Maximizing the potential of captions for primary school ESL students’ comprehension of English-language videos

Feng Teng

To cite this article: Feng Teng (2019) Maximizing the potential of captions for primary school ESL students’ comprehension of English-language videos, Computer Assisted Language Learning, 32:7, 665-691, DOI: 10.1080/09588221.2018.1532912

To link to this article: https://doi.org/10.1080/09588221.2018.1532912

Published online: 10 Jan 2019.

Submit your article to this journal

Article views: 126

View related articles

View Crossmark data

Citing articles: 2 View citing articles
Maximizing the potential of captions for primary school ESL students’ comprehension of English-language videos

Feng Teng
Department of Education Studies, Hong Kong Baptist University, Kowloon Tong, Hong Kong

ABSTRACT
This study investigated the effects of captioned videos on ESL primary school students’ comprehension of video content. A total of 182 primary school students watched two short English story videos in one of three conditions: fully captioned videos (N = 62), keyword captioned videos (N = 63), and uncaptioned videos (N = 57). Each group included learners with higher and lower levels of English proficiency. Two videos were selected, and the second video was watched twice. After each video, all participants took a comprehension test, including global comprehension and detailed questions. Findings revealed that fully captioned group achieved the best results on the global comprehension questions. Significant differences between the fully captioned and keyword captioned videos on the detailed comprehension questions were not detected. Learners with a higher level of English proficiency and those who watched the video for a second time achieved better comprehension scores. These findings suggest that full captioning videos, rather than keyword captioning videos, should be considered when using video-based comprehension activities for ESL primary school learners. However, learners’ English level and the frequency of video viewing should also be considered.

KEYWORDS
Full captions; KEYWORD captions; English proficiency level; frequency; video comprehension

1. Introduction
In light of today’s growing demands for literacy in technologically advanced societies (Teng, 2017), discerning practical ways to enhance students’ English comprehension is an essential and timely goal (Cross, 2011). However, teaching primary school students learning English as a second language (ESL) to comprehend English videos may be challenging. In most cases, primary school students need a firm foundation in auditory and oral skills before they can comprehend audiovisual input, a
process that requires them to link what they have heard to what they can decode.

New evidence has supported the use of captions in enhancing learners’ comprehension (e.g., Rodgers & Webb, 2017). Captions, which turn videos into a storybook with a stream of written language presented synchronously with video and audio reinforcement, was employed for making content accessible to people who are deaf or hard of hearing. In recent years, several studies have supported claims that using captioning can robustly enhance second language learning, including listening comprehension (e.g., Sadat, Kouros, Yuya, & Tatsuya, 2017), vocabulary development (e.g., Montero Perez, Peters, & Desmet, 2018; Teng, 2018a), and reading comprehension (e.g., Muñoz, 2017). Theoretically and practically, captioned videos appear to be an effective means of instruction by helping learners grasp the gist of visual cues. In particular, visual associations created in memory along with the mnemonic power of imagery may increase the comprehensibility of video input by helping students associate words with actual objects and encourage them to learn imaging techniques; these developments could ultimately lead to better comprehension in ESL learning.

The main outcomes of previous studies focused on the effects of using captioned videos as a medium for vocabulary development (Koolstra & Beentjes, 1999) and the effects of captioned videos on learners’ attitudes toward using such videos as an aid for language development (Danan, 2004). However, research on using captions for primary school students’ comprehension remains under-explored, though captioned videos have been proposed as an effective learning medium for adults (e.g., Bean & Wilson, 1989). If a low level of vocabulary in primary school learners is a consequence of reading difficulty rather than its cause, then captioned videos may be an effective tool to encourage and support knowledge acquisition and may positively influence comprehension. Thus, the combination of visual and auditory stimuli (i.e., pictures and words) may provide theoretical as well as realistic possibilities to enhance primary school students’ comprehension of written texts.

Armed with the above knowledge, a reciprocal relationship between captioned videos and English comprehension may exist, which could increase understanding of word-based knowledge and background knowledge. These developments are necessary to overcome obstacles in school children’s reading comprehension. Such obstacles include (1) failure to figure out the connection between written and spoken words, (2) an inability to decode and recognize words from text reading, and (3) a lack of motivation to read (National Research Council, 1998). However, comprehension in ESL learners with limited English proficiency may be
hindered by difficulties involving English syntax, vocabulary, phonological representations, and text structures.

The extent to which learners can retrieve supplementary information to comprehend text information conveyed through captioned videos, particularly at the primary school level, remains unclear. The present study seeks to address this gap by examining the effects of three captioning types—full captioning, keyword captioning, and no captioning—on video content comprehension. To enhance knowledge in this area, two variables (i.e., learners’ language proficiency and the number of times the video is viewed) were also examined. It is argued that full captions may prevent learners’ capacity to comprehend videos (King, 2002), and keyword captions might be an alternative solution to this problem (Guillory, 1998). Researchers have also proposed a need to simplify full captions (Winke, Gass, & Sydorenko, 2010) but research performed to reduce the textual density through the provision of keyword captions has yielded inconclusive results (Park, 2004). For example, Montero Perez, Peters, and Desmet (2014) showed the full captioning group learners could achieve better results in listening comprehension than keyword captioning group learners. Park (2004) argued that learners’ English language proficiency level could influence the effectiveness of full captioning and keyword captioning videos. Additionally, the number of times learners are exposed to a target word during reading-while-listening activities has been shown to affect learners’ performance in comprehending and learning new words (Chen & Teng, 2017; Teng, 2016a, 2016b). Despite previous research indicating the significant influence these variables (full caption vs. keyword captioning, English proficiency level, word exposure frequency) have on learning effects, no study has thoroughly and simultaneously examined their effects for enhancing primary school ESL learners’ English comprehension.

