Language Learning & Technology
ISSN 1094-3501

June 2019, Volume 23, Issue 2
pp. 84–104

ARTICLE

How captions help people learn languages: A working-memory, eye-tracking study

Susan Gass, *Southeast University* and *Michigan State University*

Paula Winke, *Michigan State University*

Daniel R. Isbell, *California Polytechnic State University*

Jieun Ahn, *Michigan State University*

Abstract

Captions provide a useful aid to language learners for comprehending videos and learning new vocabulary, aligning with theories of multimedia learning. Multimedia learning predicts that a learner's working memory (WM) influences the usefulness of captions. In this study, we present two eye-tracking experiments investigating the role of WM in captioned video viewing behavior and comprehension. In Experiment 1, Spanish-as-a-foreign-language learners differed in caption use according to their level of comprehension and to a lesser extent, their WM capacities. WM did not impact comprehension. In Experiment 2, English-as-a-second-language learners differed in comprehension according to their WM capacities. Those with high comprehension and high WM used captions less on a second viewing. These findings highlight the effects of potential individual differences and have implications for the integration of multimedia with captions in instructed language learning. We discuss how captions may help neutralize some of working memory's limiting effects on learning.

Keywords: *Computer-Assisted Language Learning, Eye Tracking, Listening, Multimedia*

Language(s) Learned in This Study: *Spanish, English*

APA Citation: Gass, S., Winke, P., Isbell, D. R., & Ahn, J. (2019). How captions help people learn languages: A working-memory, eye-tracking study. *Language Learning & Technology*, *23*(2), 84–104. https://doi.org/10125/44684

Introduction

An important part of the process of learning a second language (L2) is paying attention to and making sense of the stimuli available from written text or speech. In listening, segmenting speech can be a daunting task for L2 learners. If that basic task is unsuccessful, learning from listening is unlikely. Captions[1] help learners process aural language by putting printed words on a screen as the words are heard, creating a visual means for determining linguistic units (e.g., words, morphemes, grammatical components). This helps students map the aural speech stream to individual (meaningful) words and phrases.

In our previous work (Winke, Gass, & Sydorenko, 2010, 2013), we investigated whether L2 learners made use of captions as a tool for listening comprehension in a classroom context. We examined how caption use was mediated by the learners' familiarity with the video content and with the teacher's ordering of video showings (with and without captions). We also investigated the impact of L1–L2 orthography differences. In post-task interviews, we found that captions helped learners with listening comprehension, and that they aided learners in segmenting speech streams into meaningful components. In this article, we extend our earlier work and look more into the how. Our main concern is with how individual differences impact caption use, with a focus on learners' working memory (WM) capacities. We do this through an investigation of two populations of learners, English as a second language (ESL) learners and foreign language learners (English speakers learning Spanish) to see what trends emerge across the language-learning contexts.

Captions Within Multimedia Learning and Working Memory Theories

Our investigation is grounded in *multimedia learning*, a cognitive theory that claims learning is deeper and longer lasting when there is both aural and visual information to support it (Mayer, 2014). Multimedia learning theory is based on the *multimedia principle*: "people can learn more deeply from words and pictures than from words alone" (Mayer, 2014, p. 1).

In his 2014 work, Mayer outlined underlying assumptions. First, multimedia learning occurs when there are multiple input channels (generally, visual and aural). And second, individuals have a limited capacity for information processing in a single mode. Thus, learners must be selective and attend to only certain input in a particular modality. Not being selective can result in a cognitive-processing overload (Paas & Sweller, 2014; Sweller, 2005). Components of WM (Baddeley, 2000) serve key functions. Information first enters temporary memory stores, including the phonological loop (responsible for encoding the sounds we hear) and visuospatial sketchpad (responsible for encoding visual information and spatial relations). The *central executive* controls the flow of information between the loop and sketchpad, and can call on other memory stores (e.g., long-term memory) for higher-level cognitive functions (e.g., to compare or contrast). These information pieces come together within the *episodic buffer*, and eventually, the information is more permanently stored in long-term memory for later retrieval. Thus, within the episodic buffer, a mental representation is constructed. Learning involves sifting through, organizing, and integrating information (sometimes across channels) to create new information (in particular, see Desjarlais, 2017, p. 125). As Wiley, Sanchez, and Jaeger (2014) wrote, "[e]ffective use of the working memory system is critical for successful learning" (p. 598). Similarly drawing on WM, language learners map linguistic form to meaning, aided by the central executive's functioning and directing of input. WM shows considerable associations with L2 comprehension and learning (Linck, Osthus, Koeth, & Bunting, 2014), and its relevance to both multimedia learning and second language acquisition (SLA) motivate our present focus.

The general notion is that captions, within a multimedia learning context, may counter limitations in WM that hamper the processing of spoken language. From the language learning side, comprehending L2 videos is a difficult task, requiring extra attention to linguistic form when compared to L1 comprehension, and thus, L2 comprehension taxes WM and other cognitive resources. In line with multimedia learning principles, captions visually signal relevant linguistic information (e.g., word boundaries, forms) to aid comprehension of speech in the aural channel. Of course, captions do not exist in visual isolation, and must be considered alongside moving images in a multimedia learning environment. At first glance, captions may appear to compete with video in the visual channel, risking cognitive overload for learners who split their attention too thinly (Ayres & Sweller, 2014). However, if a learner's attention is usefully drawn to captions to facilitate the processing of linguistic information, the learner is likely to temporarily ignore the video imagery. And because captions are physically and temporally overlaid on a video, switching between and integrating visual media can be facilitated. Per Ayres and Sweller (2014), when the cognitive load of video comprehension is manageable, such as when content is not completely novel or when a video's difficulty level is appropriate for language learners, integrating multiple input modes will likely be manageable and beneficial. But when the content itself is more difficult and includes too much novel input, the processing load increases, making the intake and use of the multiple modalities more difficult. If the content is familiar and not difficult, the learner may not need captions to facilitate processing; in such cases, captions could be seen as superfluous or even annoying.

Second and Foreign Language Multimedia Learning: The Role of Captions

Multimedia learning has been approached from a variety of perspectives, including research on web use (Chun, 2001), annotations for vocabulary use (Chun & Plass, 1996), picture annotation (Chun & Plass, 1998), and captioning in extended foreign-language television watching (Rodgers & Webb, 2017). This research has been important in applied linguists' understanding of multimedia learning and has set the stage for research on captions for L2 learners.

A meta-analysis by Montero Perez, Van den Noortgate, and Desmet (2013) included 18 experimental studies that examined the effects of captions on listening comprehension and vocabulary learning. They found captions statistically and significantly influenced listening comprehension and vocabulary learning. Indeed, in our study (Winke et al., 2013) on the effects of captions on listening comprehension and vocabulary learning, we supported these general findings, showing that captions had a large effect on vocabulary learning and comprehension as compared to no captions.

With the general utility of captions for L2 learning well-established (Rodgers & Webb, 2017; Vanderplank, 2016a, 2016b), we turn to other factors that may play a role in the instantiation of multimedia learning. Vanderplank (2016b) alluded to the vast precipice upon which L2 acquisition and captioning researchers now stand. He noted that researchers are yet to recognize (yet alone study) how learners use the language made available in captioned videos and how they use the multimedia material (with captions) to build their language abilities. He noted that learners apply different strategies when viewing captioned material. Vanderplank claimed that most of what applied linguists know about captions "is still largely anecdotal" (p. 246). These situations, factors, and contexts in relation to captions need investigation.

In a previous study (Winke et al., 2010), we considered caption use in repeated viewings, and investigated how L2 (Arabic, Chinese, Russian, and Spanish) and proficiency level influenced learning. For captions in repeated viewings, we found that watching a video with captions two times was most effective, a somewhat predictable outcome. More interestingly, when only having captions available for one of two viewings, there was a difference for script-similar L1–L2 combinations and script-dissimilar combinations: L1 English learners of Spanish and Russian performed better when captions were available in the first viewing, whereas learners of Arabic and Chinese performed better when captions were available second. In both cases, learners noticed new information (Schmidt, 2001) in the first viewing and used the second viewing to confirm what they had noticed. For learners without orthographic dissimilarity, during the first viewing, they were able to take advantage of the written word as initial intake; but for those where the L1 and L2 had significant orthographic distance, the aural information served as the initial intake. This latter group was not able to easily utilize the written script and relied on aural input to recognize new information.

