Computer Assisted Language Instruction Consortium

CALICO JOURNAL (ONLINE) ISSN 2056–9017

Article

The Effects of Captions on EFL Learners' Comprehension of English-Language Television Programs

Michael P. H. Rodgers[1] and Stuart Webb[2]

Abstract

The Multimedia Principle (Fletcher & Tobias, 2005) states that people learn better and comprehend more when words and pictures are presented together. The potential for English language learners to increase their comprehension of video through the use of captions, which graphically display the same language as the spoken dialogue, has been documented in previous research. However, studies have generally used short videos (Markham & Peter, 2003; Montero Perez, Peters, & Desmet, 2014) or videos designed for language learning (Chung, 1999) rather than episodes of L2 television programs that students are most likely to watch on their own outside of the classroom. The present study aimed to fill this gap by investigating the comprehension of 372 Japanese university students who watched ten 42-minute episodes of an American television program with and without captions. While viewing the episodes, the participants completed comprehension tests. Analysis indicated that although the participants who viewed the episodes with captions had comprehension scores that were slightly higher across all episodes, their scores were only significantly different for three of the ten episodes. The results revealed that captions are likely to aid comprehension when episodes are most difficult. Explanations for the findings and pedagogical applications are offered.

KEYWORDS: COMPREHENSION, CAPTIONS, TELEVISION, LANGUAGE LEARNING

Affiliations

[1]University of Nottingham.
email: Michael.Rodgers@nottingham.ac.uk

[2]University of Western Ontario.
email: swebb27@uwo.ca

CALICO JOURNAL VOL 34.1 2017 20–38

doi : 10.1558/cj.29522

equinoxonline

Introduction

It is well established that suitable and sufficient language input is a vital component of language learning. This input, in both its written and spoken form, should be authentic and comprehensible. However, in the English as a Foreign Language (EFL) setting it may be a challenge to provide second language (L2) learners with sufficient aural input meeting these conditions. Television programs intended for an English-speaking audience may be a potential source of L2 aural input, and the authenticity of episodes of television is indisputable. However, whether they are comprehensible for all but the highest proficiency language learners (LLs) is a concern. One feature of television programs that may make them comprehensible for less proficient learners of English is captions.

Captions are a transcription of the spoken text that appears simultaneously at the bottom of the screen as a video plays. Originally, captions were intended as a service for the hearing impaired, but they have long been used in language-learning situations. The option of displaying captions while viewing television programs is standard on commercial DVDs and streaming videos.

Background

Support for language learning through viewing video with captions emanates from the Multimedia Principle and its application to second language learning (Fletcher & Tobias, 2005). The idea behind this principle is that people learn better and comprehend more when words and pictures are presented together. The combination of aural and visual input gives viewers the opportunity to comprehend information through different channels and make connections between them. Studies specific to captioned video support this by demonstrating that captions improve comprehension, because they allow learners to utilize their reading skills to enhance their aural comprehension (Garza, 1991). This is particularly important because lower proficiency LLs have been shown to have listening comprehension levels lower than their reading comprehension (Hirai, 1999). Through captions, learners can break down aural input from video into meaningful units and visualize it (Ellis, 2005), which is especially important if the input is to some extent beyond their comprehension ability (Danan, 2004). Research has also indicated that the presence of captions distracts little from the observation of onscreen details that support comprehension, and consequently captions may not compromise the value of imagery in television programs (Danan, 2004). In this way, the presence of captions may lead to increased comprehension of episodes of television for LLs, but there appears to be some variability in the degree to which comprehension improves depending on the study.

A number of studies have investigated the effects of captions on the comprehension of video (Chung, 1999; Guillory, 1998; Huang & Eskey, 1999; Markham & Peter, 2003; Markham, Peter, & McCarthy, 2001; Montero Perez, Peters, & Desmet, 2014). Findings from this research point to a comprehension advantage for learners who have access to captions while viewing videos. Throughout this survey of caption studies, the percentage difference between the results for the Captions Group (CG) and those of the No Captions Group (NCG) is used to illustrate the degree to which the presence of captions affected comprehension. This provides a means for direct comparison of studies that use different measurement instruments.

Guillory (1998) compared the use of captions, keyword captions, and no captions on comprehension of two educational videos for French language learning. Keyword captions are words identified as important to the video. To measure comprehension, participants ($N = 202$) in each treatment group completed 14 short-answer items that focused on the recall of details and inferencing from the information presented in each video. The results of the CG ($M = 10.1$) and the Keyword Captions Group ($M = 9.2$) were significantly higher than those of the NCG ($M = 7.3$), but there was no significant difference between the two captions groups. The CG's mean score was 38.1% higher than the NCG's.

Using a video designed for English language learning, Huang and Eskey (1999) had half their participants ($N = 30$) view a captioned version of a 21-minute educational video twice while the other half viewed an uncaptioned version twice. A comprehension test, presented aurally, followed the second viewing. The CG ($M = 10.87$ out of 16) scored significantly higher than the NCG ($M = 7.67$). The participants who viewed the video with captions had 41.7% higher comprehension scores than those who did not have access to captions.