2. Literature review

2.1. Theoretical framework for captions and comprehension

Captions, a brief L2 text or explanation accompanying a video, augment pictorial and audio input contained in a video. Empirical research suggests that captioned videos help build a link between the written and spoken words, thus leading to better English comprehension irrespective of learners’ educational level or language background (e.g., Peters, Heynen, & Puimège, 2016; Sadat et al., 2017; Winke et al., 2010). As Vanderplank (2016) argued, the process of watching captioned videos does not overwhelm viewers with bi-modal input but rather serves as a
support, providing multiple representations of the same input and helping learners better comprehend video texts.

In terms of the theoretical framework supporting the use of captions on ESL learning, Paivio’s (1986) dual-coding theory serves as an elegant explanation. According to this theory, two separate representational systems—the verbal system and the imagery system (i.e., mental images and pictures)—can be activated by each other because they are both dually coded. In line with this perspective, the text can be assumed to potentially supply a synopsis of dynamic speech; therefore, learners may attain better language recall when a word is learned in association with an appropriate nonverbal referent, as in captioned videos. In addition, according to the Multimedia Principle (Fletcher & Tobias, 2005), learners can better comprehend texts when words and pictures are presented together. Combining aural and visual input has been argued to offer learners opportunities to segregate aural input from video input into a meaningful unit while facilitating visualization (Ellis, 2005), thus enabling them to connect aural and visual inputs. These connections help learners comprehend input beyond their acquired language level as they receive input through two different channels (Danan, 2004). As such, captions appear to be effective in enhancing learners’ comprehension.

Comprehension is defined in the present study as a skill in the language learning process, specifically the preliminary stage of learning, noticing, and decoding word meaning while watching captioned videos. In the present study, listening comprehension refers to an understanding of the meaning of spoken words while reading comprehension denotes a process to construct meaning from texts. The comprehension tests in the present study were designed around Buck’s (2001) competency-based definition for comprehension. The present study measured three 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’ (Buck, 2001, p. 114). To this end, items were designed to measure global and detailed comprehension. Hence, comprehension of video content, as used in the present study, required learners to identify the main idea and details of spoken text viewed and heard in a video. This is an aspect that has been featured in taxonomies of listening skills.

Comprehension of video texts in a second language (L2) is a complex cognitive process. To comprehend written and spoken language in a video, L2 learners must think, infer, and discern what the speaker means, and relate an incoming message to their background knowledge. According to Gernsbacher (1995), comprehension is a structure-building
framework, involving at least three stages: first, learners need to act as a comprehension agent and lay a foundation of holding structures and initiating the comprehension process by processing received information. Second, learners must verify whether the incoming information is consistent with that in their formed structures. Third, learners must initiate the formation of a new substructure when the incoming information is not congruent with their previously formed structure. One feature of this framework is that learners may differ in their comprehension skills (Gernsbacher, Varner, & Faust, 1990) in terms of language as well as learners’ differences in building coherent mental representations (Derry, Sherin, & Sherin, 2014). Hence, comprehension of L2 video texts extends beyond a receptive process to constitute an interactive and dynamic process for which the combination of verbal and imagery systems may lead to rich and meaningful comprehension (Paivio, 1986). As noted in Mayer’s (2005) Principle of Multimedia Learning, learners may better comprehend texts presented in a multimedia format compared to any simple presentation form (e.g., either written words or images). Given that learners’ processing capacity in comprehending content knowledge is limited (Gernsbacher et al., 1990), activating two channels of sight and hearing (i.e., via the use of captioned videos) may help learners process information for better comprehension (Derry et al., 2014).

2.2. Captions and its potential effects on language comprehension

The effectiveness of captions in enhancing learners’ comprehension has been explored. Results suggest that captioned videos provide a greater depth of word knowledge processing, which may be useful in improving L2 comprehension (e.g., Goldman & Goldman, 1988; Koskinen, Wilson, & Jensema, 1985; Linebarger, 2001; Markham & Peter, 2003; Markham, Peter, & McCarthy, 2001; Rodgers & Webb, 2017). For example, Markham et al. (2001) sought to measure comprehension of a documentary and randomly assigned 169 university students learning Spanish to one of three treatment conditions: English captions, Spanish captions, or no captions. Two measures were considered: writing a summary of passage content based on information received from the videos (i.e., global comprehension) and answering 10 multiple-choice items (i.e., detailed comprehension). The findings revealed that learners in the English caption group outperformed those in the Spanish caption group and dramatically outperformed those in the no caption group on both measures. Although the study recruited students learning Spanish, ESL students could presumably enhance their English comprehension through caption exposure.
Linebarger (2001) focused on the effects of captions in learning to read, defined as the ability to comprehend written texts. Seventy-six children who had recently completed second grade participated in a 2 (captions; with vs. without) × 2 (narration; with vs. without) research design. Captions helped the students recognize more words, identify critical story elements in the videos, and improve their reading rates. As suggested by Linebarger (2001), recognition of new words, identification of video elements, and improved reading rates are key issues in enhancing comprehension. However, one caveat of Linebarger’s (2001) work was that the supporting texts prepared for learners were designed so average students in the same grade could read them with no difficulty. In practice, scripts for videos may not always be developed in this manner. This disparity suggests a need to use captions containing some words beyond learners’ reading competence levels, as was done in the present study.