Continuing to explore factors that affect caption use, we investigated how L1–L2 differences and content familiarity affected caption viewing (Winke et al., 2013). Using eye-tracking technology with learners of Arabic, Chinese, Russian, and Spanish, we found that Arabic learners spent more time reading captions than Russian and Spanish learners. The Chinese learners, while not distinct from the other groups on average, showed greater within-group variation in caption use. For content familiarity, only Chinese learners watching an unfamiliar video had significantly higher caption reading time, though other L2 groups showed a trend in that direction. This was consistent with the suggestion by Ayres and Sweller (2014) that multimedia learning may be more difficult for novel content. Ultimately, we concluded that major L1–L2 script differences are likely to affect how captions are used by L2 learners, at least at earlier stages of proficiency. Additionally, we found that when understanding the video was particularly difficult (i.e., due to unfamiliar content or lexis), learners directed their resources to fewer input modalities (e.g., largely ignoring imagery and using captions more or vice versa).

Other researchers have addressed the role of proficiency in how learners use and benefit from captions. By introducing mismatching information in captions and audio, Leveridge and Yang (2013) determined that more proficient English learners relied less on captions to process information, a finding also stated by Pujola (2002). A greater reliance on captions seems to require attentional trade-offs, as lower proficiency learners in Taylor (2005) reported difficulty in simultaneously processing audio, visuals, and captions. In terms of benefits, findings have been mixed. Lwo and Lin (2012) found lower-proficiency learners benefited more than higher-proficiency learners in terms of comprehension when provided with L2 English captions. However, other researchers reported larger benefits for intermediate and advanced L2 users when compared to beginners (Montero Perez et al., 2013; Neuman & Koskinen, 1992; Taylor, 2005;

Vanderplank, 1988).

Speed of caption presentation also impacts viewing behavior. Kruger (2013) found that as the speed of captions (manipulated by adjusting the total time each caption is visible) increased, learners shifted their focus to more static information, which in this case included presentation slides and pictures of a lecturer's face. Similarly, Mayer, Lee, and Peebles (2014) observed that while adding video to a slower-paced audio narration enhanced comprehension, adding captions to faster narration with video had no beneficial effect on understanding. In high-burden multimedia situations, learners cope by tuning out one or more modalities.

Captions have often been criticized as crutches when they are present but not needed (Mitterer & McQueen, 2009). For instance, Markham, Peter, and McCarthy (2001) wrote that "captions may provide too tempting a crutch for those students who are developmentally ready to test their listening skills in a captionless environment" (p. 341). Likewise, learners in the study by Danan (2016) suggested captions "might become a distraction or a crutch" if they "focused too much on the words on the screen instead of listening" (p. 14). Thus, there is a sense that at a certain point, captions may be redundant, detrimental to the process of learning to segment or parse speech, or not needed for comprehension. Researchers have noted that some learners may focus on the aural input and close their eyes to eliminate the visuals, thus reducing the written input stream (see also Winke et al., 2010), or ignore the captions to reduce the amount of incoming information (Taylor, 2005). Thus, there may be a zone within which captions are most useful, perhaps centered where the content difficulty level is not too far above or below a learner's ability level. But this zone may be larger or smaller, contingent on learning goals, individual goals, and linguistic factors relative to the text, the learner, or both. Indeed, Vanderplank (2016a) outlined how captioned video for language learners can help the learners not only segment speech and learn about culture, but also develop their literacy, especially when video watching is sustained over a long period of time (through watching series, or entire sets of film over a period of time as part of a learning strategy or as part of a language-learning curriculum). However, he noted that "the relative effectiveness of captioned viewing varies according to language level" (p. 4). We suspect that this is true with the added caveat that the viewer's language level must be not too far from the difficulty level of the captioned material for captioning to be effective and helpful. Other researchers have picked up on these notions. For example, Mirzaei, Meshgi, Akita, and Kawahara (2017), in a study of Japanese university students studying English in Japan, used partial captioning (akin to key-word captioning; see Montero Perez, Peters, & Desmet, 2014). However, in their study, the key-word captions were produced based on corpus linguistics and word frequency lists. They synchronized the captions with the video's speech through automatic speech recognition software to investigate whether English captions with only difficult or less-frequent words would be more beneficial for comprehension. They found that captions were more beneficial, no matter if they were full captions or partial and synchronized captions, but that partial and synchronized captions might be preferred, theoretically and pedagogically, because they encouraged more listening. Such research is important for understanding how captions work within the cognitive load theory, but more work is also needed to understand how learners' individual differences affect the process of using captions—any type of captions—beneficially.

A fuller understanding of the factors affecting L2 and foreign language caption use and resultant learning could be profitably used in both the design of instructional materials and the understanding of language learning more generally. We now turn to the present study which utilizes eye-tracking and WM measures to understand how full captions are used by L2 Spanish learners and L2 English learners.

The Current Study

In this article, our primary focus is on how language learners make use of captions and how WM influences learners' global caption reading behavior and video comprehension. To broaden the scope of previous research, we conducted two parallel experiments with participants sampled from two distinct populations: Spanish foreign-language learners (Experiment 1) and English second-language learners (Experiment 2). We did this in order to sample learners from two different learning environments, one with limited exposure to the target language and the other with greater exposure. The following research questions (RQs) guided both experiments:

1. Does captioning aid comprehension?
   Prediction: Captioned videos will result in better comprehension of video content (Experiment 1 only).
2. What is the relationship between WM and L2 video comprehension? Do learners with high WM capacity comprehend more than learners with low WM capacity?
   Prediction: Those with higher WM will have higher comprehension scores, given that WM aids the integration of information for overall comprehension.
3. What is the relationship between WM and caption-reading behavior? Are there differences in caption-reading behavior between learners with high WM and learners with low WM?
   Prediction: Learners with greater WM capacity will exhibit somewhat less reliance on captions, as they are better able to cope with multimedia input.
4. What is the relationship between caption-reading behavior and L2 video comprehension? Are there differences in caption-reading behaviors between learners who demonstrate high video-comprehension and learners who demonstrate low video-comprehension?
   Prediction: Based on our prior work (Winke et al., 2010, 2013), learners with low comprehension will likely use captions less than those with high comprehension.

General Method

For both experiments, we utilized eye-tracking for data collection. The assumption underlying eye-tracking is a strong association between eye movement and the human mind, referred to as the *eye-mind link* (Reichle, Pollatsek, & Rayner, 2012). In other words, one’s eye focuses on what one is thinking about, thereby giving researchers insight into the mind.

A handful of researchers have utilized this technology to study caption-reading behavior. In several eye-tracking studies, d’Ydewalle and colleagues (d’Ydewalle & De Bruycker, 2007; d’Ydewalle & Gielen, 1992; d’Ydewalle, Praet, Verfaillie, & Van Rensbergen, 1991) and Bisson, Van Heuven, Conklin, and Tunney (2014) found an attentional bias toward subtitling or captions during video watching. As explained by Ghia (2012), “subtitled audio visual texts are semiotically and communicatively complex works, and are characterized by a constant interplay among their aural and visual components. … The exploration of the acquisitional potential of all such components is of paramount importance” (p. 2). While eye-tracking technology can be used to examine the effect of specific conditions or linguistic features and trial-level reading behaviors (e.g., individual captions, d’Ydewalle & De Bruycker, 2007; individual words in a novel, Godfroid et al., 2018), we focus on global (i.e., aggregate) caption reading behaviors and global video comprehension to align the present study with our previous work on captions (Winke et al., 2010, 2013) and the work by other scholars (Bisson et al., 2014).

In these experiments, we used an EyeLink 1000 desk-mounted eye-tracking camera sampling at 1000 Hz (SR Research Ltd.) to observe caption reading behavior. Additional details can be found in the Appendix.

Materials

Video

For both experiments, we used a 4-minute 37-second video clip from a nature documentary, originally produced in English (also used by Winke et al., 2010, 2013). The subject matter was a story about wild bears, and it concerned a particular bear that was protecting her cubs from another bear. The video featured narration as well as on-camera commentary from a biologist. For Experiment 1, we translated and re-recorded the narration in Spanish (with a female voice actor) and added Spanish captions. For Experiment 2, we added captions in English to the original movie clip.