Chung (1999) compared the comprehension of English-language educational videos viewed by learners ($N = 183$) under four treatment conditions: advance organizers preceding the videos, captioned video, both advance organizers and captioned video, and uncaptioned video. Comprehension was measured through a set of 10 multiple-choice items for each video. The captions treatment ($M = 7.66$ out of 10) and the advance organizer with captions treatment ($M = 7.98$) had significantly higher results than those from the advance organizer-only ($M = 6.98$) and the no captions treatments ($M = 6.69$). The results from the captions treatment were 14.4% higher than the results from the no captions treatment.

In a pair of studies, Markham et al. (2001) and Markham and Peter (2003) compared the effects of captions on LLs' comprehension of a documentary. The CG significantly outperformed the NCG in both studies. In the 2001 study

(N = 169), comprehension was measured with a written summary which was scored by the number of idea units produced. On this test, the mean scores were 10.97 and 8.47 for the CG and the NCG, respectively. The CG's mean score was 29.5% higher than the NCG. In the 2003 study (N = 213), comprehension was measured with a multiple-choice listening test. The mean scores (out of 20) were 10.12 and 7.81 for the CG and the NCG, respectively. The CG's mean score was 29.6% higher than the NCG.

Montero Perez et al. (2014) investigated the effect of captions and keyword captions on the comprehension of three short French videos by Flemish-speaking university students (N = 226). Comprehension was measured by 43 items across the three news videos. The CG had a mean score of 29.8, the Keyword Captions Group had a mean score of 27.11, and the NCG had a mean score of 26.45. The CG scored significantly higher (12.7%) than the NCG.

Taken as a whole, the results from previous research have indicated that providing LLs with captions can be an effective method for increasing comprehension of video. Learners across a variety of target languages and a range of proficiency levels had substantial gains in comprehension compared to learners who were not provided with captions. Across the six studies presented here, scores on comprehensions tests were on average 27.7% higher for learners who viewed videos with captions. Two features of these studies, however, are striking: the types of videos and the length of the videos. Three of these studies used a documentary, two used videos intended for LLs, and one used news clips. Arguably, these types of videos may not be the kind that learners would choose to watch for language learning. Learners may be more inclined to watch episodes of authentic television of the type they typically view in their first language (L1) (Webb, 2014). Episodes of television, by and large, tell a complete story which in turn relates to the development of a story arc across a season of episodes. This building of a story arc over multiple episodes may aid comprehension as the learner develops greater knowledge of the characters and the setting with each viewing. This cumulative build-up of knowledge means that learners can approach later episodes better prepared for comprehension than if they were to view distinct and unrelated videos. The amount of total viewing time (where stated) in these studies ranged from 7 to 30 minutes. In order to obtain the large amounts of necessary L2 aural input, learners would be advised to view full-length television episodes (ranging in length from 22 to 42 minutes). The relative brevity of the input videos and the types of videos viewed in previous studies indicate that further research is needed to investigate the effects of captions on comprehension of full-length television programs.

This study was designed to answer the following research questions:

1. Does the presence of captions affect comprehension gains from the first episode to the tenth episode of English-language television viewed?
2. Does comprehension of successive episodes of English-language television change when viewed with and without captions?

Method

Participants

There were 488 participants in the first and second year of a Japanese university from 15 separate classes in this study. All participants had studied English for a minimum of seven years, and their level of proficiency can be considered pre-intermediate to intermediate within the context of the university. The participants were from a range of majors: business, commerce, engineering, law, language, pharmacy, and physical education. All classes were taught by the first author.

Participants were divided into two treatment groups: Captions Group and No Captions Group. At the beginning of the study, there were 73 participants (44 male; 29 female) in the CG and 415 participants (282 male; 133 female) in the NCG. Participants were excluded from the study if they missed viewing the initial or final episode, or more than one episode. It was believed that missing a single episode would not be a serious detriment to subsequent comprehension but missing two or more would have a negative effect. After exclusions, there were 51 participants in the CG and 321 participants in the NCG. The different sample sizes for the groups came about as a result of the practicalities of the research situation. The NCG data was from a larger data set collected at multiple instances, and the CG data was collected only once. Care was taken, however, to choose participants for the CG with similar proficiencies to those in the NCG.

As an indicator of language proficiency, the participants were administered the Vocabulary Levels Test (VLT) (Schmitt, Schmitt, & Clapham, 2001) at the 2,000-, 3,000-, and 5,000-word levels. These tests provide a measure of receptive vocabulary knowledge. An independent-samples *t*-test ($t(370) = .939$, $p = .370$) indicated no significant difference between the combined scores of these tests for the CG ($M = 49.0$, $SD = 5.0$) and NCG ($M = 48.8$, $SD = 6.7$).