Montero Perez et al. (2014) investigated the effect of captions on the comprehension of three short French videos. They assigned 226 university-level students to one of three conditions: fully captioned videos (Group 1; \( N = 81 \)), keyword captioned videos (Group 2; \( N = 75 \)), and uncaptioned videos (Control group; \( N = 70 \)). Comprehension was measured through global and detailed questions focusing on whether learners could comprehend key information in the videos. Results revealed no differences between the three groups on the detailed comprehension questions; for the global questions, however, Group 1 (\( M = 18.51 \)) outperformed Group 2 (\( M = 16.63 \)) and the control group (\( M = 16.37 \)). They contended that learners in the fully captioned group exhibited better overall comprehension than the other two groups. However, word frequency and lexical coverage were not considered in their study. The present study included these components and explored whether failure to perform well on the detailed question test could be attributed to the existence of low-frequency words in the videos.

More recently, Rodgers and Webb (2017) focused on L2 television programs with captioning, rather than the short videos often used in previous studies (e.g., Markham & Peter, 2003; Montero Perez et al., 2014). A total of 372 Japanese university students were recruited and divided into a captioning group (\( N = 51 \)) and a no captioning group (\( N = 321 \)) in Rodgers and Webb’s (2017) study. Learners watched ten 42-minute episodes of an L2 television program. The measures used in the study included true/false items, multiple-choice items, and sequencing items, all of which were designed to measure learners’ comprehension of details and their inferencing ability. Results indicated that the learners who watched captioned videos improved their comprehension at the beginning of the viewing process compared to those who watched the
uncaptioned videos. However, by the tenth episode, the learners in the no captions group could leverage their accumulated knowledge to a level where additional support from captions may not have exerted a substantial difference in comprehension. Rodgers and Webb (2017) found that comprehension varied from episode to episode, extending the findings of Montero Perez et al. (2014), in which learners in the captions group outperformed those in the no-captions group when viewing a video. However, participants in Rodgers and Webb’s (2017) study were required to watch 10 episodes, which may have equalized the overall results between the two groups. Participants presented with captions might have lost interest or patience after watching ten 42-minute videos. Thus, Rodgers and Webb’s (2017) argument that captions may not lead to the expected effect on participants’ comprehension by the final episode should be interpreted cautiously.

Others have suggested that having learners watch the same video repeatedly may be beneficial when comprehending video content. For example, Skouteris and Kelly (2006) studied children who repeatedly viewed an animated movie and found that their comprehension of the material improved. Mares (1998) studied four- and five-year-old children and found out that they demonstrated better comprehension of characters’ motives, emotions, and actions after watching a simple story four times rather than only once.

Winke et al. (2010) distributed an English multiple-choice test to various students to measure whether learners could comprehend the main points of the video stories. Results showed that the group who watched captioned videos twice demonstrated significantly better comprehension scores than those who watched the videos twice with no captions. However, Russian and Spanish students did better on comprehension when captioning was presented first, whereas Arabic and Chinese students performed worse. These findings suggested two trends: (a) learners exposed to captioned videos outperformed those who did not watch captioned videos in terms of English comprehension and (b) the effect of watching videos twice and with captions was influenced by whether the orthography of learners’ first language was similar to the target language.

Overall, several issues can be identified from the above studies. First, findings pinpoint a comprehension advantage for L2 students who are exposed to captions while viewing videos. L2 students across different target languages and proficiency levels have appeared to achieve substantial gains in English comprehension compared to those who were not exposed to captions. However, the extent to which keyword captioning leads to better comprehension over full captioning remains unclear. Guillory (1998) suggested that, to make videos beneficial for learners
with low proficiency levels, only presenting keywords as captions presents an effective approach for learners because it does not impose a heavy cognitive load. However, this outcome is inconsistent with that of Montero Perez et al. (2014), who indicated that full captioning is more effective for comprehension than keyword captioning. Hence, utilizing full captioning and keyword captioning videos in language learning require further exploration.

Second, controversy persists over using captioned videos for global comprehension and detailed comprehension. For example, Markham et al. (2001) suggested that captioned videos enhance global comprehension and detailed comprehension, whereas Montero Perez et al. (2014) argued that captioned videos only enhance global comprehension. Global comprehension, referring in the present study to understanding of the general meaning of audiovisual input, differs from detailed comprehension (i.e., comprehension of the details in language input). Discrepancies in these two types of comprehension may be related to learners’ ability to process text, their English proficiency levels, and whether they can integrate the knowledge with what they already know. These differences also prompt more research to enrich current findings.

Third, previous studies tended to focus on measuring university students’ comprehension through captioned videos, leaving young learners (e.g., primary school students) as a research gap. Context, environment, setting, and learners’ beliefs can all shape differences between young learners and adult learners in terms of language acquisition and progress. Teachers may need techniques to engage primary school ESL learners during in-classroom or out-of-classroom learning. In addition, videos from television programs and documentaries have been used (Markham et al., 2001; Rodgers & Webb, 2017); these choices leave storytelling videos, which may interest primary school ESL learners and particularly those with low proficiency, as another research gap.

Fourth, previous studies controlled learners’ language proficiency. Although homogenous language proficiency may render group comparisons more reliable, learners in any L2 or EFL context likely have different English proficiency levels. Differences in English proficiency may lead to varied results when using captioned videos. Research suggests that captioned videos offer ideal comprehensible input for L2 learners (Vanderplank, 2010), but tracking genuine language intake for better comprehension from captioned input remains challenging. For example, although advanced learners may be better able to access and obtain more comprehensible input, learners with lower proficiency seem to encounter more difficulty in comprehending videos (Chung, 2002). As argued by Vanderplank (2016), learners’ cognitive and affective engagement in
video comprehension varied substantially, potentially due to differences in learners’ proficiency in their target language. Captioning may not be beneficial for low-proficiency learners who are less able to connect auditory input to visual (i.e., they lack form-meaning mapping), which is essential to language comprehension (Taylor, 2005).