Comprehension Test

In both experiments, comprehension was measured by asking participants to recall the story and type a detailed summary. No time limit was given; most participants took 15–20 minutes. Participants could write in whichever language they wished. Following Winke et al. (2013), responses were scored polytomously (0–0.5–1) by idea units (*K* = 36; to read more on idea unit scoring, see Winke & Gass, 2016).

Working Memory Test

We used a reading span (RSPAN) test[2] to capture verbal WM, computer-delivered via E-Prime 2.0 software (Psychological Software Tools) and described in Conway et al. (2005; for further description, see the Appendix).

Procedure

We conducted the experiments one participant at a time in an eye-tracking laboratory. First, participants filled out a questionnaire that included information about their academic and language background. We began the experiment by calibrating the eye-tracker to each individual's eye movements. Next, participants watched the video on a computer screen twice while we tracked eye movements. After a short break between viewings, we recalibrated for the participant's eye movements. Immediately after video watching, participants completed the comprehension test, followed by the WM test. We compensated participants $20.00 USD for their time. Figure 1 is a schematic of the procedure.

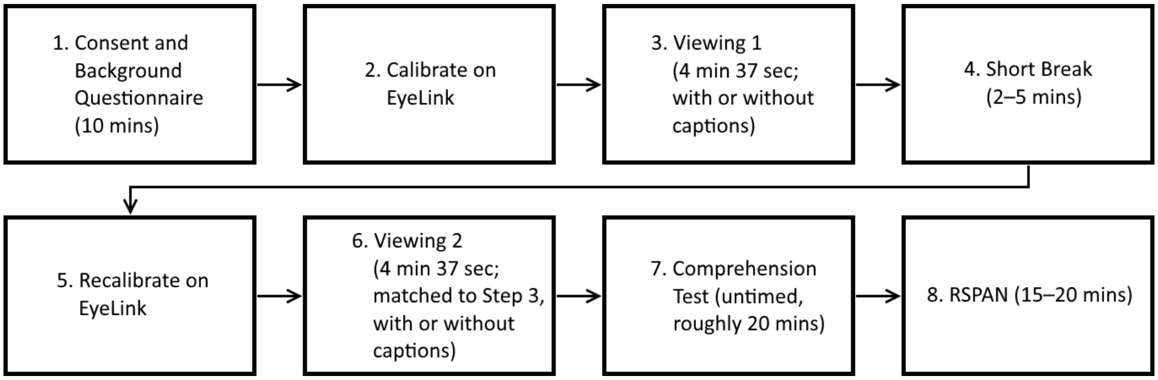
1. Consent and Background Questionnaire (10 mins)
2. Calibrate on EyeLink
3. Viewing 1 (4 min 37 sec; with or without captions)
4. Short Break (2–5 mins)
5. Recalibrate on EyeLink
6. Viewing 2 (4 min 37 sec; matched to Step 3, with or without captions)
7. Comprehension Test (untimed, roughly 20 mins)
8. RSPAN (15–20 mins)

Figure 1. Procedure for the current study

Analysis

To investigate the role of global caption reading behavior in video-based learning (RQ1), we compared sample means via a *t* test. Additionally, we computed change scores to make comparisons between first and second viewing caption reading measures. For the remaining RQs, which focus on the role of WM and global caption reading behaviors, we calculated descriptive statistics for all measures. We used Pearson correlations to examine relationships among the variables. To further explore these relationships, we divided the samples into (a) high and low WM and (b) high and low video-comprehension groups (Payne & Ross, 2005). As these subgroups were small, we used a Wilcoxon rank sum test (Mann-Whitney U), a nonparametric analogue to the two-sample *t* test. The effect size related to this test was reported as *r*.

To investigate how the participants used or read the captions when they were on screen, we created a rectangular area of interest (AOI) and several interest periods *where* and *when* (respectively) captions appeared on the screen. EyeLink software recorded the location and duration of eye fixations within the captions' AOI and interest periods. In both experiments, we used three global eye-tracking metrics, aggregated across captions in each viewing, to calculate attention and reading within the caption area AOIs:

- *Percent Caption Viewing Time* is the proportion of time a participant spends looking at captions (in ms) in relation to the total amount of time captions are on screen. A person who does not read the captions very often has a low percentage, while a person who makes frequent use of the captions a has a high percentage.

- *First Pass Time* is a metric used to demonstrate initial or early attention to an AOI (Conklin & Pellicer-Sánchez, 2016). Thus, we use First Pass Time to measure early or first, sustained attention to captions. This is the amount of time (in ms) participants spend reading an individual caption area before their gaze passes out of the caption area. The time of subsequent revisits to the AOI are not counted in this metric.
- *Rereading Time* is an eye-tracking metric "that signals more effortful and/or conscious processing" in reading (Conkin & Pellicer-Sánchez, 2016, p. 455). In this study, Rereading Time is the total amount of time (in ms) participants spend reading a caption after the first pass. For example, a participant may read a caption when it first appears, then look at an object in the video, and then reread the caption while it is still on the screen. This measure plus First Pass Time constitutes the *total gaze duration*. However, in this study, we are more interested in first pass and rereading separately, because they may indicate caption reading intentionality (first pass) and overall sustained caption use (rereading).

Experiment 1. Spanish Language Learners

Participants, Materials, and Procedures Unique to Experiment 1

Forty-six learners of Spanish (32 female) enrolled in fourth-semester courses at a large U.S. university participated in Experiment 1. They were undergraduate students pursuing 29 different academic specializations, with a mean age of 20.04 years (SD = 1.86); all participants but one (a Russian speaker) were native speakers of English.

We assigned the participants randomly to either a caption group (*n* = 23) or a no-caption group (*n* =23). The caption group watched a video with Spanish captions while the non-caption group watched the same video without captions. As illustrated in Figure 1, each participant watched the video two times.

On the free-recall comprehension test, the participants were free to write in English, their first language, or in Spanish. Two researchers scored the comprehension tests, and a third person adjudicated in case of discrepancies. The inter-rater reliability (total score correlation) was .98 before third ratings. For the Spanish version of the video, we set up 42 AOIs, synced to the times in which the captions were on the screen.

Experiment 1 Results and Discussion

RQ1. Captions Improve Video-Based Learning

We first investigated whether those who saw the videos with captions understood more (n = 23, M_{recall} = 8.27, SD = 4.16, min = 2.00, max = 16.25) than those who saw the videos without captions (n = 23, M_{recall} = 2.59, SD = 2.12, min = 0.00, max = 7.00). This difference was statistically significant (d = 1.72, t = 5.83, df = 32.2, p = 0.001). This result aligns previous research on whether captions aid in video comprehension.

With this confirmatory evidence of captions' beneficial effects on comprehension, we shifted our focus to better understanding the relations among WM, caption reading behavior, and video comprehension. In Table 1, we list recall test results, eye tracking metrics (First Pass Time in ms, and Rereading Time in ms), and RSPAN scores for those who saw the videos with captions. Change scores illustrate differences in caption reading behavior across the two viewings, Time 1 (T1) and Time 2 (T2). On average, this group, who watched the videos with captions, showed little change from T1 to T2, though there was considerable variation among the participants.

Table 1. *Recall, WM, and Caption Viewing Behaviors for Spanish Caption Viewers*

Measure	***N***	***M***	***Mdn***	***SD***
Recall	23	8.27	7.00	4.16
RSPAN	23	52.96	53.00	11.26
Caption Viewing % T1	23	74%	81%	23%
Caption Viewing % T2	23	74%	82%	22%
Change (T1-T2)		0%	-1%	9%
Mean First Pass T1 (ms)	23	2,049.07	2,325.05	627.87
Mean First Pass T2 (ms)	23	2,058.42	2,266.14	537.99
Change (T1-T2)		9.35	-53.81	259.85
Mean Reread T1 (ms)	23	323.72	282.05	148.51
Mean Reread T2 (ms)	23	317.15	324.48	146.66
Change (T1-T2)		-6.57	4.53	139.44

As a preliminary analysis to explore the relationships among recall, WM, and caption reading behaviors, we ran Pearson product-moment correlations (see Table 2). Aside from inter-dependent caption reading measures, we found few substantial correlations among the variables. More interesting were how comprehension (Recall) and WM (RSPAN) were related to caption reading behaviors. We did not find statistically significant results (meaning we cannot generalize these data beyond the current dataset), but, within this dataset, viewing captions, and particularly rereading them during T2, was negatively associated (although not significantly) with Recall scores to a small degree (r = -.22). WM also had small, negative correlations with rereading measures: -.21 for T1, and -.34 for T2.