A pilot study was conducted with 32 students from the same university. They had an English proficiency level and language-learning background similar to the participants. The purpose of the pilot was to determine the time requirements of the study, the suitability of the materials, and to provide data for an item analysis of the comprehension tests.

Materials

The Television Program

The television program that served as the input text for this study was called *Chuck* (Schwartz, Fedak, & McG, 2007). It is a series first broadcast in the USA in 2007 and released on DVD in 2008. The genre of the series is drama, but it also has elements of action and comedy. This series was chosen for four reasons. First, *Chuck* is an American drama and had a comparable lexical load to other series in that genre. American dramas were found to be a less lexically demanding genre in earlier research (Rodgers & Webb, 2011; Webb & Rodgers, 2009). The second reason was that the first season is serial in nature which may allow viewers to acquire background knowledge more easily than in a series where episodes are only loosely connected. Third, *Chuck* was received favorably in the pilot study, with 94% of participants rating it as *very enjoyable*. The fourth reason was that the series was not broadcast or available in Japan at the time of the study. This lessened the probability of participants having seen any of the episodes. If participants had previously seen episodes, they may have acquired some knowledge of the series, which may have improved their performance on the comprehension tests.

Ten episodes from the first season of *Chuck* were used in this study: Episodes 1 through 8, Episode 12, and Episode 13. The first eight episodes were used because they were successive episodes and each was part of a general story arc. Episode 12 and Episode 13 were selected for testing purposes because they were more self-contained and more independent from the first season's story arc. The episodes used in this study have an average running time of 42 minutes and 49 seconds, ranging from 41 minutes and 15 seconds to 43 minutes and 18 seconds.

There were two viewing orders for the NCG: (1) Episode 12 → Episodes 1 to 8 → Episode 13, and (2) Episode 13 → Episodes 1 to 8 → Episode 12. This was because, prior to analysis, it was unknown if Episode 12 or Episode 13 was more difficult than the other. Accordingly, this counterbalanced design was implemented to control for the possibly different levels of difficulty and to allow for an analysis of comprehension gain. For the NCG, analysis indicated that there was no significant difference between the comprehension scores of these episodes, and thus the CG group had only one viewing order, and scores for episodes in the same viewing position are analyzed together and referred to as Initial and Final Episode.

Comprehension Tests

To measure the participants' comprehension of *Chuck*, a comprehension test was created for each of the ten episodes. Before creating test items, it

was decided that each episode would be divided into six viewing sections of approximately seven minutes each. The rationale for this was that, if participants were to answer the comprehension questions following the episode, it would be a challenge for the participants to remember details from throughout the episode. Sets of comprehension questions were based on a single viewing section.

The listening comprehension tests created for the episodes were designed around Buck's (2001) default listening construct. Buck proposes a competency-based definition in which he identifies three required abilities: the ability to automatically process extended samples of realistic spoken language in real time, the ability to comprehend the information explicitly included in the text, and the ability to make inferences from unambiguously presented information in the text (p. 114). To this end, items designed to measure the comprehension of details and inferencing ability were included. The listening construct outlined by Buck was also augmented by including a characteristic important to this listening situation: the ability to understand topics contained in relatively lengthy viewing sections and in the episodes as a whole. Items that call for identifying the topic or main idea of a text is an aspect of listening comprehension that has commonly been featured in taxonomies of listening skills (cf. Richards, 1983) and previous listening comprehension research (cf. Wagner, 2002).

There were a number of considerations in the choice of item-type to be used on the comprehension tests based on the listening texts used in this study. First, when testing listening comprehension, 42 minutes is a relatively long text with a lot of content to test, so an item-type that could be quickly answered during and following viewing sections was desirable. The item-type also had to function suitably as detail, inference, and topic questions. With these conditions in mind and because listening comprehension tests that present a variety of item-types better reflect the trait of listening comprehension (Shohamy & Inbar, 1991) and are considered fairer to test takers (Spaan, 2007), it was decided to use a number of different item-types. To test comprehension of the viewing sections, a combination of true/false and multiple-choice items were used. Examples of comprehension test items are shown in Figures 1 and 2.

T *or* F Most of the people at the party are doctors.
T *or* F Chuck's major in university was accounting.
T *or* F Chuck had a girlfriend in university named Jill.

Figure 1. First three true/false items on the comprehension test for Episode 1 of *Chuck*

How does Ellie feel about Morgan?

A. She doesn't know who he is.
B. She thinks he is annoying.
C. She likes him as much as her brother.

Figure 2. Item #10 on the comprehension test for Episode 1 of *Chuck*

There was a third type of question used on the comprehension tests: sequencing items. Sequencing items measure whether participants recognize the overall order of ideas in a text and were included to measure global comprehension and the participants' ability to process input video as a whole (Richards, 1983). Comprehension of a sequence of events has been described as an important facet of listening comprehension (Brett, 1995). An example of sequencing items is shown in Figure 3. Participants needed to indicate the order of the sequence in the blanks and the first and seventh items were always provided.