Finally, controversy exists over the effects of captioned videos on L2 learners’ comprehension when videos are played a second time. This issue is important, as it informs pedagogical decisions regarding whether it is meaningful to have L2 learners watch captioned videos a second time. However, available empirical data are insufficient in revealing whether the effects of repeatedly viewing a captioned video on L2 learners’ comprehension of video content are robust or simply related to learners’ English proficiency levels. For example, viewing a captioned video a second time may help learners determine a string of sometimes-incomprehensible input and draw attention to language in the video (Skouteris & Kelly, 2006)—but learners may also focus simply on particular aspects of language, leading them to lose interest and patience, as described by Rodgers and Webb (2017). Therefore, given that learners are required to attend to L2 forms to facilitate comprehensible input (Schmidt, 2001), the variable of frequency in video viewing requires further research.

2.3. Literature summary and research questions

To summarize, captions can be an effective approach to increasing L2 learners’ language comprehension while watching a video. However, different types of videos have been used in different studies. Videos, including television reports, television programs, and documentary videos, may not be types young learners would prefer to watch for language learning; primary school students may wish to watch episodes containing a storyline (e.g., Koolstra & Beentjes, 1999), which could lead to better comprehension by developing cumulative knowledge of the characters and storyline every time they watch. Such research may enrich the scholarship on primary school ESL learners, a relatively ignored group of learners compared to university students. In addition, more research needs to be conducted on the potential effects of captions (i.e., full captioning and keyword captioning) on English comprehension for primary school students with different English proficiency levels. Although some researchers may argue that captioning has theoretical benefits for foreign and second language learning, infusing praxis-based student-focused pedagogy into empirical investigations of global comprehension and detailed comprehension in the context of primary school ESL education is sorely
needed. Furthermore, the connection between learners’ language proficiency and the frequency with which they watch captioned videos has not been investigated in previous studies. These unsolved issues motivated the present study, which focuses on L2 primary school students by measuring global comprehension and detailed comprehension through a 2 (low vs. high English proficiency) × 2 (watching captioned videos once or twice) × 3 (fully captioned videos, keyword captioned videos, and uncaptioned videos) analysis of variance (ANOVA) design. Bearing these purposes in mind, the present study attempts to address the following questions:

1. Are any of the treatments (i.e., full captioning videos, keyword captioning videos, and videos without captions) superior to the others in terms of video content comprehension for ESL primary school students with different proficiency levels?
2. To what extent do learners with higher and lower English proficiency differ under each of the three captioning conditions?
3. Are any of the treatments (i.e., full captioning videos, keyword captioning videos, and videos without captions) superior to the others in terms of video content comprehension when videos were watched once or twice?
4. Does repeated watching of captioned videos lead to better performance in video content comprehension under any of the three captioning conditions?

3. Method

3.1. Research design

Data were subjected to a 2 × 2 × 3 ANOVA using three treatments (i.e., full captioning videos, keyword captioning videos, and videos without captions), two levels of English proficiency (low vs. high), and two videos (frequency: watched once vs. twice). Each group was split into two subgroups (i.e., low vs. high proficiency) according to an internal standardized English test. Dependent variables for measuring participants’ possible enhancement in language comprehension included a written recall protocol instrument and an open-ended test. The Newman–Keuls method, a stepwise multiple comparisons procedure, was employed to identify whether group differences were statistically significant at the 0.05 level.
3.2. Participants

In total, 182 sixth-grade students from three primary schools in Hong Kong were recruited for this study. In these schools, English language instruction begins in kindergarten, and English and Chinese are the main languages of instruction. All participants ($M_{age}=11.57$, $SD =1.08$) were enrolled in a required English course that focused on basic skills in reading, writing, speaking, and listening. The students reported Chinese (Cantonese or Putonghua) as their first language (L1).

At the start of the study, a total of 300 sixth-grade students from the three schools were invited to participate with the help of three primary school teachers belonging to the same Christian church as the author. The preliminary pool of willing participants included 220 candidates (70, 75, and 75 from schools A, B, and C, respectively). Two students from school B and six students from school C reported their mother tongue as English or Hindi; thus, their data were excluded from the analysis. In addition, prior to the experiment, all students in the three schools had taken a standard Cambridge English (YLE Flyers) test. According to the scoring criteria, eight students in school A, 10 in school B, and 12 in school C were at the intermediate level. Due to the limited number of students at this level, their data were also excluded from the analysis. Hence, the final number of participants from the three schools was 62, 63, and 57, respectively. Each school was randomly assigned to a caption group. Table 1 presents the characteristics of each group and their test results.

3.3. Materials

Two different English-language videos were selected for the three experimental groups. Video 1 (10’28’’) was a short video from the story ‘Snowy day’; Video 2 (10’20’’) was a short video from the story ‘To catch a thief’. Rodgers and Webb (2017) suggested that, to obtain a large amount of necessary L2 aural input, learners should be exposed to full-length television episodes running approximately 22–42 minutes. In the present

<table>
<thead>
<tr>
<th>Group 1 (school A, fully captioned group)</th>
<th>Group 2 (school B, keyword-captioned group)</th>
<th>Group 3 (school C, no captions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>Male</td>
<td>30</td>
<td>31</td>
</tr>
<tr>
<td>Language proficiency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low achievers</td>
<td>30</td>
<td>31</td>
</tr>
<tr>
<td>High achievers</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>Total number</td>
<td>62</td>
<td>63</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group 1 (school A, fully captioned group)</th>
<th>Group 2 (school B, keyword-captioned group)</th>
<th>Group 3 (school C, no captions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>Male</td>
<td>30</td>
<td>31</td>
</tr>
<tr>
<td>Language proficiency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low achievers</td>
<td>30</td>
<td>31</td>
</tr>
<tr>
<td>High achievers</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>Total number</td>
<td>62</td>
<td>63</td>
</tr>
</tbody>
</table>
study, the author assumed that primary school learners would find it difficult to remember all details when asked to answer detailed comprehension questions following each episode if videos were too long. The videos included a single narrator providing background information.

