learners complete the test. The mean of the scale with outliers ($M = 20.42$) suggests that, overall, participants considered that imprinted questions were useful. The mean of the scale without outliers ($M = 21.47$) revealed the same fact even more clearly. The results support the hypothesis and they are in line with those obtained in a previous study with SFL learners (Casañ Núñez, 2016a). In that investigation the mean of the scale was 11.50 out of 15 ($SD = 2.565$).

Research question 3.2 explored whether any subtitled question was consistently considered more or less helpful than the others by the experimental group. The results of the Friedman test (see Table 13) indicated that there were statistically significant differences in the ranking of the embedded questions, $\chi^2(6, n = 19) = 16.653, p = .011$. Two embedded questions had particularly low mean ranks: 4 and 7. They differed

from the others in two ways. First, they measured the ability to recognize feelings (see table 3) and second, they did not have a focalising character: they implied global comprehension and the subtitles appeared at the end of the videos. Wilcoxon Signed Rank tests were conducted to follow up the findings. A Bonferroni adjusted alpha value was used to reduce the chances of obtaining false-positive results. As recommended by Field (2009) and Pallant (2011), selective comparisons were chosen to keep the alpha at a manageable level. Two Wilcoxon Signed Rank tests were carried out. Thus, the alpha level of significance was .025. Embedded question 4 (focusing on recognizing feelings, not focalised in the pertinent scene, level A2/B1) was compared with embedded questions 2 (focusing on identifying general ideas, synchronized with the relevant fragment, level A2/B1) and 3 (focusing on extracting specific information, synchronized with the pertinent fragment, level A2). Wilcoxon tests revealed that embedded question 4 was ranked significantly lower than question 2, $z = -2.311$, p (exact, two tailed) = .024, $r = -.375$, and question 3, $z = -2.443$, p (exact, two tailed) = .018, $r = -.396$. These facts could be due to the lack of focalization of embedded question 4. However, it would be necessary to ask test-takers to disclose *how* they evaluated the helpfulness of the imprinted questions.

Table 13. Results from the Friedman test.

Ranks	
Questionnaire item / Subtitled question	Mean Rank
2.1. / 1	3.79
2.2. / 2	4.74
2.3. / 3	4.68
2.4. / 4	3.11
6.1. / 5	3.92
6.2. / 6	4.55
6.3. / 7	3.21

Test Statistics*	
N	19
Chi-Square	16.653
df	6
Asymp. Sig.	.011
*Friedman Test	

Research question 3.3. investigated whether students in the experimental group were in favour of having the questions embedded in the video in the audiovisual comprehension test of the final exam. In order to address that matter, test-takers were asked directly after completing each task of the test (items 3 and 7 of the second questionnaire). The answers were virtually the same both times: most informants were in favour and very

few against this (see Tables 14 and 15). These results suggest that students considered that embedded questions were useful.

Table 14. Item 3 (task 1). Are you in favour of having the questions embedded in the image (in addition to having them on paper) in the audiovisual comprehension test of the final exam?

		Frequency	Percent	Valid Percent
Valid	No	2	10.5	10.5
	I am not sure	4	21.1	21.1
	Yes	13	68.4	68.4
	Total	19	100.0	100.0

Table 15. Item 7 (task 2). Are you in favour of having the questions embedded in the image (in addition to having them on paper) in the audiovisual comprehension test of the final exam?

		Frequency	Percent	Valid Percent
Valid	No	2	10.5	10.5
	I am not sure	3	15.8	15.8
	Yes	14	73.7	73.7
	Total	19	100.0	100.0

4. Conclusion

Listening, watching, reading and writing at the same time in a foreign language is a cognitively demanding task. This paper is part of wider research which investigates the use of audiovisual comprehension questions imprinted in video images in the form of subtitles and synchronized with the relevant fragments, for the purpose of language learning and testing. Compared to viewings where the task is available only on paper, this technique may provide some benefits. Among them, it could reduce the conflict in visual attention between watching the video and completing the task, by spatially and temporally approximating the questions and the relevant scenes. The procedure is mainly intended for students with a low proficiency level.