Table 2. *Correlations Between WM and Caption Viewing Behaviors for Spanish Caption Viewers*

Measure	**Recall**	**RSPAN**	**CVT1**	**CVT2**	**1PassT1**	**1PassT2**	**RereadT1**
RSPAN	.16						
Caption Viewing % T1	.01	.12					
Caption Viewing % T2	-.14	.10	.91*				
Mean First Pass T1	.04	.06	.89*	.80*			
Mean First Pass T2	.03	.05	.80*	.83*	.91*		
Mean Reread T1	.14	-.21	.19	.03	.15	-.05	
Mean Reread T2	-.22	-.34	-.28	-.31	-.25	-.38	.55*

Note. N = 23

*p < .05

RQ2. Working Memory and Video-Based L2 Comprehension

Seeing that verbal WM and video comprehension (recall) had a weak correlation (r = .16), we next explored comprehension differences between learners with relatively low WM and learners with relatively high WM. We created two subgroups within the caption group: high WM (top third, n = 8, M_{WM} = 64.50, SD_{WM} = 4.87, M_{recall} = 8.78, SD_{recall} = 2.29) and low WM (bottom third, n = 8, M_{WM} = 41.00, SD_{WM} = 7.27, M_{recall} = 6.38, SD_{recall} = 5.25). The groups were significantly different in their verbal WM (W = 0, p < .001, r = .83), but there were no substantial differences on free-recall scores between the two groups, with the difference being small and not statistically significant (Wilcoxon rank sum test: W = 22.5, Z = 0.63, p = 0.528, r =

0.16). These data show that there were several participants with low WM who nonetheless were successful in summarizing (recalling) content from the video.

RQ3. Working Memory and Caption-Reading Behavior

To address RQ3 concerning the relationship between WM and caption reading, we first calculated descriptive statistics (see Table 3), of three eye-movement measures for the high- and low-WM groups.

In terms of total time spent reading captions, the high-WM group spent less time on T1 compared to the low-WM group, but this difference was neither statistically significant (*W* = 40, *p* = 0.431) nor particularly large (*r* = -0.20). However, the two groups moved in opposite directions during T2. The high-WM group spent slightly less time reading captions while the low-WM group attended to them more. While the two groups were not significantly different in their change scores (*W* = 47, *p* = .128), for this group of learners the effect was medium sized (*r* = -0.38).

We next observed that the differences in caption reading appeared to be associated with rereading, our measure of sustained caption use. Average First Pass Times for the groups did not differ on T1 (*W* = 32, *p* = 1, *r* = 0), nor was there a difference in change scores on First Pass Time (*W* = 36, *p* = .713, *r* = -.09). In average Rereading Times, however, there was an appreciable but non-significant difference at T1 (*W* = 21, *p* = .270, *r* = 0.28). Changes in rereading behavior from T1 to T2 at the group level were visible (i.e., a mean of -28.10 ms for the high-WM group and 45.60 ms for the low-WM group), but ultimately neither statistically nor practically significant (*W* = 35, *p* = .793, *r* = -.07). Again, we noted large individual variation in Rereading Times across the two groups.

Table 3. *Spanish Caption Reading Behaviors of High- and Low-WM Subgroups*

	High RSPAN (*n* = 8)			**Low RSPAN (*n* = 8)**		
Measure	***M***	***Mdn***	***SD***	***M***	***Mdn***	***SD***
Caption Viewing % T1	68%	78%	27%	76%	80%	20%
Caption Viewing % T2	65%	76%	29%	81%	86%	12%
Change (T1-T2)	-3%	-4%	8%	5%	2%	13%
Mean First Pass T1 (ms)	1,915.63	2,062.27	604.15	1,985.32	2,022.80	587.03
Mean First Pass T2 (ms)	1,937.30	2,106.63	530.73	2,057.82	2,271.88	574.43
Change (T1-T2)	18.67	-37.75	270.97	72.49	21.82	296.26
Mean Reread T1 (ms)	349.24	314.85	182.23	284.29	245.27	99.97
Mean Reread T2 (ms)	321.14	336.29	204.17	329.89	272.65	125.98
Change (T1-T2)	-28.10	21.14	163.02	45.60	18.38	121.15

RQ4. Caption-Reading Behavior and Video-Based L2 Comprehension

To explore the relationship between caption-reading behavior and video-based L2 comprehension, we divided the caption group into two subgroups (top third and bottom third) based on the free-recall test scores: A high-comprehension group (*n* = 9, *M* = 12.64, *SD* = 2.24) and a low-comprehension group (*n* = 8, *M* = 4.19, *SD* = 1.86; see Table 4). The difference between the two groups' free-recall test scores was statistically significant (*W* = .00, *Z* = -3.48, *p* = .001, *r* = .84).

Table 4 presents descriptive statistics of the high- and low-comprehension groups' caption-reading behaviors. The high-comprehension group's Percent Caption Viewing Time and First Pass Time were longer than the low-comprehension group; however, these differences were small and not significantly different (Percent Caption Viewing Time: *W* = 33, *p* = .810, *r* = .06; First Pass Time: *W* = 31, *p* = .665, *r* = .11). The high-comprehension group's Rereading Time was appreciably shorter than that of the low-comprehension group, but this difference was non-significant (*W* = 48, *p* = .136, *r* = -.37).

Table 4. *Spanish Caption Reading Behaviors of High- and Low-Comprehension Subgroups*

	High Comprehension (*n* = 8)			**Low Comprehension (*n* = 8)**		
Measure	***M***	***Mdn***	***SD***	***M***	***Mdn***	***SD***
Caption Viewing % T1	83%	84%	6%	71%	80%	26%
Caption Viewing % T2	83%	82%	6%	75%	84%	20%
Change (T1-T2)	-1%	-2%	7%	4%	4%	13%
Mean First Pass T1 (ms)	2,265.28	2,389.88	377.28	2,016.76	2,322.72	740.51
Mean First Pass T2 (ms)	2,230.38	2,277.63	281.73	2,094.71	2,302.87	586.53
Change (T1-T2)	-34.90	-205.88	314.56	77.95	26.87	245.17
Mean Reread T1 (ms)	283.86	268.05	96.93	375.74	354.90	149.86
Mean Reread T2 (ms)	235.25	269.05	117.06	385.05	387.61	132.84
Change (T1-T2)	-48.61	4.53	140.90	9.34	-9.57	170.28

Looking at changes in caption viewing from T1 to T2, the two groups trended in opposite directions in terms of Percent Caption Viewing Time: The high-comprehension group spent less time while the low-comprehension group spent more time. This difference, however, was neither significant nor very large (*W* = 45, *p* = .413, *r* = -.20). The high-comprehension group spent less time on the first pass reading and less time rereading. Conversely, the low-comprehension group spent more time on first pass reading and more time rereading during T2 than T1. These differences were not statistically significant, but for the change in First Pass Time, the effect size was *r* = -.28 (W = 48, *p* = .269). The change in Rereading Time had an effect size of *r* = -.13 (W = 42, *p* = .597).

Below we summarize the findings from Experiment 1:

- RQ1 - Captions promoted L2 video comprehension.
- RQ2 - Individual differences in verbal WM had little to no effect on video comprehension.
- RQ3 - Individual differences in verbal WM had some effects on caption-reading behavior. High-WM learners appeared to spend less time reading captions during T2 and appeared to spend less time rereading captions in general.
- RQ4 - Learners with relatively low comprehension (low recall scores) spent more time rereading captions than high-comprehension learners.

Experiment 2. English Language Learners

Participants, Materials, and Procedures Unique to Experiment 2

Twenty-four English learners (16 female) at the same large U.S. university participated in Experiment 2. We recruited them from an intensive English program (*n* = 5) or credit-bearing English for academic purposes courses (*n* = 19; undergraduates = 12, graduate students = 7). Participants' fields of study represented a broad range, from music to epidemiology. Their ages ranged from 18 to 28, with a mean age of 20.96 years (*SD* = 2.96). The group consisted of 15 Chinese speakers (1 also reporting Taiwanese as an L1), 5 Arabic speakers, 2 Malay speakers, 1 Urdu speaker, and 1 Bengali speaker.