English teachers from each of the three schools were invited to discuss potentially suitable videos. They reached a consensus on the difficulty level and appropriateness of the videos ultimately chosen for the study. The teachers regarded the video images as supportive of the content but not highly explicit. The author transcribed the words and added full or keyword captions for all the videos through MAGpie. Captions were recorded as a highly synchronized, nearly verbatim text of the soundtrack (Teng, 2018a). Two videos were confirmed as having similar levels of difficulty based on the following criteria: First, the two videos had similar caption speed (approximately 90 words per minute), as suggested by Tyler et al. (2009). Second, the syntactic complexity of the captions consisted mostly of main clauses with limited subordination (Markham, 1999). Third, each video was composed of approximately 1,000 words (tokens), nearly 95% of which were 2,000 word-level tokens as measured by Cobb’s Lextutor’s VocabProfile (http://www.lextutor.ca/vp/). Figure 1 presents a snapshot of a fully captioned video.

Keywords were defined as those words deemed important for understanding the meaning of a sentence. The present study used the same keyword determination procedures as previous studies (e.g., Montero Perez et al., 2014; Park, 2004). In addition, the author invited the three experienced English teachers to watch the videos, read the transcripts, and underline the words they regarded important for understanding the
main idea of a sentence. Upon comparing the keyword determination procedures (Park, 2004) and the teachers’ suggestions, the author determined a final set of keywords (single or multiword), representing 15% of the entire transcript. The keywords were synchronized with the videos. Keywords appeared at the center of the captions, immediately before the word was spoken in the video. The duration of the presentation, depending on the length of the presented words, was set at 1–3 seconds. Figure 2 presents a snapshot of a keyword-captioned video.

3.4. Measures

Two comprehension tests were developed for each video: a written recall protocol instrument and a multiple-choice test. The first dependent measure determined students’ global comprehension. Participants were required to write a summary of the video content in English. This test was intended to measure participants’ ability to recall their comprehension of the text materials without the constraints of responding to questions posed by others.

In the second dependent measure, learners were asked to choose the best alternatives from four options for each question. The learners completed 20 multiple choice items immediately after watching a video. All questions focused on details of the video content. The items were developed with a focus on keywords but could not be answered directly without understanding the content; this format ensured that the questions addressed important information and discarded irrelevant information.
Also, group differences could not be plausibly compared without focusing on the captioned keywords. One example of this test (correct answers underlined) was as follows:

1. Why was Kate unwilling to join the party?
   A. She needs a break  B. She is driving a car
   C. She needs to work on her application to school  D. She is making a call

To determine the suitability of each video, a pilot study was conducted with 16 learners with similar educational backgrounds and intermediate English level. Pilot participants attempted the 20-item multiple-choice test and the written summary instrument. The learners in the pilot study expressed that the text and item difficulty levels were appropriate. The learners could also select the correct multiple-choice alternative at least 65% of the time at the lower end of the test performance range after viewing the video. An item analysis was performed for each of the pilot-tested multiple-choice items, and items that failed to perform at an acceptable level were altered or eliminated. The pilot learners also generated a reasonable set of written idea units with a mean score of 12 out of 20 within a 30-minute time limit.

### 3.5. Scoring

The scoring of the written recall protocol instrument was based on counting the number of learner-generated written idea units, elaborations, and distortions. This scoring system was adopted from that used by Markham et al. (2001), which included four categories: idea units, elaborations, distortions, and errors. An idea unit refers to a single thought from a passage; however, a sentence can contain more than one idea unit. Elaborations are not found within the passage but do have some relationship to the content of the passage. While distortions are like elaborations in that their contents are not found within the passage, they are unlike elaborations in that they are inaccurate because they do not have any relationship to the content of the passage. Errors are referred to as any grammatical or lexical errors found within the expressed ideas. The identified idea units and elaborations were scored for gist; fewer distortions and errors produced higher scores. A maximum of 20 points (5 points per category as determined by the three teachers in a joint session) were awarded to learners who could produce a reasonable number of idea units and elaborations without distortions and errors. Two teachers initially rated this test and achieved an overall inter-rater reliability of 0.92 when counting idea units, 0.82 for
elaborations, 0.85 for distortions, and 0.81 for errors. A third teacher was invited to rate if disagreements between the two raters occurred. Final decisions were based on majority opinion.

In terms of the scoring system for the multiple-choice test, one point was allotted to each correct answer with a possible total score of 20. Following a joint meeting of the three teachers and the author, all parties agreed that 20 items would adequately reflect the relatively short video content without consuming excessive classroom time. This test was also scored independently by two teachers who came to full agreement on answers.

3.6. Procedures

The author solicited the consent of participants and their parents prior to conducting the experiment. With approval from the three schools, students’ attendance would be considered a regular class. The author explained to the participants that they would be participating in a study regarding the use of videos and would need to take some tests. They were ensured that participation was voluntary and the results of the test would have no effect on their course grade. Participants were blinded to the main objective of the experiment. All participants were told to watch the videos and focus on the content. Each student worked individually at a computer with a headset. Students could take notes while watching the videos if they wished. Videos were available via a hyperlink. For Video 1, the hyperlink allowed students to watch the video only once; for Video 2, the hyperlink allowed learners to watch twice. Although students with poor English ability may need to watch a video several times or change the viewing speed to enhance understanding, the time students spent watching Videos 1 and 2 was almost the same in this study; students were not allowed to control the video speed or replay it, which ensured that the author could examine learners’ English proficiency levels, second-time video viewing, and the relationship with treatment effects.