This pilot multimethod study (Morse, 2003) investigated for the first time whether questions embedded in videos as subtitles had any impact on FL learners' audiovisual comprehension test performance. The results suggest that they do not have any effect on test performance. Test-takers' attitudes towards this technique are positive, however. Overall, participants in the experimental group agree that the imprinted questions help them to complete the tasks. Furthermore, most of them are in favour of having the questions embedded in the video in the audiovisual comprehension test of the final exam. Test-takers' opinions can be paralleled to studies that research experts', SFL students' and SFL teachers' views about this methodology (Casañ Núñez, 2015a, 2016a, in press-b). Globally, all three groups have positive opinions. Experts agree or strongly agree that the technique can be useful for the teaching of listening/audiovisual comprehension, and that it can provide various benefits compared to viewings where the activity is available only on paper; among them, that it can minimize the conflict in visual attention between watching a video and completing a task at the same time, and

that it helps FL students to focus their attention towards the viewing objectives. SFL teachers think that embedded questions are beneficial for FL learning, and they coincide with the experts in some of the advantages. SFL students agreed that imprinted questions helped them to complete a testing task; besides, their comments imply that subtitles focalise attention in the relevant moments, and that they minimize the conflict of visual attention. All in all, although these studies are limited and further empirical research is needed, the positive opinions suggest that this technique has potential benefits for FL learning and testing.

The current study has a number of shortcomings. First, SFL students were not selected by probabilistic sampling. As Dörnyei (2007) points out, this weakness is present in most experimental studies in the social sciences. Second, treatment diffusion constituted a potential threat to internal validity because the groups took the audiovisual test consecutively in the same classroom. Two circumstances reduced the possibility of the exchange of information: (a) there was little time for it to happen, since test-takers took the tests consecutively and they had different class schedules; and (b) in principle, there was no obvious interest in knowing the test content, as it had no impact on the final grade. A further limitation consists in the impossibility of guaranteeing that the groups were of equal ability. Although they were similar in many aspects and there were no statistically significant differences between the groups' performance on two tasks from a listening test, an audiovisual comprehension test to find out the participants' level of audiovisual comprehension in SFL was not administered.

This leads us to some directions for future research. First, it is important to investigate whether embedded questions have an impact on FL learners' viewing behaviour with regard to the video image. Compared to viewings where the task is available only on paper, imprinted questions may reduce the conflict in visual attention between watching the video and completing the task, by spatially and temporally approximating the questions and the relevant fragments. Based on this, it is hypothesized that embedded questions may increase the amount of time devoted to viewing the video by the learner. Second, it would be beneficial to replicate this study with a larger sample and compare the results. Third, it would be useful to carry out similar studies with tests for different target language use domains, other text types and other text lengths. Lastly, as this work explored embedded questions in a testing situation, further research is needed to investigate this technique in a learning context.

Dedication

This article is dedicated to my father.

References

- Alderson, J. C. (2005). *Diagnosing Foreign Language Proficiency: The Interface Between Learning and Assessment*. New York: Continuum.
- Alderson, J. C., Clapham, C. & Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.
- Allen, I. E., & Seaman, C. A. (2007). Likert Scales and Data Analyses. *Quality Progress*, 40(7), 64-65.
- Batty, A. O. (2014). A Comparison of Video- and Audio-mediated Listening Tests with Many-Facet Rasch Modeling and Differential Distractor Functioning. *Language Testing*, 32(1), 3–20. doi: 10.1177/0265532214531254.
- Bejar, I., Douglas, D., Jamieson, J., Nissan, S. & Turner, J. (2000). *TOEFL 2000 Listening Framework: A Working Paper*. Princeton, New Jersey: Educational Testing Service. Retrieved from <http://ets.org/Media/Research/pdf/RM-00-07.pdf>.