___	A ninja tries to steal Chuck's computer.
___	Casey and Sarah realize that all of their secrets of the Intersect are in Chuck's head.
___	Chuck and Morgan try to escape from Chuck's birthday party.
___	Casey starts work at BuyMore.
___	Chuck uses a computer virus to stop a bomb.
___	Chuck and Sarah go onto the roof of a building.
__1	The episode begins.
___	Chuck meets Sarah again at the BuyMore and they make a date.
__7__	Chuck and Sarah go to a nightclub.
___	Chuck gets an email message from Bryce and images flash in his head.
___	Chuck takes Sarah and Casey to a hotel.
___	Sarah comes to the BuyMore to get her phone fixed.

Figure 3. Sequencing items for Episode 1 of *Chuck*

After all the items on each comprehension test were created, they were translated into Japanese. The items were presented in the participants' L1 for three reasons: comprehension questions presented this way can make a test easier, questions in the L1 can reduce test-taker anxiety, and questions presented in the L2 may measure reading comprehension as much as they are measuring listening comprehension (Shohamy, 1984). All translations were done by a single Japanese native speaker who had viewed each episode of

Chuck. The translations were done in consultation with the first author to ensure that the translated items were asking the same questions intended in the original English items.

The results from the pilot study were used to perform an item analysis of the original comprehension tests. This involved investigating whether the test items actually measure the intended underlying trait they were designed to measure: listening comprehension ability. The analysis of the responses to the test items was carried out via single parameter item response theory (Rasch analysis). This is essentially a search for items that do not fit the model and provides a rationale for their modification or exclusion. Items that do not fit the model are suspect items, and their inclusion on a comprehension test may be detrimental to construct validity. Problematic items were either rewritten, when an obvious issue was identified, or deleted, when none was apparent. Prior to the item-analysis procedure, each comprehension test had 82 items in total (30 true/false items, 42 multiple-choice items, and 10 sequencing items). Following the item-analysis procedure, the number of items on the comprehension tests ranged from 70 (Episode 1) to 78 (Episode 8) with an average of 74.2 items per episode test. There was an average of 5.6 topic-based items, 37.3 detail-based items, and 21.3 inference questions per comprehension test. The sequencing items made up the remainder of each comprehension test.

Procedures

This study took place over 13 sessions with each usually separated by a week. In each of these sessions, participants either completed tests or administrative procedures for the study, or viewed an episode of television and completed a corresponding comprehension test.

In the first session, the participants completed the ethics approval procedure. At this point, any questions about what it meant to be a participant in the study were addressed. The participants also completed the three levels of the VLT. In the second session, the participants completed a practice session of television viewing and comprehension question answering using an unrelated video. The comprehension test for this video was made to resemble those used in the study, and care was taken to make sure that the participants understood the format of each item-type and knew how to fill in the answer sheets correctly.

From session 3 through 12, participants viewed episodes of *Chuck* and completed the corresponding comprehension test. The procedure for each of these teaching sessions was identical. The comprehension items for the viewing sections were distributed and the participants were given approximately 30 seconds to preview the questions. This was followed by the viewing section. At the

end of the section, the video was paused and the participants were given 3.5 minutes to complete the questions. When participants had finished answering a set of comprehension questions for a viewing section, those questions were collected so that participants could not use information from items in that section to answer items for subsequent viewing sections. This procedure continued for each viewing section until the episode was completed. When the comprehension questions from the final viewing section were collected, participants received the sequencing items. The participants were given ten minutes to complete this task.

Analysis

Participants who were absent from a single teaching session for Episodes 1 to 8 had their missing comprehension test scores replaced using the Expectation Maximization Algorithm. Missing comprehension test data was replaced 55 times over the eight episodes ranging from 3 to 12 times per episode.

The participants' scores on the comprehension tests are expressed as CHIPs. These are units of measurement produced when test results are analyzed using the Rasch model, where measurement is often expressed in logits. A logit is the natural log of the odds of a participant successfully answering different items on a test. The CHIPs scale is a modified version of the logit scale which has no negative numbers the way the logit scale does, and has the more familiar range of 0 to 100. The midpoint of the scale is set at 50 CHIPs, which represents the average difficulty of all the items on the test. Therefore, a participant's score on a comprehension test can be easily interpreted in relation to the average difficulty of the items on that test. Scores on the CHIPs scale are interval data, which is an assumption in statistical analyses such as ANOVAs. Whenever possible, scores for tests are also expressed as percentages for ease of reference.

Results

To compare how comprehension of *Chuck* changed from the first to the tenth episode with and without captions, the comprehension test scores for the Initial Episode and Final Episode were analyzed. As shown in Table 1, the mean score for the first episode viewed by the CG was 60.1% (52.3 CHIPs) and on the final episode viewed 63.0% (53.5 CHIPs). For the NCG, the mean score was 53.7% (51.0 CHIPs) on the Initial Episode and 61.7% (53.2 CHIPs) on the Final Episode. The group that viewed the episodes with captions had a mean gain of 2.9% (1.2 CHIPs) and the group that viewed the episodes without captions had a mean gain of 8.0% (2.2 CHIPs).