Two teachers monitored the laboratory to ensure that each participant could watch the videos without difficulty and that the procedure was followed precisely. Once participants finished watching video, they completed two paper-and-pencil tests (the written recall protocol followed by the multiple choice test). Tests were administered in this order to avoid potential contamination of students’ written summaries by multiple-choice items. The time for the summary written test and the multiple choice test, as determined through the pilot study, was 30 and
20 minutes, respectively. The experiment took approximately 2 hours to complete.

### 4. Results

#### 4.1. Analysis of written recall protocol instrument (global comprehension)

Table 2 presents the results of the $2 \times 2 \times 3$ ANOVA on the written recall protocol test scores, measuring learners’ global comprehension. The main effects of treatment, performance, video, treatment $\times$ performance interaction, treatment $\times$ video interaction, performance $\times$ video, and treatment $\times$ performance $\times$ video were highly significant.

Table 3 presents the means and standard deviations for the treatment $\times$ performance interactions as determined by the written recall protocol instrument. The Newman-Keuls test showed that the treatment with fully-captioned videos resulted in significantly higher scores than that with keyword-captioned videos ($p < 0.05$); and the treatment with keyword-captioned videos produced significantly higher scores than that without captions ($p < 0.05$). This outcome was consistent across high and low proficiency learners, suggesting that fully captioned videos constituted the most effective means of increasing learners’ global comprehension, regardless of proficiency level. The Newman-Keuls test also showed that the learners with higher English language proficiency scored significantly higher than those with lower English language proficiency ($p < 0.05$). This result was consistent across the three groups, indicating that high proficiency learners outperformed low proficiency learners in global comprehension irrespective of the captioning condition. Overall, learners with high English proficiency in the fully captioned group achieved the best global comprehension.

The next step was to separate and compare the results of Videos 1 and 2. Table 4 presents the means and standard deviations related to the treatment $\times$ video interactions on the written recall protocol test after watching each video. The Newman-Keuls test results revealed that the
treatment using fully-captioned video resulted in significantly higher scores than that with keyword-captioned videos ($p < 0.05$), and the keyword-captioned videos produced significantly better scores than uncaptioned videos ($p < 0.05$). This finding was consistent between Videos 1 and 2, implying significantly better effectiveness of the fully captioned approach than keyword captioning and no captions in global comprehension, whether learners watched the video once or twice. The Newman-Keuls test results also revealed that watching Video 2 for the second time yielded better performance than watching Video 1 once ($p < 0.05$). This result was consistent across the three groups, implying that repeated viewing of captioned videos led to better global comprehension in each of the three captioning conditions. Overall, learners who watched the video twice in the fully captioned condition achieved the best results on the written assessment.

### 4.2. Analysis of multiple-choice test (detailed comprehension)

Table 5 presents the $2 \times 2 \times 3$ ANOVA results for the multiple-choice test scores for Videos 1 and 2. The main effects of treatment, performance, video, treatment $\times$ performance interaction, treatment $\times$ video interaction, performance $\times$ video, and treatment $\times$ performance $\times$ video were significant, suggesting that the treatment condition and video frequency each had a strong mediating role on multiple-choice test performance.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Video 1, mean (SD)</th>
<th>Video 2, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully-captioned videos</td>
<td>14.09 (2.94)</td>
<td>15.08 (2.85)</td>
</tr>
<tr>
<td>Keyword-captioned videos</td>
<td>10.56 (2.45)</td>
<td>12.11 (2.24)</td>
</tr>
<tr>
<td>Uncaptioned videos</td>
<td>5.52 (2.34)</td>
<td>7.16 (2.28)</td>
</tr>
</tbody>
</table>

Note. Possible score for each video (20 points).
fully-captioned videos and that with keyword-captioned videos ($p = 0.68$). However, the treatment with keyword-captioned videos produced significantly higher scores than uncaptioned videos for the learners with low proficiency levels ($p < 0.05$). The fully-captioned videos yielded the best scores for learners with high proficiency levels ($M = 32.78, SD = 3.01$), followed by the group watching keyword-captioned videos ($M = 28.64, SD = 2.98$) and the group watching uncaptioned videos ($M = 17.94, SD = 2.21$). The Newman-Keuls test showed that the differences between the three groups were significant ($p < 0.05$). Irrespective of captioning condition, learners with high proficiency levels presented significantly higher scores than those with low proficiency levels ($p < 0.05$). Overall, learners with high proficiency in the fully captioned condition achieved the best results in the detailed comprehension test.

The means and standard deviations related to the treatment × video interactions on the multiple-choice test are listed in Table 7. The Newman-Keuls test showed that when learners watched Video 1 only once, the treatment using fully-captioned videos did not produce a significantly better score than the treatment with keyword-captioned videos ($p = 0.61$). However, the score for the treatment with keyword-captioned videos was significantly higher than watching uncaptioned videos ($p < 0.05$). When the learners got a chance to watch Video 2 for the second time, the treatment of using fully-captioned video scored significantly higher than the treatment with keyword-captioned videos.
(\(p < 0.05\)), and the treatment of using keyword-captioned video scored significantly higher than the treatment with the group without captions (\(p < 0.05\)). Irrespective of the treatment, the learners who watched videos twice obtained significantly better scores than the learners who watched videos only once (\(p < 0.05\)). Overall, learners who were allowed to watch video for the second time in the full captioning group achieved the best results.

### 4.3. Key findings for research questions

The first question explored the different effectiveness of the three treatments (full captioning videos, keyword captioning videos, and videos without captions) on video comprehension for ESL primary school students with different proficiency levels. Results show that fully captioned videos were more effective for high proficiency learners’ video comprehension, including global and detailed comprehension. Although fully captioned videos were also effective for low-proficiency learners’ global comprehension, no significant difference appeared between full captioning videos and keyword captioning videos for learners with low proficiency.