- Bisquerra, R. & Pérez-Escoda, N. (2015). ¿Pueden las escalas Likert aumentar en sensibilidad? *REIRE, Revista d'Innovació i Recerca en Educació*, 8(2), 129-147. doi: 10.1344/reire2015.8.2828.
- 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.
- Brown, J. D. (2001). *Using Surveys in Language Programs*. Cambridge: Cambridge University Press.
- Buck, G. (2001): *Assessing Listening*. Cambridge: Cambridge University Press.
- Casañ Núñez, J. C. (2015a). Un marco teórico sobre el uso de preguntas de comprensión audiovisual integradas en el vídeo como subtítulos: un estudio mixto. *MarcoELE*, 20, 1-45. Retrieved from <http://marcoele.com/comprencion-audiovisual-y-subtitulos/>.
- Casañ Núñez, J. C. (2015b). Subtitulación de preguntas de comprensión audiovisual: ejemplificación en una secuencia de *Ópera prima* de Fernando Trueba. *Foro de profesores de ELE*, 11, 45-56. Retrieved from <https://ojs.uv.es/index.php/foroele/article/view/7095>.
- Casañ Núñez, J. C. (2016a). Actividades de comprensión audiovisual con preguntas integradas en forma de subtítulos: la opinión de catorce estudiantes universitarios de español lengua extranjera. *Skopos*, 7, 19-38.
- Casañ Núñez, J. C. (2016b). Desarrollo de una prueba de comprensión audiovisual. *MarcoELE*, 22, 1-70. Retrieved from http://marcoele.com/descargas/22/casan-prueba_audiovisual.pdf.
- Casañ Núñez, J. C. (in press-a). Diseño y fiabilidad de un cuestionario sobre la comprensión auditiva/audiovisual. *Bellaterra Journal of Teaching & Learning Language & Literature*.
- Casañ Núñez, J. C. (in press-b). Tareas de comprensión audiovisual con preguntas subtituladas: valoraciones de cinco profesores universitarios de español como lengua extranjera. *E-JournALL, EuroAmerican Journal of Applied Linguistics and Languages*.
- Celce-Murcia, M. & Olshtain, E. (2000). *Discourse and Context in Language Teaching: A Guide for Language Teachers*. Cambridge: Cambridge University Press.
- Cohen, L., Manion, L. & Morrison, K. (2011). *Research Methods in Education* (7th ed.). New York: Routledge.
- Coniam, D. (2001). The Use of Audio or Video Comprehension as an Assessment Instrument in the Certification of English Language Teachers: A Case Study. *System*, 29(1), 1-14. doi: 10.1016/S0346-251X(00)00057-9.
- Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
- De Vaus, D. A. (2002). *Surveys in Social Research* (5th ed.). New South Wales: Allen & Unwin.
- Dörnyei, Z. (2007). *Research Methods in Applied Linguistics*. Oxford: Oxford University Press.
- Field, A. (2009). *Discovering Statistics Using SPSS* (3rd ed.). London: Sage Publications.
- Field, J. (2008). *Listening in the Language Classroom*. Cambridge: Cambridge University Press.
- Flowerdew, J. & Miller, L. (2005). *Second Language Listening*. New York: Cambridge University Press.
- González-Vera, P. & Hornero Corisco, A. (2016). Audiovisual Materials: A Way to Reinforce Listening Skills in Primary School Teacher Education. *Language Value*, 8(1), 1-

25. doi: 10.6035/LanguageV.2016.8.2. Retrieved from <http://www.e-revistas.uji.es/languagevalue>.
- Green, R. (2013). *Statistical Analyses for Language Testers*. New York: Palgrave.
- Gruba, P. (1993). A Comparison Study of Audio and Video in Language Testing. *JALT Journal*, 15(1), 85-88.
- Gruba, P. (1997). The Role of Video Media in Listening Assessment. *System*, 25(3), 335-345. doi: 10.1016/s0346-251x(97)00026-2.
- Harris, T. (2003). Listening with Your Eyes: The Importance of Speech-Related Gestures in the Language Classroom. *Foreign Language Annals*, 36(2), 180-187. doi: 10.1111/j.1944-9720.2003.tb01468.x.
- Hidalgo de la Torre, R. (Coord.), (2013). *Prepara y practica el DELE B1*. Barcelona: Octaedro.
- Jacoby, J. & Matell, M. S. (1971). Three-Point Likert Scales Are Good Enough. *Journal of Marketing Research*, 8, 495-500.
- Krashen, S. D. & Terrel, T. D. (1983). *The Natural Approach: Language Acquisition in the Classroom*. Oxford: Pergamon Press.
- Larson-Hall, J. (2010). *A Guide to Doing Statistics in Second Language Research Using SPSS*. New York: Routledge. doi: 10.4324/9780203875964.
- Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, 22(140), 1-55.
- Londe, Z. C. (2009). The Effects of Video Media in English as a Second Language Listening Comprehension Tests. *Issues in Applied Linguistics*, 17(1), 41-50.
- Lynch, T. (2009). *Teaching Second Language Listening*. Oxford: Oxford University Press.
- 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: Winter.
- Mayer, R. (2001). *Multimedia Learning*. Cambridge: Cambridge University Press.
- Mayer, R. (2014). *The Cambridge Handbook of Multimedia Learning*. (2nd ed.). Cambridge: Cambridge University Press.
- Martín Peris, E. (2007). La didáctica de la comprensión auditiva. *MarcoELE*, 5. Retrieved from <http://marcoele.com/la-didactica-de-la-comprension-auditiva/> (Original work published in 1991).
- Mendelshon, D. J. (1994). *Learning to Listen: a Strategy Based Approach for the Second Language Learner*. Carlsbad, California: Dominic Press.
- 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, California: Sage.
- Ockey, G. J. (2007). Construct Implications of Including Still Image or Video in Computer-Based Listening Tests. *Language Testing*, 24(4), 517-537. doi: 10.1177/0265532207080771.
- Osborne, J. W. & Overbay, A. (2004). The Power of Outliers (and Why Researchers Should Always Check for Them). *Practical Assessment, Research & Evaluation*, 9(6), 1-12. Retrieved from <http://PAREonline.net/getvn.asp?v=9&n=6>.
- Pallant, J. (2011). *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using SPSS* (4th ed.). Maidenhead: Open University Press/McGraw-Hill.