Table 1
Mean Gains from the Initial Episode to the Final Episode of *Chuck* for the CG (n = 51) and NCG (n = 321)

	% Score	CHIPs score	*SD*	Min.	Max.
Captions Group					
Initial episode	60.1	52.3	1.68	47.2	55.3
Final episode	63.0	53.5	2.25	47.9	58.3
Mean gain	2.9	1.2	2.27	-5.7	5.8
No Captions Group					
Initial episode	53.7	51.0	2.11	44.9	55.8
Final episode	61.7	53.2	2.62	43.4	59.9
Mean gain	8.0	2.2	2.44	-6.1	9.3

A paired-samples t-test was conducted to compare comprehension scores of the first and tenth episode viewed for both treatment groups. Prior to performing all t-tests on the data, informal analyses of the distribution of the scores using histograms and using normal Q-Q plots revealed no serious threats to the assumption of normality or homogeneity of variance. There was a significant difference for the CG in the comprehension scores between the Initial Episode (M = 52.3, SD = 1.682) and Final Episode (M = 53.5, SD = 2.248); $t(50) = 3.815$, $p < .001$. The effect size as measured by d was 0.61, a value corresponding to a medium treatment effect (Cohen, 1977). For the NCG, there was also a significant difference in the comprehension scores for the Initial Episode (M = 50.701, SD = 2.140) and the Final Episode (M = 52.921, SD = 2.422); $t(225) = 14.679$, $p < .001$. The effect size was 0.97 (large treatment effect).

To investigate whether there was a difference in the comprehension increase between the groups, gains between the Initial and Final episodes viewed were compared with an independent-samples t-test. Participants who viewed the episodes with captions (M = 1.212, SD = 2.268) had a significantly smaller gain in comprehension than those participants who viewed the episodes without captions (M = 2.250, SD = 2.439); $t(370) = 2.850$, $p < .01$. The effect size was 0.44 (small).

To examine the comprehension of the ten episodes with and without captions, the comprehension test scores from the two treatment groups were compared. Tables 2 and 3 present the results for each comprehension test for the Captions and No Captions groups, respectively.

Table 2
Mean Comprehension Scores on the Comprehension Tests for All Episodes of *Chuck* for the CG ($n = 51$)

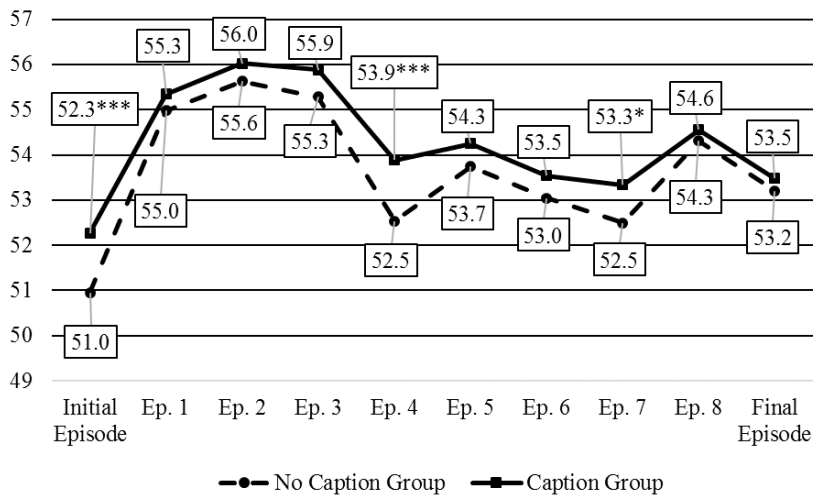
	% score	CHIPs score	*SD*	Min.	Max.
Initial episode	60.1	52.3	1.68	47.2	55.3
Episode 1	71.3	55.3	3.17	46.4	61.9
Episode 2	73.0	56.0	2.83	48.7	62.1
Episode 3	72.0	55.9	3.73	46.8	62.8
Episode 4	65.5	53.9	2.77	47.2	59.6
Episode 5	66.7	54.3	3.42	45.4	61.5
Episode 6	63.7	53.5	3.29	45.3	60.0
Episode 7	62.6	53.3	2.99	46.1	59.2
Episode 8	67.2	54.6	2.88	47.1	62.2
Final episode	63.0	53.5	2.25	47.9	58.3

Table 3
Mean Comprehension Scores on the Comprehension Tests for All Episodes of *Chuck* for the NCG (n =321)