The second question explored comprehension differences between learners with higher and lower English proficiency under each of the three captioning conditions. Results indicated that students with higher proficiency outperformed those with lower proficiency in each condition.

The third question explored the differences of the three captioning types in comprehension when a video was watched once or twice. Results imply that when the video was watched twice, the effectiveness of the fully captioned condition was outstanding for global and detailed comprehension. However, when the video was watched only once, the effectiveness of the fully captioned condition on global comprehension was detected, whereas it was negligible on detailed comprehension.

The fourth question explored whether repeated watching of captioned videos led to better performance on video comprehension in each of the three captioning conditions. Results demonstrated the effectiveness of watching the video twice rather than once in each captioning condition.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Video 1, mean (SD)</th>
<th>Video 2, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully-captioned videos</td>
<td>14.09 (3.13)</td>
<td>17.83 (3.14)</td>
</tr>
<tr>
<td>Keyword-captioned videos</td>
<td>13.46 (2.86)</td>
<td>15.74 (2.84)</td>
</tr>
<tr>
<td>Uncaptioned videos</td>
<td>7.01 (2.34)</td>
<td>9.09 (2.38)</td>
</tr>
</tbody>
</table>

Note. Possible score for each video: \(20 \times 1 = 20\).
5. Discussion

In light of these results, ESL primary school students appeared likely to demonstrate the best comprehension in the fully captioned video condition. When combining visual (i.e., pictures and words) and auditory stimuli, captions appeared to help ESL primary school students establish connections between spoken words and printed words and to expand their comprehension by drawing upon background knowledge and comprehension strategies (Goldman & Goldman, 1988; Teng, 2018a; Vanderplank, 2016). In particular, watching fully captioned videos seemed to help children construct better connections between auditory and visual input pertaining to events, characters, and other information in a story (Montero Perez et al., 2014). Through full captions, children may better comprehend key information about various subjects, thus facilitating comprehension of embedded knowledge presented in the videos. As noted by Markham and Peter (2003), full captions may help learners form strategies to determine how to summarize the main idea, predict events and outcomes in a story, draw inferences, and monitor coherence and misunderstanding. Using full captions thus seems to be an effective means of helping young learners concentrate on essential elements embedded in a video story.

By contrast, the mean scores achieved by learners in the keyword-captioning group were lower than in the fully captioned group. This discrepancy may be attributed to the fact that a sole focus on certain aspects of the videos (e.g., the keywords) did not help learners grasp the main content of a video story. Although the research questions did not address keyword structure, learners may have noticed the keywords and tried to focus on their meaning while missing other important information in the videos. As proposed by Goh (2000), learners may neglect some important parts of a story if they concentrate too much on the meaning of other parts. Learners provided with keyword captions may have been unable to direct their attention to certain elements or to interpret meaning holistically. In this case, students produced written summaries based on fragmented comprehension of the video story. This result is in line with Guillory (1998) but contradicts that of Montero Perez et al. (2014), who suggested that keyword captions do not yield better performance than no captions. The results from Montero Perez et al. (2014) may have occurred because, as the authors acknowledged, certain parts of the videos in their study were too challenging for learners due to low-frequency collocations. In the present study, approximately 95% of words were 2,000-word-level tokens. In addition, Montero Perez et al. (2014) used authentic television reports from a broadcasting house, whereas the present study used story videos. This difference may explain
why participants in the study by Montero Perez et al. (2014) failed to adequately interpret the meaning of the videos.

Given the positive results obtained across two differently captioned groups for the written recall protocol instrument, full captions may have reduced learners’ decoding load and allowed them to devote attention to meaning interpretation, as has been found in other empirical studies (e.g., Montero Perez et al., 2018; Pulido, 2007; Rodgers & Webb, 2017). In particular, the use of full captioning appeared to promote ESL primary school learners’ engagement in comprehending important elements in a story and thus helped them recall the video content more thoroughly. However, the effectiveness of fully captioned videos on English comprehension, including global comprehension and detailed comprehension, was more significant for learners with high proficiency and those who watched the videos twice. For learners with low proficiency and those who watched the videos once, the effectiveness of fully captioned videos on detailed comprehension was not significant. This finding indicates that primary school ESL learners, who may have trouble understanding a rapid, authentic native speaker in a video until they have gained considerable exposure to it (Linebarger, 2001; Markham et al., 2001), differed in their performance on global comprehension and detailed comprehension. Fully captioned videos may help primary school ESL learners improve global comprehension, which is in line with Montero Perez et al. (2014) who found that full captioning led to greater gains in global comprehension (i.e., writing the summary of a video story when measured immediately after viewing a video), compared to keyword captioning and no captioning. Nevertheless, a quantitative analysis of detailed comprehension (i.e., the multiple-choice test) in this study revealed different results. Discrepancies in multiple-choice test scores between the fully captioned and keyword captioned groups did not fully support the findings of prior research, which found that learners with access to fully-captioned videos had significantly higher comprehension scores than those in the keyword-captioned group (e.g., Guillory, 1998; Park, 2004). However, the multiple-choice test in the present study focused on information presented in keywords, which could explain why learners in the keyword-captioned group also earned good scores on the corresponding test. Additionally, even if learners recognized and memorized some details through the videos, they did not necessarily develop a better understanding of the meaning of the content. This hypothesis is beyond the scope of the present study, however, as only a comprehension test focusing on global and detailed information was included.