As in Experiment 1, Experiment 2 participants followed the procedures in Figure 1, except we only showed them videos with captions. For the English video, there were 49 captions and corresponding AOIs. On the recall test, the participants wrote what they recalled from the video using English or their native language. Non-English responses were translated by native speakers into English. Two raters scored the responses by consensus; recall scores had an internal consistency (Cronbach's α) of .77.

Because administration of the RSPAN in English for non-native speakers represents a considerable threat to reliability (Sanchez et al., 2010), we worked with applied linguists who were bilinguals, and we oversaw their translation of the RSPAN into Simplified Mandarin Chinese and Modern Standard Arabic for the Chinese and Arabic speakers in this study (for details, see the Appendix). The KR-21 reliability index for all translated RSPANs was .86 (*n* = 19).

Experiment 2 Results and Discussion

Descriptive statistics for the recall, WM, and caption viewing measures are in Table 5. All participants successfully completed the recall test; 5 participants (all L1 Chinese) recalled in their native language. For the RSPAN, only 19 participants were able to take a version of the test in their L1. One participant's eyes could not be successfully calibrated; this participant's eye-tracking data were excluded from analysis.

Table 5. *Recall, WM, and Caption Viewing Behaviors for English Caption Viewers*

Measure	***N***	***M***	***Mdn***	***SD***
Recall	24	16.25	16.00	5.11
RSPAN	19	61.84	65.00	9.90
Caption Viewing % T1	23	55%	56%	10%
Caption Viewing % T2	23	51%	55%	14%
Change (T1-T2)		-4%	-3%	10%
Mean First Pass T1 (ms)	23	1,382.06	1,305.35	303.08
Mean First Pass T2 (ms)	23	1,236.05	1,314.08	415.84
Change (T1-T2)		-146.01	-117.49	312.31
Mean Reread T1 (ms)	23	340.09	308.53	165.63
Mean Reread T2 (ms)	23	372.15	395.49	182.22
Change (T1-T2)		32.06	34.57	135.45

The average recall score for the English learners was higher than that of the Spanish learners in Experiment 1, likely due to differences in global language proficiency. The English learners also spent less time reading captions overall, compared to the Spanish learners. The participants in this experiment tended to spend more time reading captions during T1, but they reread for longer periods of time during T2.

Correlations among variables are in Table 6. Comprehension (Recall) weakly correlated with most variables, but it correlated moderately with WM (RSPAN). WM, Percent Caption Viewing Time, and First Pass Time correlated moderately to strongly with each other.

Table 6. *Correlations Between WM and Caption Viewing Behaviors for English Caption Viewers*

Measure	**Recall**	**RSPAN**	**CVT1**	**CVT2**	**1PassT1**	**1PassT2**	**RereadT1**
RSPAN	.33						
Caption Viewing % T1	.11	.58*					
Caption Viewing % T2	-.13	.21	.70*				
Mean First Pass T1	.13	.54*	.85*	.54*			
Mean First Pass T2	-.12	.25	.66*	.91*	.66*		
Mean Reread T1	-.03	.11	.31	.31	-.23	.03	
Mean Reread T2	-.05	-.08	.16	.32	-.22	-.10	-70*

Note. Pairwise correlations are reported.

*p < .05

RQ2. Working Memory and Video-Based L2 Comprehension

Verbal WM had a weak correlation (r =.33) with video comprehension. To explore the differences between participants of distinct WM capacities, we again created high- (n = 8, M_{WM} = 70.50, SD_{WM} = 2.78) and low-WM (n = 8, M_{WM} = 51.00, SD_{WM} = 6.85) groups (based only on those who took an L1 RSPAN test). These two groups had significantly different verbal WM ability (W = 64, p <.001, r = .83).

Learners with high WM performed better on the free-recall test (M_{recall} = 19.44, SD_{recall} = 5.26) than learners with lower verbal WM (M_{recall} = 11.75, SD_{recall} = 5.06), although a Wilcoxon rank sum test did not quite cross the threshold of statistical significance, (W = 50.5, p = .058, r = .47).

RQ3. Working Memory and Caption-Reading Behavior

We next calculated descriptive statistics for caption-reading behaviors of the high- and low-WM groups (see Table 7). The group with higher verbal WM spent more time reading captions during T1 compared to the low-WM group, but a Wilcoxon rank sum test showed this was not a significant difference (W = 41.00, p = 0.147, r = .36). For First Pass Time T1, the high-WM group spent more time than the low-WM group did, but this difference was not statistically significant (W = 39.00, p = .224, r = .30). Similarly, the high-WM group's larger mean of caption Rereading Time during T1 was not significantly larger than the low-WM group's mean (W = 36, p = .385, r = .22).

Table 7. *English Caption Reading Behaviors of High- and Low-WM Subgroups*

	High RSPAN (*n* = 8)			**Low RSPAN (*n* = 8)**		
Measure	***M***	***Mdn***	***SD***	***M***	***Mdn***	***SD***
Caption Viewing % T1	59%	60%	9%	50%	53%	11%
Caption Viewing % T2	53%	54%	11%	49%	55%	21%
Change (T1-T2)	-5%	-6%	7%	-1%	2%	15%
Mean First Pass T1 (ms)	1,490.87	1,419.96	274.70	1,284.13	1,191.18	400.06
Mean First Pass T2 (ms)	1,344.05	1,322.43	399.78	1,203.26	1,373.22	535.96
Change (T1-T2)	-146.82	-163.87	323.64	-80.88	-115.85	394.97
Mean Reread T1 (ms)	350.33	312.75	180.23	288.66	262.53	161.75
Mean Reread T2 (ms)	331.34	387.27	173.40	361.46	413.02	212.41
Change (T1-T2)	-18.99	-20.20	100.14	72.80	34.57	167.23

Looking at the changes in caption reading behavior on T2, the high-WM group reduced their Percent Caption Viewing Time, while the low-WM group largely viewed the captions for a similar proportion of time. This difference in caption viewing changes was small and not statistically significant (*W* = 20.5, *p* = .416, *r* = -.20). Both groups reduced their First Pass Time (and to a similar degree) on T2 (*W* = 24, *p* = .685, *r* = -.10). The high-WM group reduced their average reading time on T2, while the low-WM group increased their average reading time on T2, but the differences between the high- and low-WM groups behaviors was small (*r* = -.25) and not statistically significant (*W* = 19, *p* = .325).

RQ4. Caption-Reading Behavior and Video-Based L2 Comprehension

Overall, most of the caption reading variables correlated weakly with comprehension (measured via free recall), and none of these correlations were statistically significant. However, what participants did during T2 was interesting: Participants who spent less gaze time on captions on T2 (Percent Caption Viewing Time on the first pass) tended to exhibit greater video comprehension.

To further explore this trend, we created two subgroups based on high and low recall scores, with the middle third of participants excluded. The high-comprehension group (*n* = 8, *M* = 22.12, *SD* = 2.33) recalled almost twice as much story content than did the low-comprehension group (*n* = 8, *M* = 11.31, *SD* = 2.62), and this difference was significant (*W* = 64, *p* < .001, *r* = .83). Descriptive statistics of the caption reading measures for each group are in Table 8. Both the high- and low-comprehension groups had similar behaviors in terms of overall caption viewing time, with each group spending less time on captions during T2. On T1, the groups exhibited similar caption reading behavior, with no visible or statistically significant differences in Percent Caption Viewing Time (*W* = 34, *p* = .874, *r* = .04), First Pass Time (*W* = 33, *p* = .958, *r* = .01), or Rereading Time (*W* = 33, *p* = .958, *r* = .01).