	% score	CHIPs score	*SD*	Min.	Max.
Initial episode	53.7	51.0	2.11	44.9	55.8
Episode 1	69.2	55.0	2.82	45.9	62.3
Episode 2	70.8	55.6	2.84	47.4	62.5
Episode 3	70.2	55.3	2.65	45.0	66.2
Episode 4	60.8	52.5	2.57	42.6	58.9
Episode 5	64.0	53.7	2.65	44.8	61.9
Episode 6	62.2	53.0	2.65	46.3	60.1
Episode 7	60.2	52.5	2.50	45.3	60.2
Episode 8	66.1	54.3	2.70	45.3	61.4
Final episode	61.7	53.2	2.62	43.4	59.9

For all ten episodes of *Chuck*, the CHIPs scores for both groups show that comprehension varied by episode, but after the Initial Episode the general

trend was an increase in comprehension. The average difference between the two groups across the ten episodes was 2.62% (.65 CHIPs). Results on the comprehension tests followed a similar pattern, with episodes that produced lower or higher comprehension scores for the CG also producing lower or higher comprehension scores for the NCG. For both groups of participants, the episode with the lowest comprehension score was the Initial Episode and the episode with the highest comprehension score was Episode 2. Figure 4 plots the mean scores on the comprehension tests measured in CHIPs across the ten episodes for the Captions and No Captions groups.



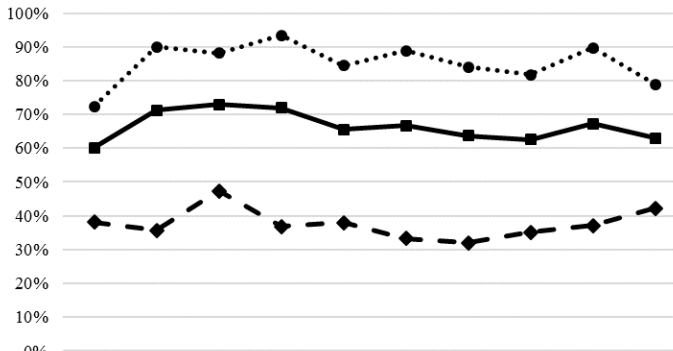
Note. *$p < .05$, ***$p < .001$

Figure 4. Mean CHIPs comprehension scores across all ten episodes of *Chuck* for the CG and the NCG

To determine whether there were any statistically significant differences between the Captions and No Captions groups, a series of one-way ANOVAs was conducted comparing the comprehension scores for each of the ten episodes. The ANOVAs revealed that there was a significant difference between those who watched with captions and those who watched without captions for three of the episodes: Initial Episode [$F(1,370) = 17.864$, $p < .001$], Episode 4 [$F(1,370) = 11.882$, $p = .001$], and Episode 7 [$F(1,370) = 4.798$, $p < .05$]. The effect size as measured by η^2 was 0.046 for the Initial Episode, 0.031 for Episode 4, and 0.013 for Episode 7, all three being a small effect size.

There was a good deal of variation in the participants' comprehension test results. This is illustrated in Figure 5, which plots the maximum, average, and

minimum comprehension test score for each of the ten episodes for the CG. The mean comprehension score is consistently over 60%, and over 70% for three episodes. The minimum comprehension test score ranged from 32.0% (Episode 6) to 47.4% (Episode 2) while the maximum comprehension test score ranged from 72.4% (Initial Episode) to 93.4% (Episode 3). This variation between participants' comprehension test scores of those who viewed the episodes with captions indicates that there were members of the sample who were able to achieve a considerable level of comprehension while others were not.



	Initial	Ep. 1	Ep. 2	Ep. 3	Ep. 4	Ep. 5	Ep. 6	Ep. 7	Ep. 8	Final
Maximum	72.4%	90.0%	88.2%	93.4%	84.5%	88.9%	84.0%	81.8%	89.7%	78.9%
Mean	60.1%	71.3%	73.0%	72.0%	65.5%	66.7%	63.7%	62.6%	67.2%	63.0%
Minimum	38.2%	35.7%	47.4%	36.8%	38.0%	33.3%	32.0%	35.1%	37.2%	42.3%

Figure 5. Maximum, mean, and minimum comprehension scores expressed as percentages across all ten episodes of *Chuck* for the CG

Discussion

Comprehension Gains

Both the Captions Group and the No Captions Group made significant gains in comprehension from the first to the tenth episode of *Chuck*. Participants in the NCG had significantly greater gains, which can be attributed to the lower comprehension scores they had for the Initial Episode. By the Final Episode, there was no significant difference between the comprehension scores of the treatment groups. This suggests that for the first episode viewed, the CG may have been able to make use of the captions to improve comprehension. The NCG did not have this support. However, by the tenth episode viewed, it appears that the NCG was able to make use of accumulated knowledge of the series to a level where the added support of the captions did not make a significant difference in comprehension.

Comprehension of All Episodes

A comparison of the comprehension results for the CG and the NCG across the ten episodes viewed reveals some apparent differences between the treatments. The mean comprehension scores for all episodes were higher when captions were present. The comprehension results for the CG, however, were only shown to be significantly higher than those of the NCG for three of the ten episodes. The Initial Episode was one of the significantly higher episodes, indicating that when the participants are viewing a television series for the first time, the captions are beneficial for comprehension. The two other episodes with a significant difference between comprehension scores (Episodes 4 and 7) were episodes where the NCG had their lowest comprehension scores after the first episode. This indicates that, for certain episodes, regardless of the participants' familiarity with a series, the presence of captions can significantly assist comprehension, particularly when the episode is comparatively difficult.