Finally, students’ language proficiency appeared to be a key factor in determining whether captions effectively improved content
comprehension. Previous research has shown that children’s ability to comprehend language through captions is influenced by their level of linguistic competence (e.g., Koolstra & Beentjes, 1999). Therefore, ESL primary school learners with better English skills would benefit more from captions than those with lower English proficiency. This trend suggests that students with better English knowledge are competent to such a degree that their language ability exceeds the threshold at which lower- and higher-level students can make sense of embedded meanings in certain content aspects (Taylor, 2005). This explanation is supported by the finding that learners with stronger English skills scored significantly higher than those with lower English skills, irrespective of the treatment condition. Students’ performance with higher English proficiency in the keyword-captioned group was better than that of learners with low proficiency in the keyword-captioned group but worse than learners with low proficiency in the fully captioned condition. High-proficiency learners in the fully captioned condition achieved the best results in global and detailed comprehension. In addition, learners who were allowed to watch the video for a second time achieved higher mean scores on comprehension questions. Watching a captioned video for a second time appeared to make it easier for learners to comprehend certain parts of the videos that were challenging when they watched the video the first time. This is a tentative outcome that complements the study by Montero Perez et al. (2014), as well as others (e.g., Pujola, 2002). Little research has explored the frequency with which captioned videos are watched, which would be a worthwhile focus for future studies (Teng, 2018a).

In terms of theoretical explanations of the results, learners’ proficiency levels played a key role in dictating whether they could be engaged with captioned videos. As suggested by Yeldham (2018), more proficient learners may be more likely to engage with the multiple cues embedded in captions and visuals. Differences in comprehension between high-proficiency and low-proficiency learners may enrich Mayer’s (2009) cognitive load theory. First, L2 learners seem to have constraints on their working memory when processing information as noted by Baddeley (2000). Such constraints on working memory may hinder learners’ ability to simultaneously handle the verbal and imagery systems proposed in Paivio’s (1986) dual-coding theory. Second, L2 learners with low proficiency may have limited working memory when processing aural and visual information. However, Vanderplank (2016) contended that watching captioned videos does not overwhelm the viewer with bimodal input but instead offers multiple representations of the same input to help learners better comprehend video information. In the present study, learners with low proficiency levels may have been overloaded and
unable to process multiple representations of information required per the multimedia principle (Fletcher & Tobias, 2005). This finding is inconsistent with the parallel processing model (Taylor, 2005), wherein learners can use various cognitive channels to process multiple sources of information. This pattern may apply to learners with high proficiency, as this model characterizes learners as being able to comprehend information automatically without devoting much working memory to handling multiple cues. Third, learners who watched the video for a second time may have developed strategies to compensate for difficulties comprehending the video information. This development likely involved strategies of planning, monitoring, and evaluating the comprehension process; predicting forthcoming information; and discerning unknown meaning (Teng, 2018b; Teng & Huang, 2018). Finally, detailed comprehension has been shown to be challenging for L2 learners, as they may need to narrow their cognitive focus to evaluate detailed information. In practice, L2 learners in such a position may favor overall comprehension over attending to or recalling detailed information. In a setting where video images and captioned information are presented together, learners would be expected to exploit the aural conduit; however, the benefits of resorting to captions for detailed comprehension, a skill requiring a complex cognitive process and exerted effort for primary school ESL learners, may warrant further study.

6. Implications and limitations

From the perspective of pedagogical utility, ESL teachers should consider using captioned videos to help students improve comprehension in the target language. Certain implications can be drawn from the better performance of the captioning group over other groups, namely, that English language input could be a powerful contributor to general comprehension if learners can cognitively process visual hints contained in captioned videos. In a classroom setting, if teachers want to use videos during English comprehension tasks, then fully captioned videos should be selected; they are more efficient in enhancing global comprehension in primary school ESL learners. However, teachers should always seek to ascertain whether ESL children possess sufficient English skills to ensure adequate comprehension of L2 video input. Furthermore, teachers may consider allowing primary school ESL learners to watch challenging videos a second time, which may consolidate knowledge grasped from watching the video the first-time.

The present study had some limitations. First, the soundtrack was presented in the target language only. Using a multilingual soundtrack may
yield better results considering that students likely have different language abilities. Likewise, multilingual captioning should be investigated more thoroughly in light of its potential to enhance comprehension. Second, ESL children in this study were not told about the educational nature of either medium (i.e., captions or no captions), nor were they instructed to use the captions to help them learn English. Making children aware of the rationale for using on-screen captions may inspire them to work harder to process input when captions are presented. One could explore this possibility in further studies targeting young learner groups. Such a targeted population could be then be taught media literacy skills. Gaining these skills could help them to become more critical in their viewing habits while also helping them to get the maximum benefits from captioning (e.g., improved comprehension). Third, although interactive advance-organizer strategies have been found to affect how language learners use L2 captions (Li, 2014), this factor was not included in the present study. Fourth, performance differences between male and female students were not analyzed. All these issues deserve further investigation. Finally, an analysis of students while watching videos should be included in future work to elucidate how learners process captions. For example, in subsequent studies, eye-tracking technology could be used to examine students’ reading behavior and attention allocation, which can quantify the effectiveness of using captions during language learning.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**Notes on contributor**

*Feng Teng* is a language teacher educator with extensive teaching and research experience in China. He is now studying for a Ph.D. degree in Hong Kong Baptist University. His professional interests include L2 writing, L2 vocabulary development, and identity research. He has published widely in international flagship journals, including *Thinking Skills and Creativity, TESOL Quarterly, Applied Linguistics Review, Spanish Journal of Applied Linguistics*, among others. He is a guest editor of a special issue on L2 literacy development in the *Asian EFL Journal*, as well as an issue on metacognition and L2 writing in the *Journal of Writing Research*.

**References**