Table 8. *English Caption Reading Behaviors of High- and Low-Comprehension Subgroups*

	High Comprehension (*n* = 8)			**Low Comprehension (*n* = 8)**		
Measure	***M***	***Mdn***	***SD***	***M***	***Mdn***	***SD***
Caption Viewing % T1	54%	54%	8%	53%	55%	13%
Caption Viewing % T2	47%	42%	16%	51%	54%	16%
Change (T1-T2)	-8%	-7%	10%	-2%	-3%	10%
Mean First Pass T1 (ms)	1,367.84	1,304.49	241.64	1,344.31	1,363.22	398.00
Mean First Pass T2 (ms)	1,125.92	879.95	542.02	1,254.98	1,286.87	445.34
Change (T1-T2)	-214.91	-321.71	684.75	-89.33	-93.97	192.19
Mean Reread T1 (ms)	338.68	279.31	184.70	333.94	265.37	185.27
Mean Reread T2 (ms)	346.44	306.72	237.17	350.96	347.82	190.02
Change (T1-T2)	7.77	-38.66	138.98	17.03	-19.95	177.81

The two groups showed similar patterns of change in caption reading behavior on T2, with Percent Caption Viewing Times and First Pass Times dropping noticeably while Rereading Times changing relatively little. For Percent Caption Viewing Time, the high-comprehension group exhibited a slightly larger reduction, but this difference was not statistically significant (*W* = 22, *p* = .317, *r* = -.25). Similarly, a significant difference was not found for First Pass Times (*W* = 23, *p* = .372) across the groups, despite a visibly larger reduction for the high-comprehension group (*r* = -.22). There was no significant or substantial difference in Rereading-Time changes between the two groups (*W* = 31, *p* = .958, *r* = -.01).

Below is a summary of the findings from Experiment 2:

- RQ2 - Individual differences in verbal WM had a medium effect on video comprehension, as measured by free recall.

- RQ3 - Individual differences in verbal WM had small effects on caption-reading behavior. High-WM learners spent more time reading captions (First Pass Time) and then rereading captions on T1. High-WM learners also tended to reduce their caption reading to a greater degree than low-WM learners on T2.
- RQ4 - Learners with high- and low-comprehension (recall scores) had similar caption reading behaviors on T1. However, the high-comprehension group showed greater reduction in their overall caption reading time and their First Pass Time on T2 compared to the low-comprehension group.

Discussion

The purpose of the present study was to situate the use of captions within the broader field of SLA as well as in general learning theories. RQ1 (Experiment 1) confirmed previous work in which captions generally promoted L2 video comprehension. The other three RQs investigated individual differences (WM capacity) in samples from two different populations: Spanish foreign language learners and ESL learners. In Experiment 1, we focused on foreign language learners (learners with limited exposure to the language), and in Experiment 2, we researched L2 learners who were immersed in input-rich environments.

The first of the common RQs focused on the effect of an individual's verbal WM capacity on captioned-video comprehension, measured by a free-recall task following the watching of a short captioned video. We found that the two samples investigated in our study differed in that there was little effect of WM for the Spanish L2 learners, whereas there was a medium effect for the ESL learners. As noted earlier, the general proficiency of the ESL learners was assumed to be higher than that of the Spanish learners, supported by the fact that the recall scores of the ESL learners were higher than those of the Spanish learners. Although proficiency level was not featured in our RQs, we concur with Mirzaei et al. (2017) that it is an important variable in captioning research.

Proficiency level must be considered to fully understand how WM capacity impacts comprehension from video-based listening. In our literature review, we alluded to the idea that there might be a zone within which captions are most useful to individual learners, a zone in which the concepts referenced and the language used in the video are neither too far above nor too far below a learner's comprehension level. What we propose is that one needs to be at a certain level of proficiency before differences in WM capacity play a significant role in the comprehension of the video. In other words, there is a proficiency threshold (as described by Mirzaei et al., 2017) after which WM becomes relevant. At lower levels of proficiency, most of the effort in comprehending comes at a very basic level which consists of word-by-word interpretation (Leveridge & Yang, 2013; Pujola, 2002). It may be difficult to use individual words to create a meaningful stretch of speech; memory, therefore, may not be relevant, and one's WM capacity seems not to differentiate learners. But when closer to the proficiency threshold, WM does seem to differentiate learners' comprehension of captioned videos.

We further investigated the role of verbal WM and its effects on caption-reading behavior (RQ3). Interesting differences were found based on T1 versus T2 of watching the video. If WM played a role, one might expect that those with higher WM scores would spend less time reading captions during T2 given that one might be able to recall the text. In fact, for both groups (Spanish and ESL), behaviors were similar. When we looked at behavior at T2, the reading behavior of the Spanish and ESL learners was similar. The two WM groups went in opposite directions: the high groups reduced their caption reading time, whereas the low-WM groups increased their reading time. While these results are potentially interesting, we are cautious in our interpretation of their importance. We can only talk about a trend, but the trend does coincide with a review by Desjarlais (2017) who summarized L1 research on multimedia learning and suggested that individual differences, including WM, account for variation in information processing during multimedia learning (p. 131). A more robust sample size would likely reveal greater differences between those with low and those with high WM capacity.

RQ4 considered the relationship between comprehension and the behaviors that participants demonstrated while reading captions. We focused on the changes in reading behaviors from T1 to T2. In the Spanish L2 group, high-comprehension learners spent less time overall reading captions (First Pass Time) and less time rereading than the low-comprehension group, perhaps taking more from the audio or from the visuals than from the captions. For the ESL group, there were fewer differences between the high- and low-comprehension groups, perhaps reflecting their higher proficiency as well as their experiences living in an English-speaking country. They had learned to balance the multiple sources of input. In this study, we considered individual differences in WM capacity, specific reading behaviors when reading, and how those behaviors might relate to comprehension.

From prior research, we know that when viewers have enough experience with captions so that splitting their attention among the multiple-modes of input is not overly taxing (Ayres & Sweller, 2014), captions will be mostly beneficial, and will help viewers parse and understand the incoming speech stream. Our data support this perspective. Proficiency level (corresponding to experience) differences were found to impact two factors. First, those at higher levels of proficiency were differentiated by their WM capacity. In other words, once learners were able to balance multiple input sources, other individual characteristics came into play. Second, the two ESL groups (high- and low-comprehension) demonstrated little difference in their reading behaviors between T1 and T2. Experience in viewing captioned videos allowed them to create a more balanced approach to using multiple sources of input.

However, our results showed an apparent contradiction in the effect of WM capacity on video comprehension. The WM capacity of the Spanish L2 learners did not impact comprehension, whereas it had a medium effect on comprehension for the ESL learners. In other words, learners used captions regardless of their WM capacities, but the caption use likely depended in part on the learners' WM capacities and their L2 proficiency levels relative to the video content. WM capacity also appeared to be related to differences in caption reading behavior. With the English learners, WM capacity seemed to affect how they used captions in their second viewing: L2 learners with higher WM capacity did not use the captions as much as those with lower WM capacity. They might have been more able to hold key information effectively in the episodic buffer during T1 and, therefore, they might have gleaned more of the needed information from the captions the first time around.

Our data also show that captions are a powerful visual attraction even if one does not need them. High-WM learners tended to spend less time reading captions at T2, and less time reading captions in general, but they still read them, even if they might not have needed to. Captions appeared to be a more helpful processing aid to learners with lower comprehension relative to the difficulty level of the video. Captions seemed to provide important visual scaffolding that could be used to confirm only partially-acquired knowledge or to fill gaps in knowledge, such that a more complete picture could be obtained and retained.

To answer the larger question here, *How do captions help people learn languages?*, our results suggest that captions help with attentional control. In other words, the physical presence of written material on the screen provides an attention-drawing device in a way that normal spoken text, for example, does not. This salient, attention-grabbing written information is useful particularly when other information, such as aural information, becomes too difficult (i.e., high cognitive-load bearing) and, thereby, not accessible. Evidence for this claim came from learners in our study with more limited WM capacities who tended to use captions more. As explained by Wiley et al. (2014), "Learning from multimedia is a higher-order cognitive process that relies on many subprocesses to be successful" (p. 598). In their view, WM limitations are one of the primary motivations for using multimedia instruction. It can neutralize WM's limiting effects on learning.