The difference in comprehension scores between the CG and the NCG in this study does not fully support the findings of prior research. In previous studies, LLs with access to captions had significantly higher comprehension scores (Chung, 1999; Guillory, 1998; Huang & Eskey, 1999; Markham & Peter, 2003; Markham et al., 2001; Montero Perez et al., 2014). In these studies, comprehension scores were on average 27.7% higher when learners viewed videos with captions. In this study, for the three episodes with a significant difference between comprehension scores of the treatment groups, the mean difference was 4.5% (Initial Episode: 6.4%, Episode 4: 4.7%, Episode 7: 2.4%). The comparatively smaller differences in comprehension scores between the CG and the NCG may be a result of the type of video used in this study. In the six studies cited, educational videos, documentaries, and news clips were used. The average viewing times for each of the videos was 3.9 minutes. In this study, however, participants viewed episodes of television over 40 minutes in length. The series, *Chuck*, had a season-long story-arc, with each episode furthering that storyline. However, each episode was still designed to be understood and enjoyed as a separate entity, even with little or no background to the series, with each episode telling a complete story. This may not be the case for short documentaries or educational programs which are designed to convey information to viewers. In the absence of complete narratives developed in detail over a relatively long period of an episode, it appears that LLs in previous studies might have relied on captions considerably more for comprehension leading to larger differences in comprehension scores between the CG and the NCG. In this study, even if learners did not have access to captions, they may have been able to comprehend the story more effectively than the learners in previous research did due to the type of video viewed.

There was considerable variation between participants with minimum and maximum comprehension test scores as well as between episodes of *Chuck* viewed. That comprehension varied from episode to episode was not unexpected. Unlike the Initial and Final episodes, where procedures were taken to control for the relative difficulty of the episodes and the comprehension tests, it was never assumed that the successive episodes would be equivalently difficult for LLs. Each of the eight successive episodes can be considered a different viewing text, and it is unreasonable to think that different texts, even if they are from the same television series, would be equally comprehensible to LLs. There are many factors that might affect comprehension scores from episode to episode, including those based on listening comprehension: accent, pronunciation, hesitations, connected speech, prosody, speaker speed, language proficiency, and the length of the listening text (Buck, 2001). Other factors affecting comprehension are specific to viewing videos: visual literacy of participants, relationship of images to audio, interest in the text by the participants, and video type (Gruba, 2004; Wagner, 2002). Determining how these factors possibly contributed to the differing comprehension scores of the episodes is beyond the scope of this research. The findings, however, are important as they establish that LLs' comprehension of authentic episodes of television can vary episode by episode, regardless of the presence of captions, and could be the impetus for future research on exactly how and to what extent comprehension of television is influenced by these factors and how the presence of captions affects this relationship.

Limitations

This study made use of a single television program from a single genre at a relatively standard viewing time throughout. It is unclear whether the findings are generalizable to different television programs of differing lengths. It is difficult to say whether there was something intrinsic about the episodes of *Chuck* used in this study that diminished the effectiveness of captions for increasing comprehension that has been demonstrated in other studies. Replication of this study utilizing a variety of television programs and language measures may provide a more conclusive assessment of the effects of captions on language learning through viewing television. This study could also be replicated with learners across a broader range of language proficiencies and with learners with other L1s. There is also a need for more use of mixed-methods approaches in research on viewing television with captions, including qualitative data from observations, surveys, and interviews that may provide information on the learners' perceptions of caption use.

Pedagogical Implications

Despite these limitations, the results from this study have a bearing on learning through viewing episodes of television. The most salient effect of the availability of captions was increased comprehension of episodes near the beginning of the viewing process and for episodes where comprehension may be more difficult. Consequently, LLs should be encouraged to make use of captions when they first begin viewing a series and when they believe they are having comprehension problems with subsequent episodes. Supporting comprehension in these ways would be particularly useful for television series with interrelated episodes, where failure to fully comprehend one episode may lead to comprehension problems in subsequent episodes. Learners might believe that their understanding of authentic television without the use of captions is considerably less than with captions. They need to be made aware that this may not be the case and that even without captions it is possible to understand a considerable number of television programs, especially when they view episodes of the same program successively.

Results from this study also indicate that television programs may be suitable for extensive viewing where L2 aural input is provided by learners watching videos (see Webb, 2014 for a more detailed description of this learning approach). A wealth of viewing resources can be provided through digitally streamed seasons of television. Beyond captions, there are a number of computer-assisted methods that may improve comprehension, including onscreen glossing and hyperlinked captions. Studies on the efficacy of these are needed.