- Pardo-Ballester, C. (2016). Using Video in Web-Based Listening Tests. *Journal of New Approaches in Educational Research*, 5(2), 91-98. doi: 10.7821/naer.2016.7.170.
- Plass, J. L. & Jones, L. (2005). Multimedia Learning in Second Language Acquisition. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 467-488). Cambridge: Cambridge University Press.
- Riley, P. (1979). Viewing Comprehension: "L'Oeil Écoute". *Mélanges CRAPEL*, 10, 80-95. Retrieved from <http://www.atilf.fr/IMG/pdf/melanges/6riley.pdf>.
- Rubin, J. (1995a). The Contribution of Video to the Development of Competence in Listening. In D. J. Mendelsohn & J. Rubin (Eds.), *A Guide for the Teaching of Second Language Listening* (pp. 151-165). San Diego, California: Dominic Press, Inc.
- Rubin, J. (1995b). An Overview to A Guide for the Teaching of Second Language Listening. In D. J. Mendelsohn & J. Rubin (Eds.), *A Guide for the Teaching of Second Language Listening* (pp. 7-11). San Diego, California: Dominic Press, Inc.
- Saris, W. E. & Gallhofer, I. N. (Eds.). (2014). *Design, Evaluation, and Analysis of Questionnaires for Survey Research* (2nd ed.). doi: 10.1002/9781118634646.
- Stempleski, S. & Tomalin, B. (2001). *Film*. Oxford: Oxford University Press.
- Stoller, F. L. (1992). Using Video in Theme-Based Curricula. In Susan Stempleski & Pual Arcario (Eds.), *Video in Second Language Teaching: Using, Selecting and Producing Video for the Classroom* (pp. 25-46). New York: TESOL.
- Sueyoshi, A. & Hardison, D. M. (2005). The Role of Gestures and Facial Cues in Second Language Listening Comprehension. *Language Learning*, 55(4), 661-669. doi: 10.1111/j.0023-8333.2005.00320.x.
- 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, California. Retrieved from <http://www.lexjansen.com/wuss/2009/anl/ANL-SuhrShay.pdf>.
- Suvorov, R. S. (2008). Context Visuals in L2 Listening Tests: the Effectiveness of Photographs and Video vs. Audio-Only Format. *Retrospective Theses and Dissertations* (paper 15448). Retrieved from <http://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=16447&context=rtd>.
- Suvorov, R. S. (2015). The Use of Eye Tracking in Research on Video-Based Second Language (L2) Listening Assessment: A Comparison of Context Videos and Content Videos. *Language Testing*, 32(4), 463-483. doi: 10.1177/0265532214562099.
- Sweller, J., Ayres, P. & Kalyuga, S. (2011). *Cognitive Load Theory*. London: Springer.
- Underwood, M. (1989). *Teaching Listening*. London: Longman.
- Ur, P. (1994). *Teaching Listening Comprehension* (12th printing): Cambridge: Cambridge University Press.
- Ur, P. (1999). *A Course in Language Teaching: Practice and Theory*. Cambridge: Cambridge University Press.
- Vandergrift, L. & Goh, C. C. M. (2012). *Teaching and Learning Second Language Listening*. New York: 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. Retrieved from <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. doi: 10.1080/15434300802213015.
- Wagner, E. (2010a). The Effect of the Use of Video Texts on ESL Listening Test-Taker Performance. *Language Testing*, 27(4), 493-513. doi: 10.1177/0265532209355668.

Wagner, E. (2010b). Test-Takers' Interaction with an L2 Video Listening Test. *System*, 38(2), 280-291. doi: 10.1016/j.system.2010.01.003.

Wagner, E. (2013). An Investigation of How the Channel of Input and Access to Test Questions Affect L2 Listening Test Performance. *Language Assessment Quarterly*, 10(2), 178–195. doi: 10.1080/15434303.2013.769552.

Appendix

Preguntas integradas como subtítulos: la opinión del estudiante (original version)

El presente cuestionario tiene por finalidad conocer qué piensas sobre el uso de las preguntas de comprensión integradas en el vídeo en forma de subtítulos. No existen respuestas correctas ni incorrectas porque estás expresando tu opinión. La información que proporciones se tratará de forma confidencial y solo se utilizará con fines académicos.