In relation to general learning theory, we believe, as Plass and Jones (2005) suggested, that captions add yet another medium to the already complex picture of how language learners can benefit from multimedia input. Within multimedia learning, the assumption is that visual and aural input support one another. Learning is, therefore, aided because there are multiple sources to draw on for comprehension, reducing the burden of relying on only one source (Mayer, 2014). Multimedia learning must be designed in line with how the human mind works and, in the case of L2 learning, of how learning takes place. As Mayer (2014) stated, "[a] fundamental hypothesis underlying research on multimedia learning is that multimedia instructional messages that are designed in light of how the human mind works are more likely to lead to meaningful learning than those that are not so designed" (p. 43). Thus, reflecting on this research, and also taking into account Vanderplank's (2016a) view that captioned videos may work best when language learners are watching videos that match "their level of interest and their own level of proficiency in the foreign language" (p. 188), we conclude that captions most certainly aid comprehension, but the captioned multimedia must be chosen carefully by teachers. They have to be matched with the language learners' proficiency levels (Winke et al., 2010, 2013), and one must also consider the L1–L2 relationship. They also have to be chosen in consideration of learners' prior experience in learning with captioned multimedia. Finally, their use must be integrated into robust task-design. Most beneficially, captions appear to help counter-balance individual differences in cognitive processing ability, especially when learners are appropriately challenged (i.e., are viewing videos they can engage in and that are at the right difficulty level) and are motivated to learn from the videos. When comprehension is relatively high and strains on cognitive processing are not too great, captions support learning. Learners maximally use captions (for learning) when they perceive gaps in the aural input and use captions as a scaffold to figure out, reinforce, or confirm the aural input's meaning. The process of using captions to fill knowledge gaps (see Swain & Lapkin, 1995) in aural input helps learners with overall L2 vocabulary and forms (see Montero Perez, Peters, & Desmet, 2015; Sydorenko, 2010) and helps them, when they are successful in the gap-filling, improve their overall L2 comprehension.

Limitations

Our study had several limitations. Both of our experiments were small-scale and lab-based. We used short video clips in an eye-tracking lab. As demonstrated by Rodgers and Webb (2017), and as called upon by Vanderplank (2016a, 2016b), more longitudinal (and longer viewing time) research on captions is needed. Also, investigations into caption use by language learners in more natural (out-of-class) contexts are needed. Research in which language learners are asked to watch videos with the ability to voluntarily switch captions off or on would be informative, as caption-use information connected to WM capacities and proficiency could shed more light on what people do with captions to learn. Finally, we did not have a pre-experiment measure of proficiency. Instead, we relied on impressionistic observations, supported by evidence from recall scores. A next step is to observe how language learners with various cognitive-processing levels (relative to their peers) benefit from captions, but with the difficulty level of the captioned video matched to the general proficiency of the learners. Prior research studies on the benefits of captions relative to proficiency level (e.g., Leveridge & Yang, 2013; Lwo & Lin, 2012; Montero Perez et al., 2013; Neuman & Koskinen, 1992; Pujola, 2002; Taylor, 2005; Vanderplank, 1988) have only had the language learners, and not a set of videos, vary in terms of the trait of proficiency, and most certainly, any video itself can be rated and categorized on an external proficiency scale as well. A promising research direction is that taken by Mirzaei et al. (2017), who investigated learners' uses of smart captions: in our study, we only looked at full captions. Certainly, research on caption types and caption creation in relation to students' proficiency levels and individual differences is necessary. Researchers need to understand the complex interactions among the learners' proficiency levels, the captioned videos' difficulty levels, and the individuals' motivations and goals in watching (in connection with their cognitive abilities in multimodal processing) to fully understand how (and why and when) captioned-video enhances the L2 acquisition process.

In terms of analysis, we wish to emphasize two points. First, the sample sizes in both experiments were small, which only allowed us to conduct simple correlational and non-parametric analyses, limiting the causal interpretations and extrapolation of results. Second, our decision to carve out subgroups for WM and comprehension for exploratory analyses compromised the quality of those measures: continuous data were flattened into crude categories. While we found these subgroup analyses informative, as they showed how learners with measurable differences might interact with captioned videos, it made it more difficult to generalize the results. Future studies involving WM and captioned video comprehension would undoubtedly benefit from larger samples and more rigorous statistical analysis.

Conclusion

Mayer (2014, p. 44) emphasized the need for an “evidenced-based theory of multimedia learning” that can guide designers in the creation of effective multimedia. Understanding how multimedia input is processed by L2 and foreign language learners and understanding the role of WM capacity in relation to multimedia use are central to understanding the important role of captions in video-based listening, and hence captioning’s role in language learning. In previous research, researchers have considered what it is that learners focus on during captioned-video watching, but in this study, we wanted to shed light on how captions help people learn during captioned-video watching.

Investigating caption use by L2 and foreign language learners is well situated within multimedia learning theories. The next steps should be to better contextualize captioned learning material within language learning proficiency development and instruction.

Notes

1. In this article, we differentiate between captions and subtitles. The latter are in a different language than the language of the audio. Captions, on the other hand, are in the same language and are commonly used in language classes.
2. Two other WM tests were delivered to the participants, a visuospatial test (Experiment 1) and an operation span test (Experiment 2). In this article, we limit the discussion to the RSPAN because of its strong correlation with other measures of interest.

References

Ayres, P., & Sweller, J. (2014). The split-attention principle in multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 206–226). Cambridge, UK: Cambridge University Press.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423.

Bisson, M.-J., Van Heuven, W. J. B., Conklin, K., & Tunney, R. J. (2014). Processing of native and foreign language subtitles in films: An eye tracking study. *Applied Psycholinguistics*, *35*(2), 399–418.

Chun, D. (2001). L2 reading on the web: Strategies for accessing information in hypermedia. *Computer Assisted Language Learning*, *14*(5), 367–403.

Chun, D., & Plass, J. (1996). Effects of multimedia annotations on vocabulary acquisition. *The Modern Language Journal*, *80*(2), 183–198.

Chun, D., & Plass, J. (1998). Ciberteca [Computer software]. New York, NY: St. Martin’s Press.

Conklin, K., & Pellicer-Sánchez, A. (2016). Using eye-tracking in applied linguistics and second language research. *Second Language Research*, *32*(3), 453–467.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user’s guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786.

d’Ydewalle, G., & De Bruycker, W. (2007). Eye movements of children and adults while reading television subtitles. *European Psychologist*, *12*(3), 196–205.

d’Ydewalle, G., & Gielen, I. (1992). Attention allocation with overlapping sound, image, and text. In K. Rayner (Ed.), *Eye movements and visual cognition* (pp. 415–427). New York, NY: Springer Verlag.

d'Ydewalle, G., Praet, C., Verfaillie, K., & Van Rensbergen, J. (1991). Watching subtitled television: Automatic reading behaviour. *Communication Research, 18*(5), 650–666.

Danan, M. (2016) Enhancing listening with captions and transcripts: Exploring learner differences. *Applied Language Learning, 26*(2), 1–24.

Desjarlais, M. (2017). The use of eye-gaze to understand multimedia learning. In C. Was, F. Sansosti, & B. Morris (Eds.), *Eye-tracking technology applications in educational research* (pp. 122–142). Hershey, PA: IGI Global.

E-Prime (Version 2.0) [Computer software]. Sharpsburg, PA: Psychology Software Tools.

Gass, S., & Lee, J. (2011). Working memory capacity, stroop interference, and proficiency in a second language. In M. Schmid & W. Lowie (Eds.), *From structure to chaos: Twenty years of modeling bilingualism* (pp. 59–84). Amsterdam, Netherlands: John Benjamins.

Ghia, E. (2012). *Subtitling matters. New perspectives on subtitling and foreign language learning*. Oxford, UK: Peter Lang.

Godfroid, A., Ahn, I., Choi, I., Ballard, L., Cui, Y., Johnston, S., Lee, S., Sakar, A., & Yoon, H. (2018). Incidental vocabulary learning in a natural reading context: An eye-tracking study. *Bilingualism: Language and Cognition, 21*(3), 563–584.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford, UK: Oxford University Press.

Kruger, J.-L. (2013). Subtitles in the classroom: Balancing the benefits of dual coding with the cost of increased cognitive load. *Journal for Language Teaching*, *47*(1), 29–53.

Leveridge, A., Yang, J. (2013). Testing learner reliance on caption supports in second language listening comprehension multimedia environments. *ReCALL, 25*(2), 199–214.

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review*, *21*(4), 861–883.