Conclusion

The results of the present study indicate that the presence of captions can lead to increased comprehension of episodes of authentic television. Viewers with and without captions were able to make significant gains to their comprehension from the first to the tenth episode viewed. Unlike previous research where captions invariably led to increased comprehension, across the ten episodes viewed here captions were shown to make a significant difference in comprehension for episodes early in the viewing process and for those that may be considered more difficult.

About the Authors

Michael P. H. Rodgers is an Assistant Professor at Carleton University. His research interests include vocabulary acquisition, language learning through viewing videos, and extensive viewing and listening. He has published in journals such as *Applied Linguistics*, *Language Learning*, and *TESOL Quarterly*.

Stuart Webb is a Professor at the University of Western Ontario. His research interests include vocabulary, second language acquisition, and extensive reading and listening. His articles have been published in journals such as *Studies in Second Language Acquisition*, *Applied Linguistics,* and *Language Learning*.

References

Brett, P. (1995). Multimedia for listening comprehension: The design of a multimedia-based resource for developing listening skills. *System*, *23*(1), 77–85. https://doi.org/10.1016/0346-251X(94)00054-A

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511732959

Chung, J. M. (1999). The effects of using video texts supported with advance organizers and captions on Chinese college students' listening comprehension: An empirical study. *Foreign Language Annals*, *32*(3), 295–308. https://doi.org/10.1111/j.1944-9720.1999.tb01342.x

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

Danan, M. (2004). Captioning and subtitling: Undervalued language learning strategies. *Meta: Translators' Journal*, *49*(1), 67–77. https://doi.org/10.7202/009021ar

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition*, *27*(2), 305–352. https://doi.org/10.1017/S027226310505014X

Fletcher, J. D., & Tobias, S. (2005). The multimedia principle. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 117–133). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511816819.008

Garza, T. J. (1991). Evaluating the use of captioned video materials in advanced foreign language learning. *Foreign Language Annals*, *24*(3), 239–258. https://doi.org/10.1111/j.1944-9720.1991.tb00469.x

Gruba, P. (2004). Understanding digitized second language videotext. *Computer Assisted Language Learning*, *17*(1), 51–82. https://doi.org/10.1076/call.17.1.51.29710

Guillory, H. G. (1998). The effects of keyword captions to authentic French video on learner comprehension. *CALICO Journal*, *15*(1–3), 89–108.

Hirai, A. (1999). The relationship between listening and reading rates of Japanese EFL learners. *The Modern Language Journal*, *83*(3), 367–384. https://doi.org/10.1111/0026-7902.00028

Huang, H. C., & Eskey, D. E. (1999). The effects of closed-captioned television on the listening comprehension of intermediate English as a Second Language (ESL) students. *Journal of Educational Technology Systems*, *28*(1), 75–96. https://doi.org/10.2190/RG06-LYWB-216Y-R27G

Markham, P., & Peter, L. A. (2003). The influence of English language and Spanish language

captions on foreign language listening/reading comprehension. *Journal of Educational Technology Systems*, *31*(3), 331–341. https://doi.org/10.2190/BHUH-420B-FE23-ALA0

Markham, P., Peter, L. A., & McCarthy, T. J. (2001). The effects of native language vs. target language captions on foreign language students' DVD video comprehension. *Foreign Language Annals*, *34*(5), 439–445. https://doi.org/10.1111/j.1944-9720.2001.tb02083.x

Montero Perez, M., Peters, E., & Desmet, P. (2014). Is less more? Effectiveness and perceived usefulness of keyword and full captioned video for L2 listening comprehension. *ReCALL*, *26*(01), 21–43. https://doi.org/10.1017/S0958344013000256

Richards, J. C. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly*, *17*(2), 219–240. https://doi.org/10.2307/3586651

Rodgers, M. P. H., & Webb, S. (2011). Narrow viewing: The vocabulary in related television programs. *TESOL Quarterly*, *45*(4), 689–717.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88.

Schwartz, J., Fedak, C., & McG. (2007). *Chuck* [Television series]. Los Angeles, CA: NBC.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, *1*(2), 147–170. https://doi.org/10.1177/026553228400100203

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question type. *Language Testing*, *8*(1), 23–40. https://doi.org/10.1177/026553229100800103

Spaan, M. (2007). Evolution of a test item. *Language Assessment Quarterly*, *4*(3), 279–293. https://doi.org/10.1080/15434300701462937

Wagner, E. (2002). Video listening tests: A pilot study. *Columbia University Working Papers in TESOL & Applied Linguistics*, *2*(1), 1–39.

Webb, S. (2014). Extensive viewing: Language learning through watching television. In D. Nunan & J. C. Richards (Eds.), *Language learning beyond the classroom* (pp. 159–168). London: Routledge. https://doi.org/10.4324/9781315883472

Webb, S., & Rodgers, M. P. H. (2009). Vocabulary demands of television programs. *Language Learning*, *59*(2), 335–366. https://doi.org/10.1111/j.1467-9922.2009.00509.x