Tarea 1. Fragmento de la película *Ópera prima*.

1. Sobre la base de tu experiencia global completando la tarea, ¿en qué medida estás de acuerdo o en desacuerdo con que las preguntas subtituladas te han ayudado? Señala la respuesta con un círculo.

- a) En desacuerdo
- b) Ni de acuerdo ni en desacuerdo
- c) De acuerdo

2. De forma individual, ¿en qué medida estás de acuerdo o en desacuerdo con que las preguntas subtituladas te han ayudado? Completa la tabla poniendo equis (X).

	En desacuerdo	Ni de acuerdo ni en desacuerdo	De acuerdo
2.1. ¿Cómo se llama la chica?			
2.2. ¿De qué hablan?			
2.3. ¿Cuál es el teléfono de la chica?			
2.4. ¿Qué sentimientos puede tener el chico por la chica?			

3. ¿Estás a favor de que en la prueba de comprensión audiovisual las preguntas aparezcan en la imagen (además de en papel)?

- a) No
- b) No estoy seguro
- c) Sí

4. A continuación tienes un espacio en el que puedes comentar cualquier aspecto de la actividad. Puedes responder en español o portugués.

Tarea 2. Fragmento de la película *Los peores años de nuestra vida*

5. Sobre la base de tu experiencia global completando la tarea, ¿en qué medida estás de acuerdo o en desacuerdo con que las preguntas subtituladas te han ayudado? Señala la respuesta con un círculo.

- a) En desacuerdo
- b) Ni de acuerdo ni en desacuerdo
- c) De acuerdo

6. De forma individual, ¿en qué medida estás de acuerdo o en desacuerdo con que las preguntas subtituladas te han ayudado? Completa la tabla poniendo equis (X).

	En desacuerdo	Ni de acuerdo ni en desacuerdo	De acuerdo
6.1. ¿De qué temas habla el chico pelirrojo?			
6.2. ¿Cómo se llama la chica?			
6.3. Considerando todo el fragmento, ¿qué sentimientos puede tener el chico pelirrojo por la chica? Justifica brevemente la respuesta.			

7. ¿Estás a favor de que en la prueba de comprensión audiovisual las preguntas aparezcan en la imagen (además de en papel)?

- a) No
- b) No estoy seguro
- c) Sí

8. A continuación tienes un espacio en el que puedes comentar cualquier aspecto de la actividad. Puedes responder en español o portugués.

Questions embedded as subtitles: students' opinion (English translation)

This questionnaire aims to know what you think about the use of audiovisual comprehension questions embedded in the video in the form of subtitles. There are no right or wrong answers because you are expressing your opinion. The information you provide will be treated confidentially and will only be used for academic purposes.

Task 1. Fragment from the film *Ópera prima*.

1. Based on your overall experience completing the task, to what extent do you agree or disagree that the subtitled questions have helped you? Mark your answer with a circle.

- a) Disagree
- b) Neither agree nor disagree
- c) Agree

2. Individually, to what extent do you agree or disagree that subtitled questions have helped you? Complete the table with "X".

	Disagree	Neither agree nor disagree	Agree
2.1. What is the girl's name?			
2.2. What are they talking about?			

2.3. What is the girl's telephone number?			
2.4. What feelings may the boy have for the girl?			

3. Are you in favour of having the questions embedded in the image (in addition to having them on paper) in the audiovisual comprehension test of the final exam?

- a) No
- b) I am not sure
- c) Yes

4. Next there is a space where you can comment on any aspect of the activity. You may reply in Spanish or Portuguese.

Task 2. Fragment of the film *Los peores años de nuestra vida*.

5. Based on your overall experience completing the task, to what extent do you agree or disagree that the subtitled questions have helped you? Mark your answer with a circle.

- a) Disagree
- b) Neither agree nor disagree
- c) Agree

6. Individually, to what extent do you agree or disagree that subtitled questions have helped you? Complete the table with "X".

	Disagree	Neither agree nor disagree	Agree
6.1. What is the redheaded boy talking about?			
6.2. What is the girl's name?			
6.3. What feelings may the red-haired boy have for the girl?			

7. Are you in favour of having the questions embedded in the image (in addition to having them on paper) in the audiovisual comprehension test of the final exam?

- a) No
- b) I am not sure
- c) Yes

8. Next there is a space where you can comment on any aspect of the activity. You may reply in Spanish or Portuguese.