Lwo, L., & Lin, M. (2012) The effects of captions in teenagers' multimedia L2 learning. *ReCALL*, *24*(2), 188–208.

Markham, P. L., 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.

Mayer, R. (2014). Cognitive theory of multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 43–71). Cambridge, UK: Cambridge University Press.

Mayer, R., Lee, H., Peebles, A. (2014). Multimedia learning in a second language: A cognitive load perspective. *Applied Cognitive Psychology*, *28*(5), 653–660.

Mirzaei, M., Meshgi, K., Akita, Y., & Kawahara, T. (2017). Partial and synchronized captioning: A new tool to assist learners in developing second language listening skill. *ReCALL, 29*(2), 178–199.

Mitterer, H., & McQueen, J. M. (2009). Foreign subtitles help but native language subtitles harm foreign speech perception. *PLoS ONE*, *4*(1). https://doi.org/10.1371/journal.pone.0007785

Montero Perez, M., Peters, E., & Desmet, P. (2014). Is less more? Effectiveness and perceived usefulness of keyword and full captioned video for L2 comprehension. *ReCALL*, *26*(1), 21–43.

Montero Perez, M., Peters, E., & Desmet, P. (2015). Enhancing vocabulary learning through captioned video: An eye tracking study. *The Modern Language Journal*, *99*(2), 308–328.

Montero Perez, M., Van Den Noortgate, W., Desmet, P. (2013). Captioned video for L2 listening and vocabulary learning: A meta-analysis. *System*, *41*(3), 720–739.

Neuman, S., & Koskinen, P. (1992). Captioned television as ‘comprehensible input’: Effects of incidental word learning from context for language minority students. *Reading Research Quarterly*, *27*(1), 95–106.

Paas, F., & Sweller, J. (2014). Implications of cognitive load theory for multimedia learning. In R. Mayer (Ed.). *The Cambridge handbook of multimedia learning* (pp. 27–42). Cambridge, UK: Cambridge University Press.

Payne, J., & Ross, B. (2005). Synchronous CMC, working memory, and L2 oral proficiency development. *Language Learning & Technology*, *9*(3), 35–54.

Plass, J., & Jones, L. (2005). Multimedia learning in second language acquisition. In R. Mayer (Ed.). The Cambridge handbook of multimedia learning (pp. 467–488). Cambridge, UK: Cambridge University Press.

Pujola, J.-T. (2002). CALLing for help: Researching language learning strategies using help facilities in a web-based multimedia program. *ReCALL*, *14*(2), 235–262.

Reichle, E. D., Pollatsek, A., & Rayner, K. (2012). Using E-Z Reader to simulate eye movements in nonreading tasks: A unified framework for understanding the eye-mind link. *Psychological Review*, *119*(1), 155–185.

Rodgers, M. R. H., & Webb, S. (2017). The effects of captions on EFL learners’ comprehension of English-language television programs. *CALICO Journal*, *34*(1), 20–38.

Sanchez, C. A., Wiley, J., Miura, T. K., Colflesh, G. J. H., Ricks, T. R., Jensen, M. S., & Conway, A. R. A. (2010). Assessing working memory capacity in a non-native language. *Learning and Individual Differences*, *20*(5), 488–493.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge, UK: Cambridge University Press.

Service, E., Simola, M., Metsänheimo, O., & Maury, S. (2002). Bilingual working memory span is affected by language skill. *European Journal of Cognitive Psychology*, *14*(3), 383–408.

Swain, M, & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning, *Applied Linguistics*, *16*(3), 371–391.

Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 19–30). Cambridge, UK: Cambridge University Press.

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

Taylor, G. (2005). Perceived processing strategies of students watching captioned video. *Foreign Language Annals*, *38*(3), 422–427.

Vanderplank, R. (1988). The value of teletext subtitles in language learning. *ELT Journal*, *42*(4), 272–281.

Vanderplank, R. (2016a). *Captioned media in foreign language learning and teaching: Subtitles for the deaf and hard-of-hearing as tools for language learning*. London, UK: Palgrave Macmillan.

Vanderplank, R. (2016b). ‘Effects of’ and ‘effects with’ captions: How exactly does watching a TV programme with same-language subtitles make a difference to language learners? *Language Teaching*, *49*(2), 235–250.

Wiley, J., Sanchez, C. A., & Jaeger, A. J. (2014). The individual differences in working memory capacity principle in multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 598–621). Cambridge, UK: Cambridge University Press.

Winke, P., & Gass, S. (2016). Using free recall and idea units for evaluating second language comprehension: Methodological choices and issues. *TESOL Applied Linguistics Forum*. Retrieved from http://newsmanager.commpartners.com/tesolalis/issues/2016-11-04/5.html

Winke, P., Gass, S., & Sydorenko, T. (2010). The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology*, *14*(1), 65–86.

Winke, P., Gass, S., & Sydorenko, T. (2013). Factors influencing the use of captions by foreign language learners: An eye-tracking study. *The Modern Language Journal*, *97*(1), 254–275.

Winke, P., Godfroid, A., & Gass, S. (2013). Introduction to the special issue: Eye-movement recordings in second language acquisition research. *Studies in Second Language Acquisition*, *35*(2), 205–212.

Appendix. Additional Details on the Methodology and Procedures

RSPAN Description

Participants sat at a computer and read increasingly long sets of sentences, with each sentence within a set presented one at a time. After a sentence appeared, participants indicated (by pressing a key) whether or not it made sense (some sentences had nonsense elements which rendered the sentence absurd), and then the participant saw a letter. After a subset of sentences and letters were presented, the participant had to recall the string of letters, in the correct order. The largest span of letters from the largest sentence-set one could remember correctly is his or her *reading span score*. Conway et al. (2005, p. 776) reported an internal consistency of .79 for the RSPAN when using the partial scoring method.Because the WM model we were using involved processing as an important component, we used the participants' plausibility judgments to ensure that they processed the sentences. Following Service, Simola, Metsänheimo, and Maury (2002), we modified some of the sentences so that implausibility was not apparent until the final two words (see Gass & Lee, 2011).

Eye-Tracking Details

For each experiment, we loaded the videos into an eye-tracking experiment file. A rectangular interest area was created and positioned to encompass all the captions in the video (for the captioned video). Additionally, we set up timed interest periods that we synced with the times in which captions were on the screen. This allowed us to extract total gaze times on the captions only when they were on screen.

During data collection, participants were seated 55 centimeters away from the computer monitor. They placed their head against a forehead rest and used a chin rest. We tracked their right eye (as described by and previously done by Holmqvist et al., 2011; Winke, Godfroid, & Gass, 2013). Although we carefully calibrated participants' eyes to the screen before each viewing of the video, eye drift was an issue due to the long viewing time. We watched for clear patterns of reading (i.e., word-to-word horizontal saccades) as well as eye-focus on action sequences (e.g., bears running along a riverbank) to ensure that a person's eye calibration was not drifting off. When we did see drift, we marked that on paper, and after data collection, we conducted individual *eye-washing*: that is, we manually moved the areas of interest in relation to the eye-movements on screen for that individual to correct for his or her actual drift during the experiment.

About the Authors

Susan Gass is Professor in the School of Foreign Languages at Southeast University in Nanjing, China, and University Distinguished Professor at Michigan State University. She has published widely in the field of second language acquisition including books on second language acquisition and research methods.

E-mail: gass@msu.edu

Paula Winke is an associate professor in the MATESOL and Second Language Studies programs at Michigan State University. She co-directs the SLS eye-tracking labs with Aline Godfroid. She is the 2012 recipient of the TESOL International Distinguished Research Award. She is co-editor (with Luke Harding) of *Language Testing*.

E-mail: winke@msu.edu

Daniel R. Isbell is an Assistant Professor in the English Department at California Polytechnic State University, San Luis Obispo. He has a PhD in Second Language Studies from Michigan State University and is interested in how technological affordances can support language learning, especially for less-commonly taught languages.

E-mail: isbell.daniel@gmail.com

Jieun Ahn is a doctoral candidate in the Second Language Studies program at Michigan State University. Her research interests include implicit and explicit learning, L2 reading, individual differences, and corrective feedback. To explore her research interests, she uses psycholinguistic techniques, such as eye-tracking and reaction times.

E-mail: ahnjieun@msu.edu